Overcoming Sparsity and Technical Noise: Systematic Evaluation of Imputation and Denoising Methods for Single-Cell Metabolomics and Lipidomics Data

Master's Thesis

Presented to the Faculty of Biosciences

of the Ruprecht-Karls-Universität Heidelberg

Marius Klein

Heidelberg, 2023

This Thesis was written at the Alexandrov group at the European Molecular Biology Laboratory (EMBL) in the period from February 2023 to July 2023 under the supervision of Tim Daniel Rose and Theodore Alexandrov.

1st examiner: Prof. Julio Saez-Rodriguez, Institute for Computational Biomedicine
2nd examiner: Prof. Ursula Kummer, Center for Organismal Studies (COS)

I herewith declare that I wrote this Masters Thesis independently, under supervision, and that I used no other sources and aids than those indicated throughout the thesis.

Heidelberg, Friday 07$^{\text{th}}$ July, 2023

Preface

This thesis was completed at the group of Theodore Alexandrov, Structural and Computational Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

Dr. Tim Daniel Rose supervised and accompanied the project.

Dr. Theodore Alexandrov accompanied the project.

Bishoy Wadie, Måns Ekelöf PhD and many others provided further guidance and support.

Luisa Abreu, Jeany Delafiori PhD, Alexander Mattausch, Sharath K. Menon, Dr. Luca Rappez, Dr. Mohammed Shahraz and Dr. Mira Stadler acquired and processed the datasets investigated in this work.

Glioblastoma cells were supplied by Srijita Banerjee, Dr. Susanne Kleber and Prof. Dr. Ana Martin Villalba at German Cancer Research Center (DKFZ), Heidelberg and pancreatic cancer cells by Dr. Marija Trajkovic-Arsic and Prof. Dr. Jens Siveke at Bridge Institute of Experimental Tumor Therapy (BIT), University Hospital Essen, Germany.

Abstract

Many human diseases with significant societal impact are closely linked to metabolism, underscoring the need for a deeper understanding of this fundamental cellular process. The emerging SpaceM technology, based on MALDI imaging mass spectrometry (imaging MS) and light microscopy, now enables the generation of single-cell metabolomics and lipidomics data, facilitating comprehensive studies of metabolism in complex and heterogeneous cell populations. However, the widespread sparsity and technical noise present in this data hinder the detailed analysis of important metabolites and their relationships. To my knowledge, no imputation strategies have been proposed yet to address these challenges in this novel field.

However, in other research areas and omics layers, a host of imputation and denoising methods have been developed, including singular value decomposition (SVD) and machine learning (ML) algorithms. In this study, a selection of them was adapted to four single-cell metabolomics and lipidomics datasets and benchmarked on simulated missing values using established and new evaluation metrics. Their performance was further validated in biological applications like correlation-based network inference.

Novel single-cell MS datasets seem to suffer predominantly from at-random and not-at-random missing values (MNAR and MAR), likely due to measurement sensitivity bias and batch effects. Imputation and denoising methods demonstrate mixed performance in overcoming these influences. Specialized denoising techniques like Markov affinity-based graph imputation of cells (MAGIC) and deep count autoencoder (DCA) preserve information and population structure most accurately, while SVD imputation achieves the best results with highly sparse data matrices. These methods also show reasonable performance in recovering biological information, such as differentially abundant metabolites and reaction-based lipid relationships.

This study highlights the specific challenges of sparsity and technical noise in single-cell metabolomics and lipidomics research and discusses how particular imputation and denoising methods can help overcome these issues. Additionally, it provides practical guidance for employing these and other methods throughout the entire single-cell analysis workflow. Alternative strategies for direct analysis and interpretation of sparse data are also discussed. By considering these insights, researchers can develop more comprehensive analysis strategies that account for the unique requirements of this novel technology and the underlying biology of metabolism.

Contents

List of Figures

List of Tables

List of Acronyms

2-DG 2-deoxy-D-glucose

AE autoencoder
ALRA adaptively thresholded low-rank approximation
AMPK AMP-activated protein kinase
ARI adjusted rand-index

BBKNN batch balanced k-nearest neighbors
bbMAGIC batch-balanced MAGIC

cAMP cyclic adenosine monophosphate
CV coefficient of variation

DAM differentially abundant metabolite
DAN 1,5-diamino-naphthalin
DCA deep count autoencoder
DHB 2,5-dihydroxy-benzoic acid
DR dropout rate

EMT epithelial-mesenchymal transition
ESI electro-spray ionization

GlcCer glycosyl-ceramide

HPC high performance computing

IL-17α interleukin17-alpha
imaging MS imaging mass spectrometry
IP$_3$ inositol trisphosphate

k**NN** k-nearest neighbors

L/S location and scale
LC liquid chromatography
LOD limit of detection
lyso PC lyso-phosphatidylcholine

m/z mass-to-charge ratio

MAGIC Markov affinity-based graph imputation of cells
MALDI matrix-assisted laser desorption/ionization
MAR missing at random
MCAR missing completely at random
MICE multivariate imputation by chained equations
ML machine learning
MNAR missing not at random
MS mass spectrometry
MSE mean square error
mTOR mammalian target of rapamycin

NB negative binomial

PC phosphatidylcholine
PCA principal component analysis
PDAC pancreatic duct adenocarcinoma
PTM post-translational modification

sc single-cell
scRNA-seq single-cell RNA-sequencing
SVD singular value decomposition

TMD transmembrane domain
TPCA1 (5-(p-fluorophenyl)-2-ureido)thiophene-3-carboxamide

UMAP uniform manifold approximation and projection
UMI unique molecular identifier

WGCNA weighted gene correlation network analysis

ZINB zero-inflated negative binomial

Chapter 1

Introduction

Obesity and an unbalanced diet are well-established risk factors for a variety of diseases, including metabolic syndrome, diabetes, and cardiovascular disease [1]. Their rising prevalence in Western societies not only impairs the health of individuals but also increasingly pressures health systems and national economies [2, 3]. In many disorders, the adverse effects of obesity are mediated through changes in cellular metabolism. These changes include insulin resistance, imbalanced lipid metabolism and storage, mitochondrial dysfunction, and chronic inflammation [4, 5, 6]. Given the strong connections to other cellular processes, metabolism plays a central role in determining cell fate, and on a large scale health and disease states. Accordingly, changes in cellular metabolism are involved in the development of other heart, inflammatory, and neurodegenerative diseases [7, 6, 8] as well as in cancer formation and progression [9], even though an association with lifestyle choices might be less obvious for some of these illnesses. In either case, these implications underline the rising interest in comprehensively elucidating the precise regulation and interplay of cellular metabolism with health and disease.

1.1 Metabolomics and lipidomics research

The research area that seeks to understand complex biological systems mechanistically is often referred to as systems biology. With the rapid development of comprehensive high-throughput *omics* methods and new opportunities to quickly and quantitatively characterize cellular states, computational and systems biology gained increased attention. Genomics, transcriptomics, and proteomics have evolved sequentially with respective technological advances in sequencing and mass spectrometry. They each create precise images of the genetic predisposition, situational transcription program, or abundance of proteins within a biological sample [10]. This has later been complemented by metabolomics and lipidomics methods which provide a rather phenotypic overview of the metabolic and structural states of biological samples [11]. Altogether, omics technologies generate large amounts of data that are sought to enable a holistic understanding of physiological and pathological processes in biological model organisms and patients. This way, they provide a framework for practical medical applications such as drug and biomarker discovery and comprehensive molecular profiling for

personalized diagnostics and therapies. Therefore, the investigation of different omics layers is considered crucial for the widespread implementation of precision medicine [12].

Within this field, metabolomics and lipidomics are particularly important to bridge the gap between the genotype and the phenotype of cells and organisms. Both specialties use high-throughput qualitative or quantitative methods to measure a variety of biomolecules (amino acids, sugars, and lipids) that represent the basic building blocks of cells. These molecules supply cells with energy and materials to synthesize proteins, nucleic acids, and membranes. The system of converting and providing the required metabolites is defined as metabolism. To adjust to a rapidly changing cellular environment, the metabolism regulates itself tightly and maintains a host of reciprocal interactions with other cellular processes [13, 14]. For instance, the metabolic state of a cell is modulated by signal transduction [13]: through second messengers like cyclic adenosine monophosphate (cAMP) and inositol trisphosphate (IP_3), metabolism-related pathways like AMPK and mTOR, or metabolite sensing transcription factors [15, 16]. In turn, many lipid species act directly as signaling molecules to control other cellular functions such as angiogenesis, immunity, and differentiation [17, 18, 19]. Moreover, individual metabolites serve as co-factors and modulators of structural changes such as epigenetic modification [20] and post-translational modification (PTM) of proteins [21]. These interactions illustrate the extensive interdependence of metabolism, cellular homeostasis, and development. Their mutual influences become even more apparent in disease states, underscoring the need for comprehensive metabolomics and lipidomics studies. Thus, this line of research is crucial to identify the underlying mechanisms that drive an illness and to develop effective treatment strategies.

1.2 Spatial and single-cell studies

Most metabolomics studies are based on mass spectrometry (MS) due to its exceptional sensitivity and high speed. Using electric fields, a mass spectrometer accelerates or deflects ions inside a vacuum, separating and detecting them based on their mass-to-charge ratios (m/z). The accurate masses from the collected m/z spectra can be used to infer sum formulas and annotate metabolite species of all ions salvaged from a given sample [22].

In most setups, the mass spectrometer is paired with an additional separation system such as a liquid chromatography (LC) column or an ion-mobility spectrometer to reduce the complexity of acquired mass spectra and guide identification of molecule species based on multiple characteristics (e.g. m/z and polarity). In a targeted setting, MS can quantify metabolites from a pre-defined set by using internal standards [23]. In contrast, untargeted approaches aim to measure and identify as many metabolite species in a sample as possible. Consequently, the latter approach solely generates relative or semi-quantitative information on metabolites and is typically employed for exploratory analysis and hypothesis generation [24].

In order to make uncharged biomolecules measurable in a mass spectrometer, they need to be ionized. Many biological applications use MS in positive ionization mode where molecules

are given a positive charge. Apart from trypsinized peptides in proteomics, this is often employed to detect various lipid classes with positive head groups [25, 26]. The negative mode has recently been favored for smaller molecules [27], but a lot of the central metabolites can be detected in both modes [28]. Depending on the aim of the study, different practical ionization techniques are utilized. Many targeted systems use electro-spray ionization (ESI) that is directly connected to an upstream separation system (LC) to analyze samples dissolved in a liquid solvent. By applying a high voltage to the ESI source, the solution is introduced into the MS as a fine aerosol of charged droplets. As the solvent evaporates, charges are transferred to the biomolecules, turning them into gas-phase ions that can be measured by the MS [22]. ESI mass spectrometers typically use tandem measurement approaches that include ion fragmentation. This ability to repeatedly analyze smaller ion fragments makes ESI-based instruments particularly useful for further structure elucidation of the measured ions [29].

In contrast, most matrix-assisted laser desorption/ionization (MALDI) systems only acquire MS1 spectra (without fragmentation). These instruments analyze fixed samples, such as tissues or cells on a solid surface like a glass slide. The sample is coated with an organic matrix like 2,5-dihydroxy-benzoic acid (DHB) or 1,5-diamino-naphthalin (DAN) and then ablated using an ultraviolet or infrared laser [30]. At the focal point, mainly the matrix absorbs the laser energy and is ionized, leading to the desorption of the matrix-analyte mixture and secondary transfer of charges to the analyzed biomolecules. Yet, the exact process of this charge transfer is still unclear [31]. With accurate control of measurement positions, MALDI is well-suited for spatial analyses of biological samples, known as imaging mass spectrometry (imaging MS). Here, the laser ablates the sample at various locations in a grid pattern and the acquired metabolite intensities are combined to create ion images [32]. Recent technological advances in MALDI imaging have enabled us to study metabolomics on an increasing spatial resolution and thus aided in understanding tissue- and even cell-level heterogeneity of the metabolism. This had a great impact on research of complex, heterogeneous tissues like brain [33] or tumor [34, 35].

Further advances in laser precision and ablation mark diameter have decreased the spatial resolution of imaging MS below 5 µm [36], a range where multiple pixels overlap with single cells. By overlaying the ion images with light microscopy images, computational methods register pixel-based metabolite measurements with spatial structures like tissue areas [37] or individual cells [38]. An emerging method to resolve imaging MS data to single cells is SpaceM. It combines high-resolution MALDI imaging with light microscopy to generate single-cell metabolomics or lipidomics data. The method has been optimized for subconfluent cells in culture and works in a high-throughput manner ($>$ 1000 cells/hour). It can also incorporate fluorescence-based measurements from light microscopy and maintain spatial information about the cells. SpaceM generates a combined matrix of metabolic, fluorescence, and spatio-morphological data [38].

The SpaceM method includes four steps: i) pre-MALDI light microscopy imaging and cell segmentation register the positions of individual cells. ii) MALDI imaging detects metabolites

in an untargeted fashion and is followed by iii) post-MALDI light microscopy imaging to register ablation marks. iv) The computational analysis includes annotation of exact masses to molecular sum formulas and molecular species using a metabolite or lipid database. Finally, the single-cell data is generated by pixel-cell deconvolution, namely the conversion of ion intensities from pixels to single cells based on the colocalization of cell and ablation mark positions. The consensus method of deconvolution normalizes ion intensities to the sampling proportion (cellular overlap of the pixel) of the respective pixels and combines them weighted by their sampling specificity (proportion of cellular overlap of the pixel with the cell of interest). Pixels are disregarded if they overlap multiple cells.

Compared to bulk experiments, these single-cell studies provide a more detailed understanding of cellular structure and metabolism. By capturing the metabolic profiles of individual cells, they reveal the inherent heterogeneity in cell populations and uncover volatile and intermediate cellular states [38, 39]. In the near future, this approach could also enable the deconvolution of complex tissue samples and the dissection of the metabolic contributions of different cell types within the tissue. Taken together, single-cell metabolomics and lipidomics provide valuable insights into the dynamic landscape of cellular metabolism, leading to a deeper understanding of physiological and disease mechanisms and guiding the development of targeted therapeutic strategies.

To pave the way for these insights, the spatiomolecular data matrices generated by SpaceM can be further analyzed functionally, for example with respect to the separation of conditions and cell types or by identifying differentially abundant metabolites (DAMs) that drive this separation. However, with the relative novelty of the field, a host of technical challenges as well as a general lack of standardization still limit the reproducibility and expressiveness of the data and outcomes of downstream analysis [40]. In this regard, metabolomics and lipidomics research can profit greatly from previous efforts to standardize and harmonize data processing and analysis in the more established *omcis* specialties as well as reuse and adapt methods that were developed in the respective areas.

1.3 Technical challenges

To improve the reliability of metabolomics and lipidomics analyses and results, various technical issues throughout the data acquisition, processing, and analysis process need to be addressed. Many of these challenges are shared with proteomics due to the common technological platform, but also single-cell transcriptomics studies suffer from some of the limitations as all these methods produce similarly structured data.

While modern sequencing technologies have been optimized to distinguish a small set of nucleotide bases very consistently, MS instruments detect a huge variety of biologically relevant and irrelevant ions. External influences such as sample preparation or ambient conditions create background signals that can mask the biological expressiveness of a dataset. Moreover, a variety of isotopes and adducts contribute to noisy and complex spectra [41]. Additionally, instrument-specific characteristics like mass accuracy, resolution, and calibration, as well as

different software and experimental designs impede the comparison or integration of datasets. These limitations call for stringent quality control, calibration, and further standardization to improve the general comparability and reproducibility of MS data acquisitions across time points, instruments, and laboratories [42]. Moreover, these technical influences can introduce batch effects and underscore the need for powerful batch integration approaches more than in other omics fields [43].

In particular, MALDI-based imaging MS methods have lower sensitivity than bulk MS protocols or even sequencing experiments. In order to reach a high spatial resolution, the laser ablates only small sample volumes and thus produces a limited set of ions to be detected [44]. Moreover, matrix is deposited ubiquitously on the sample to mitigate the relatively poor ion extraction efficiency of MALDI [45]. However, this increases the risk of detecting mostly matrix ions. As a consequence, the lower sensitivity leads to a greater sparsity in the generated data matrices, which in turn complicates downstream analysis and correction of batch effects [43].

If ions are detected, their quantitative significance is limited by various technical influences. In addition to the instrument variability mentioned above, matrix effects lead to heterogeneous ionization efficiency of metabolite species and mutual ion suppression. The spatial character of MALDI amplifies this within-sample heterogeneity since sample preparation and ablation of the analyzed locations cannot be controlled to be exactly identical. In particular, this includes fixation, matrix coating, and laser focusing [46, 47].

Finally, commonly used MALDI instruments for imaging MS are usually not hyphenated to other separation technologies like LC and do not include tandem approaches (MS/MS). Consequently, the identification of molecular species only relies on accurate mass measurements and downstream metabolite identification fails to distinguish structural isomers and, depending on mass resolution, isobars. [48]

Taken together, datasets generated by MS technology are highly sparse, hence only a fraction of all ions are detected in each cell [49, 50]. Certainly, some of the intensities are missing because metabolites are in fact not present in a given condition or ablated region [44]. These biological zeros are desired and informative about the metabolic state of the assessed biological system. However, the remaining missing values do not reflect the biological state of the sample but are introduced throughout the acquisition process. They are commonly referred to as technical zeros or dropouts and, depending on their source, occur in different patterns [51]: Values are missing completely at random (MCAR) if their absence is truly random and cannot be associated with other observable or unobservable variables. This could be a result of spontaneous instrument malfunction or random variation in data processing (e.g. during image registration or cell segmentation). In contrast, values are missing at random (MAR) if their absence is biased and can be related to other observable variables. Possible reasons include instrument drift over time and variations between conditions or replicates due to deviations in sample preparation or other ambient factors. Also, ion-specific differences in ionization and measurement efficiencies could result in MAR. Finally, values are MNAR if their missingness is biased but related to unobservable variables. This is usually rooted in the

measurement sensitivity bias: Metabolites that occur in the sample at lower concentrations are more likely to fall under the limit of detection (LOD) and not be detected. This leads to a right-skewed distribution of ion intensities. If values below a certain limit do strictly not occur at all, the data is also called left-censored. Technical zeros in a dataset are usually the result of multiple missing patterns and a clear attribution to one source is often not possible [52].

In bulk MS-based proteomics and metabolomics studies, MNAR was identified as the dominant pattern of missingness [53, 44, 54]. Combined with the high proportion of missing values [49, 50], this issue has to be addressed before further downstream analysis. Otherwise, MNAR has the potential to strongly bias statistical tests and lead to misinterpretation of the data [51, 55].

Different strategies have been employed to deal with missing values in MS data: Removing (all) incomplete ions restores a complete data matrix at the cost of losing the entire biological information associated with these features [56]. Also, some analysis methods can cope with a small fraction of missing values [57], but both these strategies are unsuitable for typical metabolomics or lipidomics datasets with higher dropout rates (DRs). Consequently, replacing missing values in the dataset using an appropriate imputation method is in many instances the most promising strategy.

1.4 Available imputation methods

Data sparsity is a common problem in many fields and a host of imputation methods have been developed and compared in the last decades [58, 59, 60]. Some make few assumptions about the data and are thus broadly applicable, others were specifically designed for an application and cannot be readily applied elsewhere. Also, most imputation methods are designed to handle a certain pattern of missingness and are usually not suitable for other patterns.

The simplest imputation methods just replace missing values with static numbers, for instance, the mean of a feature (assuming MCAR) or a low value (assuming MNAR). More elaborate tools use other features or other samples to estimate missing values. For instance, k-nearest neighbors (kNN) imputation replaces missing features in a sample by the mean of the features in nearest neighbor samples [61], assuming there is no bias in the missingness pattern. In contrast, Multivariate imputation by chained equations (MICE) employs regression models to impute missing values in a feature using all other features as predictors. The predictions are updated in an iterative process until the algorithm converges [62]. This enables the method to handle MAR where bias is related to observable variables, and to some extent MNAR.

In addition to these general imputation methods, a variety of techniques have been developed to overcome specific data sparsity challenges in single-cell transcriptomics. Notably, many of the newer tools also address the general technical noise in the data and discern between biological and technical zeros [63]. Thus, instead of replacing all missing values with

suitable estimates, so-called denoising methods replace only suspected technical zeros and modify the values present in the data [64]. Although there are now more voices that advocate for accepting missing values in single-cell RNA-sequencing data instead of performing data imputation [65], MS-based omics technologies could profit greatly from the achievements in this area:

With its publication in 2018, Markov affinity-based graph imputation of cells (MAGIC) was introduced as one of the first denoising methods that borrow information across samples (cells) to modify the overall structure of the data. Being related to kNN imputation, the authors advocate for its broad applicability in other omics technologies. The only underlying assumption is that technical noise occurs at higher frequencies than biological signals and thus can be essentially smoothed out [66].

Driven by the insufficient capabilities of MAGIC and other denoising methods to distinguish biological and technical zeros, adaptively thresholded low-rank approximation (ALRA) imputation was developed. In contrast to other tools, ALRA leaves presumed biological zeros completely unchanged and only imputes presumed technical zeros using low-rank matrix approximation. The relatively simple approach using SVD was also motivated by the need for highly scalable imputation methods for increasingly large datasets [67].

Yet another strategy underlies the development of the deep count autoencoder (DCA) method. It was built on stricter assumptions about the noise that affects RNA-seq count data, modeling it as a negative binomial distribution. Using an autoencoder (AE), the data is also compressed to a lower dimensional space. However, unlike low-rank matrix approximation methods and similar to kNN-based methods the machine learning (ML) model also captures non-linear relationships between features [68].

Along with the varying motivations that led to the development of these specialized denoising methods, they come with slightly different proposed use cases and applications. Most sophisticated methods emphasize their ability to preserve biological information in the data to improve the characterization of heterogeneous cell subpopulations (also called archetype analysis) [66, 68, 69, 70]. Additionally, the recovery of gene-gene relationships for correlation analysis is mentioned frequently [66, 68]. Notably, while the authors of MAGIC warn against overestimating effect sizes when performing differential expression analysis on imputed data, the developers of other methods like DCA and scImpute use this application to prove the impact of their methods [69, 68]. Given their degree of customization, it is not always useful to transfer denoising methods from transcriptomics to metabolomics or lipidomics. Some methods rely on narrow assumptions to model unique molecular identifier (UMI) sequencing counts that do not apply to MS-generated intensities [71, 69]. Others incorporate prior knowledge or data that may not be available in the case of metabolomics or lipidomics [72]. Finally, there are other factors to consider when reusing these methods, such as computational complexity, discontinued maintenance, and the ability to seamlessly integrate the methods into standard analysis workflows.

1.5 Aims of this study

With the novelty of single-cell metabolomics and lipidomics, only a handful of single-cell datasets have been generated using the SpaceM pipeline so far. As a consequence, no systematic comparison of methods for processing and downstream functional analysis has been carried out and no consensus is available. Indeed, a myriad of methods and tools were developed for similar applications in transcriptomics and proteomics research and have the potential to be repurposed. However, in order to ensure the quality and accuracy of the conclusions drawn from the data, the suitability and impact of every processing and analysis method have to be assessed critically and systematically with datasets generated using this new technology.

Like in other lines of research, data complexity and sparsity in metabolomics and lipidomics represent two hurdles to overcome in order to translate measurements into biological insights. Many methods that tackle complexity, like dimensionality reduction and correlation network analysis, require complete data matrices and suffer from excessive missing values. Thus, the particularly sparse and noisy data in the field of single-cell metabolomics and lipidomics calls for suitable and effective methods to minimize this uncertainty and recover the buried patterns of biological information.

Other single-cell omics fields have seen considerable efforts to compare and benchmark imputation methods using different datasets and dimensions of evaluation [59, 64]. In contrast, to my knowledge, no systematic review of data imputation on single-cell metabolomics and lipidomics data has been performed so far. With this work, I investigated the patterns of missingness in available single-cell datasets and employed different imputation methods to remove zeros and technical noise from the data. Using a set of performance metrics, I evaluated the ability of imputation methods to recover missing values accurately and to preserve biological patterns. The results of this benchmark enabled me to formulate suggestions on the usage of data imputation methods for different applications, aiding the general standardization of this field. Furthermore, I developed a processing pipeline that delivers a foundation for future analysis of new datasets and imputation methods. Finally, I further explored correlation-based network inference as a field of application for imputed single-cell lipidomics data with the potential to provide more systematic insights into the metabolism.

Chapter 2

Materials and Methods

2.1 Included datasets

Four datasets were investigated in this work, three of them are lipidomics and one is a metabolomics dataset. Prior to MALDI imaging, the corresponding matrix was applied to samples evenly using a HTX TM-Sprayer (HTXImaging). All datasets were acquired within the Alexandrov group with an AP-SMALDI source (Transmit) connected to a Q-Exactive Plus mass spectrometer (ThermoFisher Scientific). Using SpaceM, ion spectra and light microscopy images were combined to infer single-cell ion intensities. Metabolite annotation was performed with the METASPACE platform [48]. All lipidomics datasets were annotated using a custom database. A tabular overview of the datasets can be found in Table 1.

Seahorse The seahorse dataset profiled primary naive $CD4^+$ T cells under different perturbation conditions. The cultured, unstimulated cells served as controls (NStim). Cells were stimulated using human anti-CD3 anti-CD28 T-Activator Dynabeads (ThermoFisher Scientific) and then cultured for 3 days. A fraction of the cultured cells were incubated with 2-deoxy-D-glucose (2-DG) or Oligomycin for 30 minutes (inhibitors of glycolysis and oxidative phosphorylation, respectively). Cells from all four conditions were analyzed using the SpaceM pipeline. In addition to removing sparse ions and cells, this dataset was also filtered to contain only ions annotated with at least one endogenous metabolite according to Human Metabolome Database (HMDB v4 [73]). The excluded ions were manually classified into exogenous and essential metabolites and the latter were also included in the analysis.

Glioblastoma Human naive glioblastoma cells were modified in the CD95 gene/Fas receptor using lentiviral transduction. In particular, between one and three point mutations were introduced into the transmembrane domain (TMD) of CD95 (single: TMD_sM, double: TMD_dM, triple: TMD_tM), the gene was knocked out completely (CD95_KO), or overexpressed (CD95_WT). The six conditions were analyzed using the SpaceM pipeline.

Pancreatic cancer Four different pancreatic duct adenocarcinoma (PDAC) cell lines, two of an epithelial phenotype (HPAF-II, HPAC) and two of a mesenchymal phenotype (Mia-

Paca2, PSN1), were analyzed without modification or perturbation. This dataset was acquired in three separate batches, each containing all analyzed cell lines.

HepaRG Differentiated HepaRG cells in culture were stimulated with different combinations of oleic and palmitic acid and cytokines for 24 hours. For one condition, cells were only treated with fatty acids (F), for the other two conditions, fatty acids were combined with interleukin17-alpha (IL-17α) (FI) or with both IL-17α and (5-(p-fluorophenyl)-2-ureido)thiophene-3-carboxamide (TPCA1) (FIT). More details can be found in the publication where it was first used [38].

Table 1: Overview of the included datasets. DR: dropout rate, FA: fatty acid, DAN: 1,5-diamino-naphthalin, DHB: 2,5-dihydroxy-benzoic acid

	Samples	**MS**	**Data**
seahorse (Metabolomics)	human CD4+ T cells, perturbations (NStim, Stim, Stim+2-DG, Stim+Oligomycin)	negative mode matrix: DAN	2-4 replicates Raw DR: 72.5%
glioblastoma (Lipidomics)	human glioblastoma cells, gene modifications of CD95: (1-3 point mutations of the TMD domain)	positive mode matrix: DHB	4-6 replicates Raw DR: 80.9%
pancreatic cancer (Lipidomics)	pancreatic cancer cell lines of epithelial and mesenchymal phenotype (HPAC, HPAF and PSN1, MiaPaca2)	positive mode matrix: DHB	3 replicates Raw DR: 86.6%
HepaRG (Lipidomics)	differentiated human hepatic cell line, perturbations (ctrl, FAs, FAs+IL17α, FAs+IL17α+TPCA1)	positive mode matrix: DHB	3-7 replicates Raw DR: 71.1%

2.2 Imputation and denoising methods

Most of the employed imputation methods were presented to tackle missing values in a wide range of fields and data types. Moreover, many methods were designed for datasets with low dimensionality and sample size. In contrast, most denoising methods were developed in recent years for the specific purpose of processing single-cell RNA-sequencing (scRNA-seq) data. Due to their different strategies and assumptions, both imputation and denoising methods can be divided into families. In the following, the employed and analyzed methods are described in brief, more details can be found in the respective original publications.

2.2.1 Simple methods

Imputation with fixed values The simplest and oldest imputation methods replace missing values with a global or feature-specific value, for instance, the mean, median, or half-minimum value of the respective feature.

2.2.2 Regression-based methods

MICE imputation estimates missing values in a feature using a regression model based on all the other features of a dataset, regardless if they also contain missing values. In one step, all missing values in the dataset are imputed that way. This step is repeated for multiple iterations, updating the estimates of missing values using the also updated features until the algorithm converges. In order to provide a measure of uncertainty, the whole iterative process is completed multiple times. MICE imputation is a flexible and widely used method that is thought to handle both MAR and MNAR patterns [62]. A python implementation adapted from the MICE R package is IterativeImputer from scikit-learn. Instead of multiple imputed data matrices, it only generates one result. The function was applied with an iteration limit of 20 and with lower (0) and upper bounds (maximum intensity of respective ion) for imputed values.

2.2.3 kNN/smoothing-based methods

kNNimpute was originally developed for missing value imputation in microarray studies. It constructs a neighborhood graph of cells based on their Euclidean distance in the feature space (disregarding features that are missing in both compared cells). Every missing value of a cell is replaced by the average of the corresponding features of the k nearest neighbors [61]. This method only preserves zeros in the data if all k nearest neighbors have zero counts for the corresponding feature as well. This work uses the implementation from scikit-learn with 1, 3, or 5 neighbors.

MAGIC denoising extends the concept of kNNimpute. It creates a kNN neighborhood graph based on the principal components of the data. Based on this graph it constructs an affinity matrix between cells by applying an adaptive Gaussian kernel. In turn, this affinity matrix is used to create a diffusion operator that can be applied iteratively (number of iterations t) to the data to propagate patterns in the gene expression between cells. That way, cells are increasingly smoothed together with their local neighborhood [66]. MAGIC performs well in many cases as the underlying diffusion process is mostly driven by highly interconnected cells due to their similar biological patterns and disregards rare spurious connections due to noise. MAGIC is available as Python, Matlab, and R implementations. In this study, the method was applied with 100 principal component analysis (PCA) dimensions and 3 neighbors to calculate the neighborhood graph. The diffusion operator was applied with 1, 2, 3, or 5 iterations.

2.2.4 SVD-based methods

SVDimpute acts similarly to MICE in an iterative fashion. Using SVD, it obtains k significant eigenfeatures that represent the most important patterns of the data matrix (k is the rank of the approximation and has to be chosen during application). Initially, missing values are replaced with the sample average. Then, an expectation-maximization (EM) method

iteratively estimates missing values by regressing the corresponding features against the selected k eigenfeatures until the matrix change falls below a threshold [61]. In this work, the `IterativeSVD` implementation from the Python library fancyimpute was employed with a chosen rank of 100 and a lower bound (0) of imputed values.

ALRA denoising also relies on SVD to generate a low-rank approximation of the acquired data matrix. However, the optimal rank k of the approximation is estimated in a data-dependent manner to include finer biological patterns while disregarding noise with even higher frequencies. This low-rank approximation represents the denoised data matrix and ALRA restores presumed biological zeros in the data by setting negative and small positive values to zero using a feature-specific threshold. This is based on an observation of the authors, that biological zeros are distributed symmetrically around zero in the low-rank approximation [67]. The authors provide an R package that was applied with default parameters.

2.2.5 ML-based methods

DCA performs denoising using generative neural networks called AEs. Instead of training the network to reconstruct the original data, DCA defines a custom reconstruction error in the form of a likelihood function for a negative binomial (NB) or zero-inflated negative binomial (ZINB) distribution. The distribution parameters of individual genes are trained unsupervised and the trained means of the distributions are used to generate a denoised data matrix. By the forced compression into a lower-dimensional manifold, DCA combines the information from co-expressed genes and only learns the most important patterns in the data [68]. The authors implemented the algorithm in a Python library. DCA requires integer counts as input, thus discretized raw intensities were supplied to the algorithm, and normalization and transformation were performed afterwards.

2.3 Batch integration methods

Batch integration is an important step in analysis pipelines where subsets of the data were acquired at different points in time or obtained from different sources. These methods can be distinguished by their level of impact: Some methods alter the data in the original feature space and others that only integrate batches in a lower dimensional manifold like PCA or uniform manifold approximation and projection (UMAP) space.

ComBat was introduced as a batch integration method of microarray experiments. It quantifies the differences between batches using a linear model and subsequently minimizes these differences by changing the location and scale (L/S) of the batch-specific subsets of the data. In contrast to other L/S methods, it uses an empirical Bayes approach to estimate data adjustments and was hence found to perform robustly on small sample sizes [74]. ComBat was applied with default parameters.

Batch balanced k-nearest neighbors (BBKNN) is a fast method for the alignment of different batches and common visualization in UMAP space. It is based on the assumption that differences between distinct cell types are larger than differences introduced by batch effects. Accordingly, BBKNN looks for the nearest neighbors of every cell (in PCA space) separately in individual batches and merges them to construct a global batch-balanced neighborhood graph. This can be further used for UMAP visualization, clustering, and lineage analysis [75]. In contrast to the other listed methods, BBKNN does not batch-integrate data matrices. The method was applied with default parameters.

2.4 Evaluation metrics

To evaluate the different data imputation methods, I employed a variety of evaluation methods and quantitative metrics. These can be interpreted as scores and combined to assess and rank imputation strategies. If not indicated otherwise, comparative metrics contrast imputed data with raw data before dropout simulation and imputation.

2.4.1 Metrics of pattern recovery

This set of metrics aims to judge the preservation of information in the imputed data.

Mean square error (MSE) of values To evaluate the general preservation of the data after different imputation methods, the MSE of all intensity values I in the imputed dataset was calculated in comparison to the raw values (both flattened). The simulation of dropouts increased the MSE, imputation methods should recover the missing intensity values and thus reduce the MSE again.

$$MSE = \frac{1}{\#ions \cdot \#cells} \sum_{i=c}^{\#cells} \sum_{i=1}^{\#ions} (I_{i,c,raw} - I_{i,c,imp})^2$$

MSE of variance The variance error score was calculated as MSE of the variances of ion intensities between again imputed and raw data. Simulation of random dropouts increased the variance of the raw data and data imputation ideally mitigates this process.

$$MSE(V) = \frac{1}{\#ions} \sum_{i=1}^{\#ions} (Var(I_{i,raw}) - Var(I_{i,imp}))^2$$

Deviation in ion-ion correlations This metric is based on the mean pairwise Pearson correlation coefficients of all pairs of ions in a dataset. The absolute difference between these values after and before imputation was calculated:

$$\Delta C_{ions} = |\text{mean}(C_{ions,imp}) - \text{mean}(C_{ions,raw})|$$

C_{ions} denotes the correlation matrix between all ions in a dataset. A large increase in ion-ion correlations could indicate that an imputation used strong dimensionality reduction and thereby introduced spurious correlations between ions. Thus, small deviations are favored.

Cell-cell correlations Analogously, the cell-cell correlation score was calculated as the mean of the pairwise Pearson correlation coefficients of all pairs of cells. For large datasets with $>$ 10k cells, a randomly sampled subset of this size was used for this calculation to reduce the computational burden. Sampling was stratified by condition to ensure adequate representation of cell types/conditions. Especially in datasets with very high DRs, cell-cell correlations were decreased. Consequently, this measure is a good indicator of the ability of an imputation method to recover missing data to realign cells with the dataset.

2.4.2 Metrics of group separation

The set of group separation metrics evaluated the performance of imputation methods at preserving or even improving the population structure of the analyzed datasets. This is an indication of their ability to transfer the information required for imputation from the relevant group of cells. Increasing DRs led to a successive loss of the population structure of datasets. Imputation methods were judged on their ability to recover this population structure.

kMeans clustering overlap kMeans clustering is an unsupervised machine learning algorithm that clusters samples based on their high-dimensional data into k distinct partitions. In this work, given the number of conditions in a dataset, the algorithm clustered cells with similar ion intensity profiles into groups. If the different conditions or cell types could be separated based on their distinct ion profiles, the data-driven clustering would largely agree with the given condition assignment. This overlap between kMeans clustering and dataset conditions was quantified by the adjusted rand-index (ARI). The metric takes values between -1 and $+1$, where $+1$ indicates a perfect agreement and 0 an agreement by chance. This metric was adapted from Mongia et al. [70]

Silhouette score The silhouette score measures the separation of named clusters. It was calculated per sample (cell) using its intra-cluster distances and the mean distance to the nearest different cluster. The overall silhouette score is the mean of all sample silhouette scores. The score also takes values between -1 and $+1$. Values close to $+1$ symbolize a perfect separation and values close to -1 a very poor group separation [76]. The silhouette score was calculated on cells in a 2-dimensional UMAP embedding with their condition metadata as cluster labels.

Calinski-Harabasz score The Calinski-Harabasz score, also known as the Variance Ratio Criterion, is another metric used to evaluate the quality of clustering results. It measured the ratio of between-cluster dispersions to within-cluster dispersions. The Calinski-Harabasz score ranges from a minimum of 0 to an arbitrarily large value and a higher score indicates better separation and compactness of clusters [77]. In this work, the Calinski-Harabasz score was calculated in the original feature space and using the conditions as cluster labels.

Davies-Bouldin score The Davies-Bouldin score was used to assess clustering quality as well. It was calculated as a ratio between cluster compactness and distance between clusters.

Lower values of the Davies-Bouldin score indicate better separation and distinctiveness between clusters [78]. It was calculated in the original feature space with conditions as cluster labels.

2.4.3 Metrics of biological information recovery

Although the metrics of group separation already give an idea of the imputation methods' abilities to recover biological information, their performance is also evaluated using other characteristics.

Comparison of DAMs In analogy to differentially expressed genes in transcriptomics, DAMs are metabolites that exhibit a significantly higher or lower abundance in one condition compared to another or the rest of the conditions (`sc.tl.rank_genes_groups`). DAMs were identified based on the log-transformed and normalized ion intensities, as imputation was not performed on raw intensities. Wilcoxon rank-sum test was used as statistical test and ions with a fold change $FC \geq 2$ or $FC \leq -2$ and an adjusted $P \leq 0.05$ (Benjamini-Hochberg) were selected as DAMs. For the preservation analysis of log-fold change ranks, ions were ranked from highest log-fold change to lowest and compared to the raw data using Kendall's rank correlation. For the analysis of the top 20 DAMs, the 20 ions with the respective highest or lowest fold change (up- and down-regulated) were selected for the baseline data and all simulated and imputed data matrices. The recovery was calculated as the fraction of the various sets from different methods, that overlapped with the set from the baseline data.

Correlation analysis Fully connected and weighted ion correlation networks were generated based on Pearson correlation and Spearman's rank correlation. To facilitate the interpretation of network edge weights, the obtained correlation coefficients were scaled from the interval of $[-1; 1]$ to $[0; 1]$. Using the linex2 tool[79], a theoretical biological reaction network was established based on all ions detected in the dataset. Based on this prior knowledge network, edges in the correlation network were classified as whether a biological reaction exists between two ions or not. The distributions of weights for edges with/without biological evidence were compared using a two-sided Wilcoxon rank sum test. P-values were adjusted for multiple testing using the Benjamini-Hochberg method. The results of the Wilcoxon test were visualized in a volcano plot where edge weight fold changes corresponded to the fold change between the average weights of edges with/without biological reaction.

2.4.4 Combination of imputation metrics

To determine the summarized imputation scores "pattern recovery" and "cluster separation" for the individual datasets, the listed metrics from sections 2.4.1 and 2.4.2 were combined in the following way: Metrics that should be minimized, namely MSE of values, MSE of variance, deviation in ion-ion correlations, and Davies-Bouldin score, were inverted (multiplication with -1). Afterward, for every dataset, the individual metrics were scaled across imputation methods and dropout rates to an interval of $[0; 1]$ to remove the impact of different scales

while preserving differences between methods and levels of sparsity. Finally, the combined scores were derived by taking the mean of the respective four scaled metrics. This process is summarized in Table 2. In order to generate the 2x2 performance matrices, the datasets were imputed and evaluated in 5 distinct replicates. The error bars represent the standard deviations between these replicates.

Table 2: Overview of the evaluation metrics and their integration to scores. Not all metrics were designed such that higher values correspond to better performance. The four metrics, where lower values were considered better, were inverted. Then, all metrics were min-max scaled to $[0; 1]$ and averaged to yield the corresponding summarized scores. MSE: mean square error, UMAP: uniform manifold approximation and projection

Score	Metrics
information recovery	**MSE of values**: Converted as a vector, all intensities of a dataset after imputation were compared to raw data using MSE. **inverted** **MSE of variance**: The variance for each ion in a dataset was calculated and the variances of the imputed data were compared to raw data using MSE. **inverted** **Cell-cell correlations**: Pearson correlations were calculated for all pairs of cells and averaged. **Deviation in ion-ion correlations**: Pearson correlations were calculated for all pairs of ions and averaged, the absolute difference between this value for raw and imputed intensities represents the deviation. **inverted**
cluster separation	**ARI of kMeans**: In feature space, cells were clustered unsupervised, and the cluster labels were compared to condition metadata. **Silhouette score**: In 2D UMAP space, the score was calculated per cell using the within-cluster and adjacent-cluster distance and averaged over all cells. **Calinski-Harabasz score**: In feature space, scores were determined as a ratio of between-cluster dispersion and within-cluster dispersion for each group according to metadata. **Davies-Bouldin score**: In feature space, scores were determined as a ratio of between-cluster compactness and between-cluster distance for each group according to metadata. **inverted**

2.5 Implementation

General data processing was done using the scanpy package [80]. This included filtering of ions and cells, normalization to a fixed count of 10^4, dimensionality reduction to PCA and UMAP space, and calculation of highly variable and differentially abundant metabolites. The underlying approaches and implementations of the imputation and denoising methods

are listed in Table 3. MSE, kMeans clustering, ARI, silhouette score, Davies-Bouldin score and Calinski-Harabasz score were implemented using the scikit-learn library [81]. Application of imputation methods, dimensionality reduction, and imputation evaluation, as well as calculation of correlation matrices for network inference, were performed on the EMBL IT services high performance computing (HPC) resources. A Python library for further evaluation of imputation and denoising methods or datasets is available from GitHub (mariusrklein/sc_imputation_denoising).

Table 3: Overview of the imputation/denoising implementations. The table summarizes all imputation approaches and the representative methods employed in this work. The respective implementations point to GitHub repository paths or specific functions within them. kNN: k-nearest neighbors, MICE: multivariate imputation by chained equations, SVD: singular value decomposition, ML: machine learning

approach	**methods**	**implementation**
fixed value imputation	ctrl_mean, ctrl_median, ctrl_random	iskandr/fancyimpute.SimpleFill
kNN imputation	knn_3, knn_5	sklearn.kNNImputer
MICE imputation	fancy_multi	sklearn.IterativeImputer
SVD imputation	fancy_itersvd, fancy_soft	fancyimpute.IterativeSVD, fancyimpute.SoftImpute
SVD denoising	ALRA	KlugerLab/ALRA
kNN denoising	MAGIC_t1, MAGIC_t3, ...	KrishnaswamyLab/MAGIC
ML denoising	dca_nb, dca_zinb	theislab/dca

Chapter 3

Results

Not many datasets have been generated using SpaceM so far. I have selected four of them to carry out all the following investigations. All datasets comprised human cells in culture. The only metabolomics dataset (called seahorse) analyzed primary T cells under different stimulation and perturbation conditions. It was complemented by three lipidomics datasets, entailing glioblastoma, PDAC (pancreatic cancer), and healthy hepatocyte (HepaRG) cell lines. While the HepaRG cells were also treated with different compounds, the PDAC experiment comprised different commercially available cell lines that reflected different states of epithelial-mesenchymal transition (EMT). Finally, the glioblastoma cells received different genetic modifications of the CD95 gene.

3.1 Ion filtering beats cell filtering in reducing data sparsity

To be able to evaluate the performance of data imputation methods against a ground truth and on multiple levels of sparsity, I aimed to introduce artificial dropouts into the SpaceM datasets. However, the available datasets already had raw DRs between 71.1% and 86.6%, which left only a small margin to simulate dropouts on top. In other omics fields, an accepted method to reduce data sparsity is the removal of ions and cells with high fractions of missing values. In other words, only cells with a certain number of ions measured and only ions that are present in a certain number of cells, are kept for the analysis. This should apply to single-cell metabolomics and lipidomics as well. Accordingly, I systematically examined the impact of different filtering thresholds on the overall structure of the data. To enable between-dataset comparisons, I used a fixed set of fractions (5%, 10%, 20%, 30%) to filter cells and ions. For instance in the seahorse dataset, filtering out cells with less than 5% unique ions would remove all cells that have less than 48 features, and for ions, it would remove ions that are present in less than 495 cells (see Figure 1a and 1b). Notably, cells and ions had very different distributions of sparsity: While very few cells had a very high or a very low number of features, the majority of ions were present in either very few cells or almost all cells. As a consequence, a low threshold of 5% reduced the sparsity of the seahorse dataset drastically to 43.4% when applied to remove ions. In contrast, no cell in the dataset had less than 5% unique ions present, such that this filter did not remove cells or affect

the dropout rate. A closer examination of the distribution of unique ions per cell revealed that cells from different conditions exhibited varying numbers of features. For instance, cells measured in the 2DG condition counted 299 unique ions on average while non-stimulated cells only had 247. This entails the risk of unequally diminishing cell populations from different conditions when filtering thresholds are set arbitrarily. An example of this can be observed in Figure 1c: Two-dimensional UMAP embeddings of the seahorse dataset are shown after applying different filtering thresholds on ions and cells. The different columns correspond to the removal of ions and rows to the exclusion of cells. Whereas removing ions had very little impact on the population structure of the dataset, the filtering threshold of 30% unique ions removed almost all cells from the non-stimulated condition.

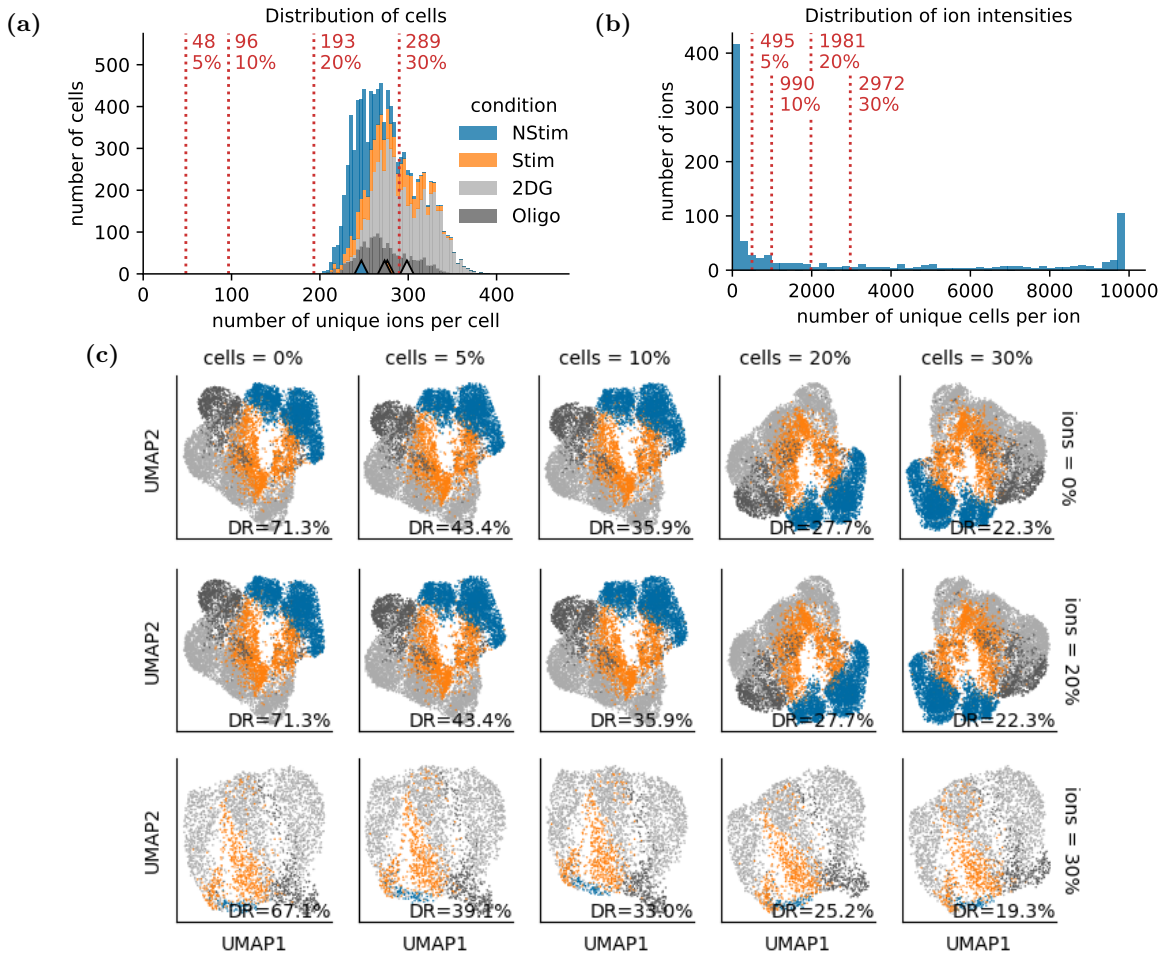

Figure 1: Systematic filtering on the seahorse dataset. Sparsity in single-cell data is commonly reduced by removing sparse ions and cells. **(a)** Systematic thresholds removed cells with fewer unique ions than specified with red lines and text (absolute number and fraction). Cells are colored by their affiliation to conditions. Colored triangles at the bottom indicate the respective mean number of unique ions per condition. **(b)** Analogous, ions that were detected in fewer cells than specified in different thresholds, were removed. **(c)** Population structure and dropout rate (DR) of the dataset after different filtering thresholds. Rows correspond to (a), and columns to (b). Ion thresholds 5% and 10% exhibited no differences to 0%, so they are not shown. Cells are colored according to conditions as in (a). The meanings of the individual conditions are explained in the methods section.

Similar observations were made for the other included datasets: For all cell filtering thresholds that excluded a considerable portion of cells, the glioblastoma dataset showed a progressive loss of the population structure (Supplementary Figure S.1). In contrast, the removal of ions hardly affected the UMAP embeddings. A closer comparison of the respective cell distributions (panel (a)) revealed that all datasets exhibit different unique ion counts (Supplementary Figures S.1, S.2 and S.3). Additionally, the minimum unique ion count differed greatly between datasets (seahorse: 203, glioblastoma: 5, pancreatic cancer: 22, HepaRG: 10) which indicates that the seahorse dataset had been filtered before.

Given these systematic insights into the effects of filtering on the overall structure of the data, reasonable thresholds can be chosen to retain the biological complexity of the datasets but reduce data DRs to a range of 36% to 63%. The chosen cutoffs are detailed in Supplementary Table S.1. For three of the datasets, filtering reduced the number of features to around 350 features which seems to be sufficient to maintain the cellular population structure.

3.2 Dropouts in single-cell metabolomics/lipidomics data are driven by MNAR

Imputation and denoising methods are usually only suitable for a certain pattern of missingness in the data. To guide an informed selection, I had to identify the relevant missing pattern first. This is challenging, as usually a combination of causes leads to technical zeros and often a mixture of missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) is observed. MNAR results from a relationship between the value of missing measurements and the fact that they are missing (visualized in Supplementary Figure S.4). Under the assumption that the majority of ions show a relatively stable abundance in the present biological system, the missing values of an ion can be roughly approximated with the average intensities of the remaining, non-missing values. This is an accepted simplification within the community [44, 82]. Across datasets, the mean logarithmic ion intensities exhibited a negative relationship with the fraction of missing values of the corresponding ions (Figure 2a), This indicates that MNAR is present as a driver of missingness in all datasets. Notably, the vast majority of ions with a mean intensity above 10^7 had very few missing values. In this range, intensity-dependent MNAR had no more influence, so only MCAR and MAR could lead to missing values among these ions. Especially the glioblastoma dataset exhibited few missing values in this high range, suggesting it was hardly affected by MCAR or MAR. In contrast, the seahorse and even more prominently the HepaRG dataset showed a larger fraction of ions scattered on the right side of the bulk of features (red asterisks). These ions with higher mean intensities and large fractions of missing values can be attributed to MCAR or MAR. Compared to the other datasets, pancreatic cancer showed generally a less prominent association between the fraction of missing values and ion mean intensities. Moreover, it did not contain a fraction of ions with high mean intensity and very few dropouts. This could indicate that in this case the effects of MNAR and MAR were overlaid.

The corresponding distributions of mean log intensities across datasets are shown in Figure 2b. These distributions are consistently right-skewed with particularly long right tails in the case of the seahorse and HepaRG datasets. Furthermore, the lower limit of mean intensities lies at 10^4. Because the presented intensities were processed by SpaceM, they have undergone a scaling step to convert pixel intensities to cell intensities. Hence, this limit does not directly translate to the LOD, but it serves as a clear indication that low-intensity ions were lost during data acquisition.

These initial observations already uncovered that although MNAR was the overall driving pattern of missingness in the given single-cell MS-based studies, every dataset has faced individual influences that shape the structure of their sparsity. This further underscores the need for a systematic application and comparison of imputation approaches, as there is no perfect solution for all single-cell metabolomics and lipidomics experiments.

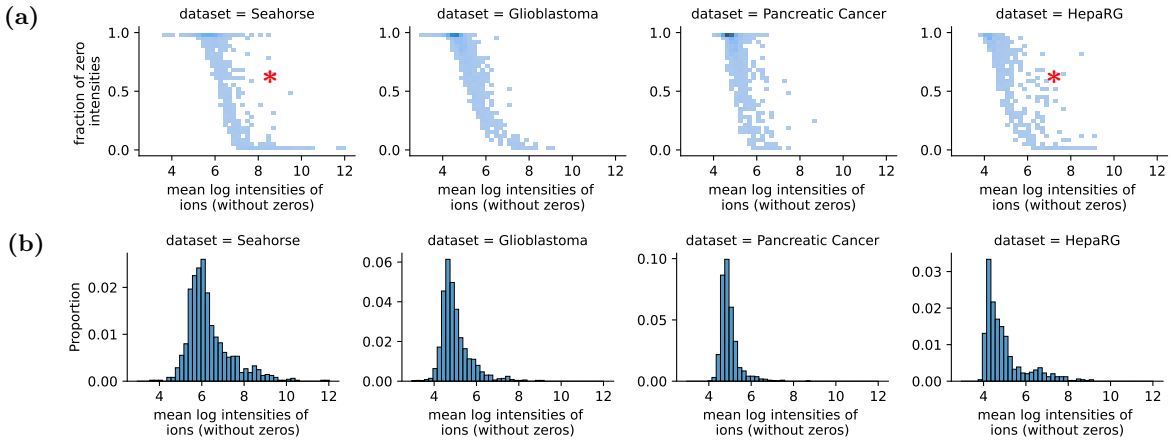

Figure 2: Relationship between ion intensities and dropout rates. (a) For all four datasets, there is a pronounced negative relationship between the fraction of missing values of ions and the mean logarithmic intensities of the remaining values, hinting at the presence of missing not at random (MNAR). Intensities were not processed otherwise, missing values were returned by SpaceM as zero intensities. Red asterisks mark ions with both high mean intensity and fraction of missing values, likely driven by missing at random (MAR) or missing completely at random (MCAR) **(b)** Corresponding histograms of mean logarithmic ion intensities for the four analyzed datasets show right-skewed distributions and a clear lower limit at 10^4.

3.3 Simulated dropouts serve as ground truth for imputation benchmark

To obtain a ground truth to test the performance and robustness of data imputation and denoising methods, I developed a method to introduce artificial dropouts at specific rates into the already sparse SpaceM datasets. This is a challenge, as the sets of existing missing values and simulated missing values usually overlap. Thus, the effective number of simulated dropouts is usually lower than expected. To overcome this, I created a score-based simulation program that can model both MCAR and MNAR.

To simulate simpler MCAR, all intensities of a dataset were assigned a removal score in the form of a random number from the interval $[0; 1]$. Originally missing intensities were assigned a zero. To simulate a DR of 65%, all intensities with corresponding random values below the 65th percentile of the set of random values were removed. Analogous, a MNAR pattern was simulated by combining the random values with a deterministic part, for example, a sigmoid function of the respective metabolite intensities. That way, lower intensities have a greater probability to be removed than higher intensities. Random and deterministic parts of the removal score can be varied (1:1, 1:5, etc.) to cater to dataset-specific differences. A graphical visualization of the dropout simulation is given in Supplementary Figure S.5.

The DRs of the different datasets were largely stable with a coefficient of variation (CV) ranging between $1.2\% - 4.8\%$ and $3.0\% - 5.9\%$ across conditions and replicates, respectively. This low level of variation within the datasets allowed a global simulation of dropouts across conditions and replicates. Taking into consideration the finding that data sparsity was largely driven by measurement sensitivity bias, I decided to simulate all dropouts using the MNAR method (deterministic part vs. random 2:1). On this corrupted data, I applied different imputation and denoising methods to evaluate them against the original intensities before simulation. Of note, all methods were only supplied with normalized and log-transformed or with raw intensities, but not with cell or ion identifiers or any other metadata.

For an initial assessment of the general impact of different methods, I inspected visually the UMAP embeddings of the corrupted and subsequently imputed data. In this regard, the glioblastoma dataset serves as an example to show the effects of four of the applied methods (Figure 3): Simple fixed value imputation with the respective ion means (ctrl_mean), SVD imputation (fancy_itersvd) as well as denoising with MAGIC and DCA (MAGIC_t3, dca_nb). The original data (no simulated dropouts, no imputation, DR=63%) enabled a clear separation of the conditions in UMAP space. This population structure was still captured with 79% biased missing values, but at 94%, the cell clusters fell apart into four mixed-condition groups (first column). Based on their approaches, the shown imputation and denoising methods had very different impacts on the data structure. At the intermediate sparsity level, they all resembled the structure of the original data. However, at the highest dropout rate, the simpler mean and SVD imputation seemed to perform better at recovering the structure from the original data than the sophisticated denoising methods. Instead, the latter also introduced new local structures in the UMAP embedding, indicating that denoising involves more complex data transformations.

Despite their expressiveness, these visual observations are less suitable for a high-level comprehensive comparison. To enable that, I complemented the qualitative visual depictions with a set of performance metrics that evaluated the population structure of differently imputed data quantitatively. These metrics assess the colocalization of unsupervised clustering with the given condition labels (ARI of kMeans), measure the dispersion (Calinski-Harabasz) and compactness (Davies-Bouldin) of the different cell populations of the raw/imputed data in the original feature space. Only the Silhouette score that measures the intra- and inter-cluster distances relies on the dimensionality reduction of UMAP. To facilitate comparison

and later summarization of the metrics, the only metric that considered lower values as better, namely the Davies-Bouldin score was inverted. That way, for all metrics, higher values corresponded to better performance. The resulting performance charts for the glioblastoma dataset are depicted in Figure 4. Notably, denoising methods reached high values at low simulated dropout rates, but their performance deteriorated with increasing sparsity. In contrast, imputation methods performed more consistently throughout the range of simulated dropout rates. Although observable in all four metrics, this became especially obvious for the Silhouette coefficient, where imputation methods achieved constant scores while denoising methods performed better at low levels of sparsity, but fell behind imputation at high levels (Figure 4b). This confirmed the observed population patterns in the UMAP embedding. Additionally, the visual impression of a highly accurate structure recovery after SVD imputation (lower center panel) was clearly reflected in the performance metrics.

Taken together, evaluating the performance of imputation and denoising from this global perspective on the data revealed strong differences between the corresponding underlying mechanisms of the tested methods. It also provided a first impression of the ability of different methods to remove technical variance in order to preserve biologically similar cells.

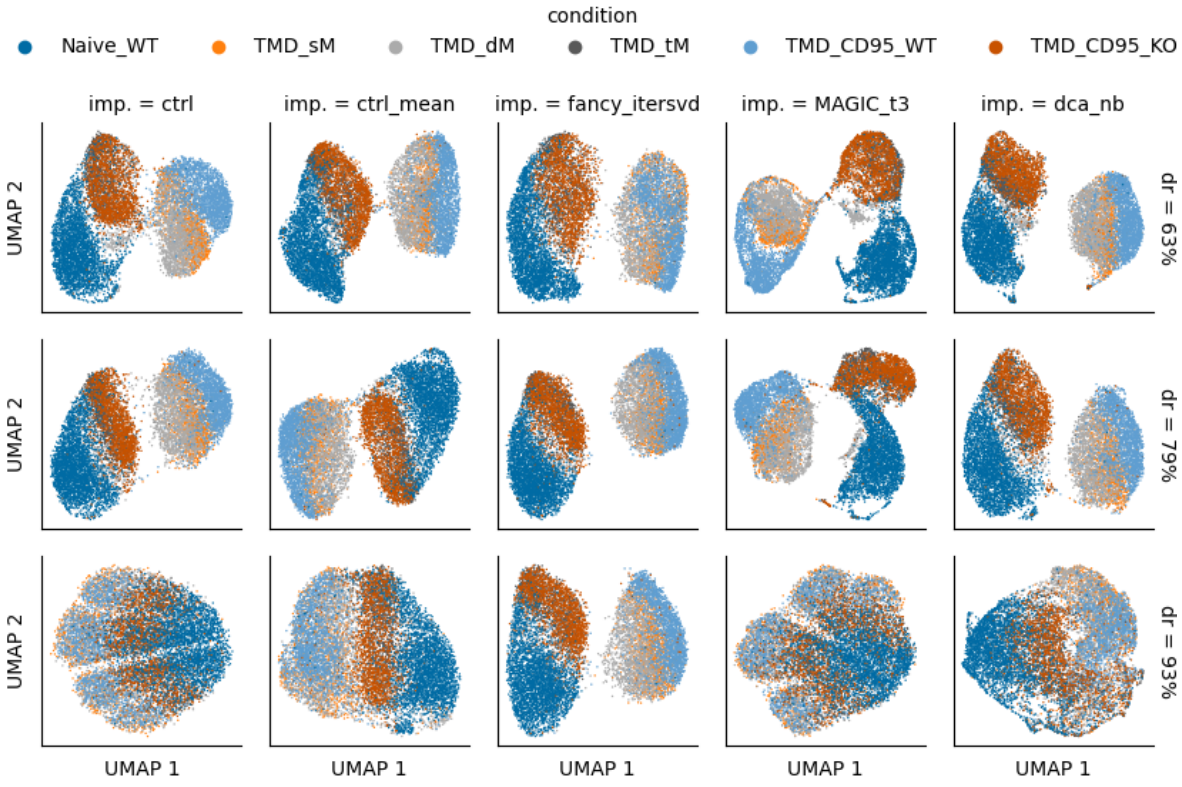

Figure 3: Recovery of population structure with imputation (glioblastoma). After simulating different levels of missing values in an missing not at random (MNAR) pattern, the dataset was processed with a variety of imputation/denoising methods. Shown is a UMAP embedding of the dataset at different levels of missing values (rows: 63% is prior to simulation) and after applying a selection of imputation/denoising methods (columns: ctrl is not imputed). With an increasing dropout rate (dr), the separation of conditions is lost. The methods abilities in recovering this structure varies greatly. The meaning of the individual conditions is explained in the methods section.

3.4 Denoising strengthens consensus within conditions

For applications that directly use the data matrices, not only does the overall structure of an imputed dataset matter but also the way and extent individual values are affected and how this could influence downstream analyses. Naturally, imputation and denoising change the supplied data in very different ways. To compare this visually, I plotted the intensities of an ion before and after imputation/denoising against each other. Taking the glutamine ion from the seahorse dataset (without prior simulation of dropouts) as an example, kNN imputation replaced missing values of this ion using information from similar ions and left present values unchanged. Thus, in a direct comparison of before and after imputation, the unchanged intensities laid on the identity line while the imputed values were found on the x-axis intercept (see Figure 5a, left). This relationship is more interesting for the various denoising methods that also modified the present intensities in the data. For some ions, including glutamine, denoising methods altered the intensities differently across conditions. For instance, the MAGIC method decreased the glutamine intensities in non-stimulated cells strongly compared to cells that were treated with 2DG. On the contrary, glutamine intensities

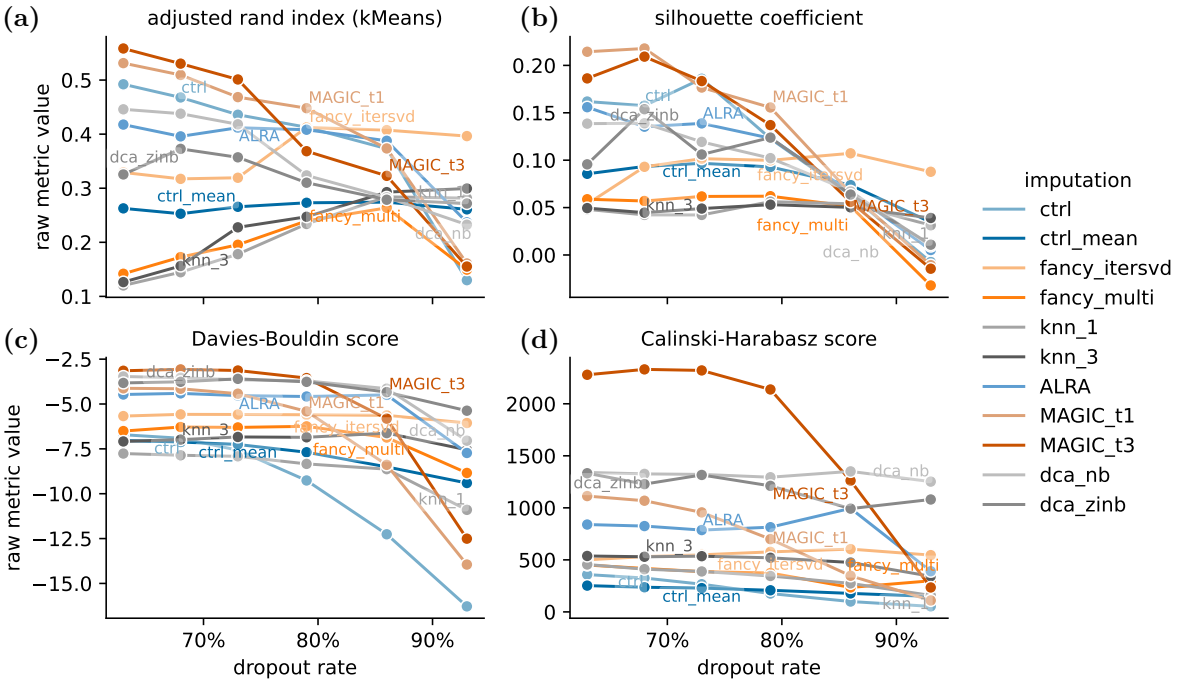

Figure 4: Metrics of population structure recovery with imputation (glioblastoma). Different performance metrics summarize the visual perception of cluster separation after simulating different levels of missing values (missing not at random) and subsequent imputation/denoising using different methods. The `ctrl` curve represents a baseline of only simulation. **(a)** The adjusted rand-index of kMeans clustering measures the agreement of unsupervised clustering with given condition labels and **(b-d)** the three other metrics measure cluster separation from different perspectives. The Davies-Bouldin score (c) favors small scores, so its values are inverted for easier comparison of the metrics.

of many 2DG-treated cells with previously low intensities were increased by MAGIC. DCA and ALRA presented similar patterns, the latter even increased intensities of 2DG-treated cells regardless of their previous value. This separation of cell intensities was observable for conditions, but not on the level of individual wells or replicates (Supplementary Figure S.6a). Hence, the process could be related to biological variability rather than technical effects. Moreover, not only cells with values present were separated, but the denoising methods also replaced zeros in cells from different conditions with values from distinct distributions (Supplementary Figure S.6b). In contrast, imputation methods using kNN, MICE, or SVD replaced zeros with similar values across conditions.

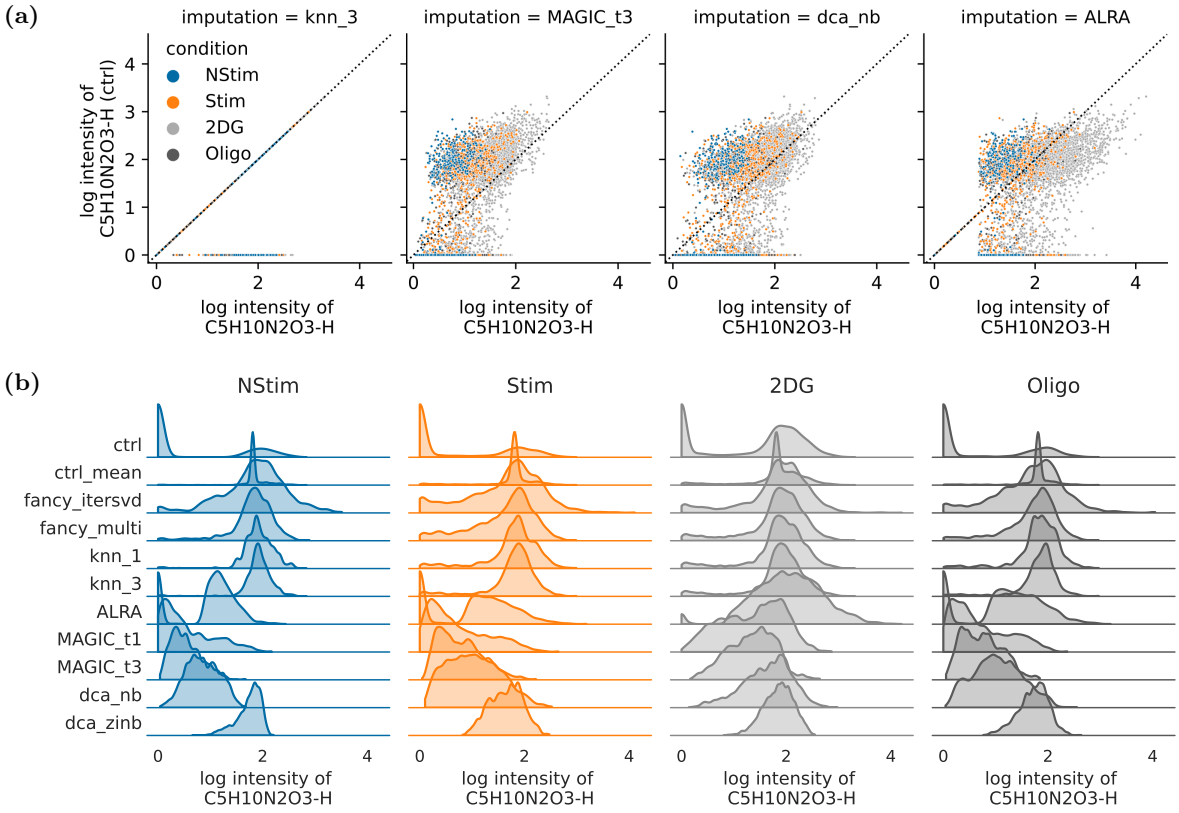

Figure 5: Effects of imputation methods on ion distributions (seahorse). Imputation and denoising methods affect the present data in different ways. **(a)** Direct comparison of raw/ctrl and imputed intensities of the ion $C_5H_{10}N_2O_3$ (Glutamine). `kNN_3` imputation (panel 1) only affects zeros and leaves non-zero values unchanged (points lie only on the dotted identity line and on the x-axis). However, denoising methods (panels 2-4) also modify non-zero values. The effects of these modifications on intensities differ across conditions. Non-stimulated cells are decreased the strongest by denoising while oligomycin-treated are affected the least. **(c)** The intensity distributions of $C_5H_{10}N_2O_3$ under different imputation/denoising methods are compared. Imputation methods (up to `knn_3`) center the intensities around the mean of non-zero values. In contrast, most denoising methods (down from `ALRA`) shift distributions down, taking the fraction of zero intensities into account.

Analogous patterns can be observed for various other ions of the seahorse dataset (see Supplementary Figure S.7). This effect of focussing intensities within and separating them across conditions suggests that the investigated denoising methods are successful in borrowing

the right information across cells to replace zeros and remove technical noise. However, an important part of this are varying fractions of missing values across conditions. They seem to have a rather large impact on denoising in comparison to imputation: When plotting the glutamine intensity distributions across conditions and imputation and denoising methods, the 2DG condition has considerably fewer missing values (26.4%) in the raw data than the other conditions (Stim: 57.8%, Oligo: 62.4%, NStim: 68.9%, Figure 5b). Obviously, all imputation methods (rows up to `knn_3`) replaced zero intensities upwards to approximate the non-missing values. In contrast, most denoising methods shifted both the present and missing values towards each other. In line with the different proportions of present and missing values across conditions, intensities from cells treated with 2DG experienced a weaker decrease than cells from the other conditions. Thus, the given denoising methods seem to incorporate the information on missingness to some extent into the returned data.

These observations raised the question of which imputation or denoising approach best approximates the original intensities and how the methods relate to each other. To this end, the matrices of raw and imputed/denoised data were compared as vectors using Pearson correlation and cosine similarity (Supplementary Figure S.8). In hierarchical clustering based on both metrics, imputation and denoising methods are separated into two distinct groups (except `dca_zinb` is assigned to imputation). Despite the severe impact of denoising methods, the original data matrix clustered together with the denoised data and showed the highest similarity to the adjusted intensities generated by MAGIC and DCA.

3.5 2x2 matrix captures performance of imputation/denoising

While this general comparison of imputation and denoising methods gives an idea about their ability to preserve information from the raw data, it does not provide insights into their performance at recovering information that was lost due to technical noise and missing values. To investigate this recovery further, I developed a set of information recovery metrics and employed them to compare how the artificially corrupted and subsequently imputed/denoised data resembled the ground truth intensities. In combination with the cluster separation metrics from Figure 4, they enable a clear performance ranking of the employed methods. The used metrics tackle different aspects of the structure of the data. To facilitate comparisons, the metrics based on MSE were inverted such that higher values correspond to better performance consistently.

The MSE of all intensity values provides a general quantitative measure of the methods' abilities to restore intensities that were removed artificially from the data. For the glioblastoma dataset, the MSE showed a clear advantage of denoising over most imputation methods (Figure 6a). Only SVD imputation (`fancy_itersvd`) restored intensities as well as the investigated denoising methods, especially at higher simulated sparsity. Besides the two ML-based denoising methods, it stands out as the only method that reduced the error when confronted with more missing values. In contrast, the other imputation methods exhibited far greater errors across sparsity levels than the denoising methods and especially than no imputation.

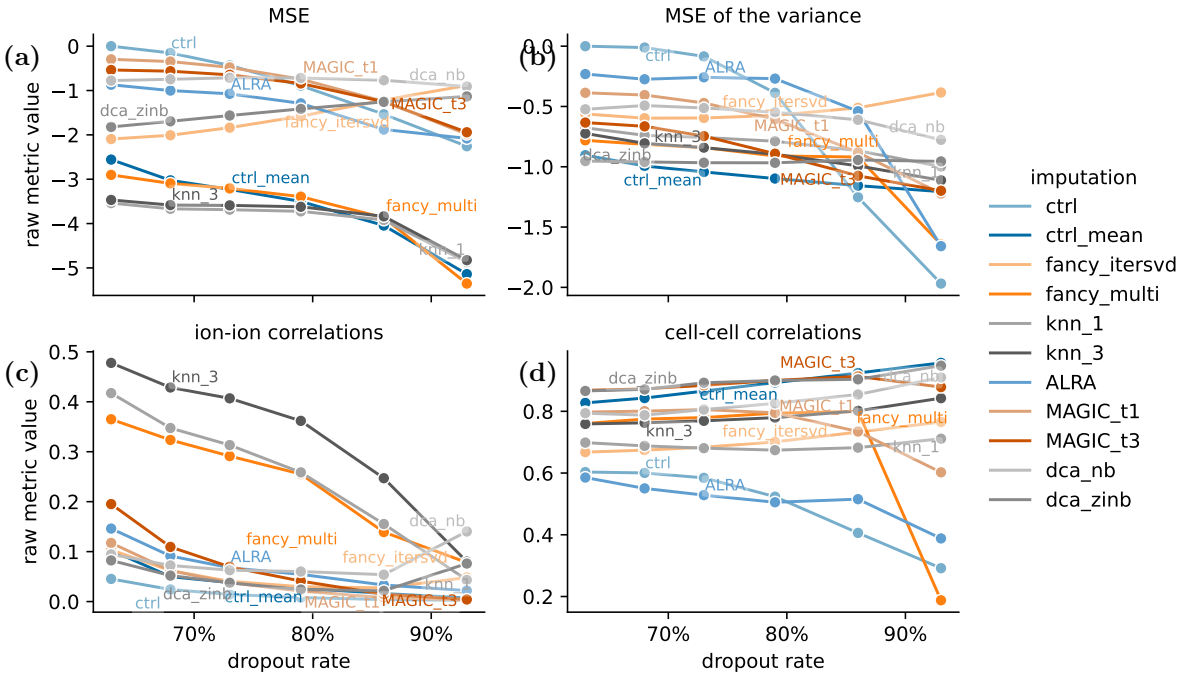

Figure 6: Information recovery of imputation/denoising methods (glioblastoma). These performance metrics show the information recovery that methods achieved with the glioblastoma dataset. To this end, data with increasing dropout rates was simulated (missing not at random), imputed using different methods, and compared to the original intensities before simulation. The `ctrl` curve represents the baseline of only simulation without imputation. To facilitate the visual comparison, the metrics based on mean square error (MSE) were inverted. Thus, higher values of these metrics translate to a favorable, lower error. **(a)** MSE measures the deviation in the individual intensities and **(b)** MSE of variance the change in ion variability. The correlation-based scores average the pairwise **(c)** ion and **(d)** cell correlations to quantify the linear dependence in the imputed data matrices.

In practice, this means that extensive replacement of previously present intensities with zero represented a less serious deviation of the original data than their replacement with imputed value. This could be explained by the simulation using MNAR: Very low intensities were more likely to be replaced by zero during simulation, creating smaller deviations than simple imputation methods that replaced these dropouts with values close to the feature mean.

The MSE of the variance measures how the simulation and replenishment of missing values affected the variability in the ions present (Figure 6b). Like individual intensity error, the variance change was aimed at the smallest possible values. A reduction in variance can indicate unwanted uniformity in the data, as produced by fixed value imputation. On the other hand, an increased variance could hint towards artificial bias introduced to individual ions or distorted relationships between them. Thus, both negative and positive deviations from the variances in the original data were measured by this metric. In contrast to the successively corrupted `ctrl` data, almost all imputation and denoising methods maintained a stable variance. Only ALRA and MICE exhibited a sudden increase in the variance error towards high levels of sparsity.

Finally, ion-ion and cell-cell correlations give an indication about the presence of linear relationships in the data. They were determined as the average of all pairwise Pearson correlation coefficients between features and samples, respectively. High cell-cell correlations were expected in a biological dataset, as they indicate a general similarity of the cellular metabolic profiles. Apart from ALRA, most imputation and denoising methods increased the already high cell-cell correlations of the non-imputed data consistently across levels of sparsity (Figure 6d).

Compared to cell-cell interactions, ion-ion correlations should not deviate greatly from the original data: Many imputation and denoising methods include some step of dimensionality reduction, essentially reducing the number of linearly independent features in a dataset. This excessive compression of the data bears the risk of introducing artificial positive relationships between features. In the glioblastoma dataset, a great increase in ion-ion correlations could be observed for the kNN and MICE imputation at lower simulated dropout rates (Figure 6c). Moreover, ALRA (SVD denoising) showed this behavior for higher dropout rates in the seahorse dataset (Supplementary Figure S.9). To punish imputation or denoising methods that exhibit this phenomenon, the metric was modified in further usage to report the absolute deviation from the control ion-ion correlation score. Ideally, small deviations serve as an indication of retained linear independence in the data. Using these metrics, the various imputation and denoising methods could be clearly distinguished with respect to information recovery and robustness. Moreover, individual desirable or concerning characteristics could be identified.

In order to enable a quick comparison of the employed tools using all available evaluation metrics, I summarized them into two general metrics: The four metrics from Figure 4 were scaled and averaged to serve as a measure of cluster separation and the four metrics from Figure 6 to quantify information recovery (ion-ion correlations adapted as described above). That way, the methods' performances could be compared visually along two dimensions in 2x2 matrix plots. This type of visualization enabled multiple observations: Since imputation and denoising methods were applied and evaluated at multiple dropout rates, their respective performance could be compared at different levels of sparsity. Moreover, the development of their performances across levels of sparsity served as an indicator of their robustness. Figure 7 illustrates this for the glioblastoma dataset. Apart from SVD imputation, the imputation methods in the top row failed to recover much information from the original data and showed a less concise population structure. However, except MICE imputation, they exhibited strong robustness across levels of sparsity (points are a lot closer together than the non-imputed control). In contrast, the included denoising methods in the bottom row appeared to be less robust, especially MAGIC and DCA concerning cluster separation and ALRA at higher sparsity in both dimensions. Thus, if the simulated sparsity was confined to a low level, MAGIC and DCA performed well at retaining information and population structure, but their advantage was lost at the highest simulated dropout rates. Here, SVD imputation stood out again for resembling the original data the closest along both dimensions at high levels of sparsity.

Most of these observations could be retraced with the other included datasets. In particular, the SVD imputation approach proved robust across all datasets and surprisingly even showed some improvements with higher dropout rates. However, there were some differences between the datasets that could be associated with their varying baseline dropout rates. For instance, the proportional change of the control data with higher sparsity could be associated with the baseline dropout rate of the respective dataset. At a lower original sparsity, the performance loss for the seahorse dataset was dominated by information recovery on the x-axis (seahorse has flatter control curve, Supplementary Figure S.10a) while the datasets with higher initial dropout rates rather lost performance with respect to cluster separation on the y axis (pancreatic cancer has steeper control curve, Supplementary Figure S.10c). Additionally, the robustness of DCA with regards to cluster separation increased for higher baseline dropout rates (Supplementary Figure S.10). Both observations are likely rooted in the varying sparsity intervals of the different datasets: Regardless of the baseline sparsity, the investigated simulated dropout rates were chosen automatically according to a logarithmic scale. Thus, the absolute range of covered sparsity levels for the seahorse dataset was much larger (48 percent points) than for the glioblastoma data (30 percent points). While population structure collectively vanished at some level of simulation, the degree of information loss should be directly associated with the absolute difference in sparsity.

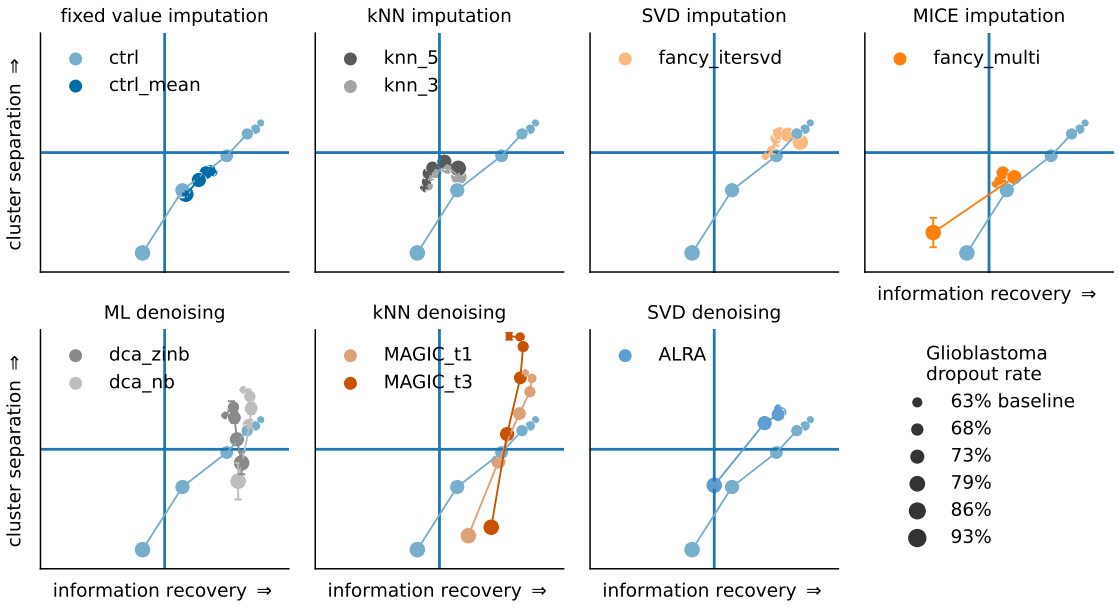

Figure 7: Summarized performance of imputation/denoising on the glioblastoma dataset. The performance of different approaches at increasing simulated dropout rates (DR) was summarized in two dimensions, each combining 4 individual metrics. While information recovery compares simulated and imputed data with original intensities (prior to the simulation of dropouts), cluster separation evaluates the respective data matrices individually. Thus, imputed datasets outperform the control (dull blue series in all plots) on the y-axis but not on the x-axis. Increasing simulated DRs are represented by growing circles. Error bars indicate the standard deviation between 5 technical replicates of imputation and evaluation. *k*NN: *k*-nearest neighbors, MICE: multivariate imputation by chained equations, ML: machine learning, SVD: singular value decomposition

Taken together, the results of this benchmark were consistent across datasets and aided the further dissection of imputation and denoising mechanisms as well as the characteristics of the investigated datasets. MAGIC denoising performed best for lower simulated dropout rates and SVD imputation for particularly high sparsity. This shows that the individual characteristics of a dataset can influence the suitability of imputation and denoising methods.

3.6 Denoising methods and SVD imputation preserve biological patterns

So far, I have shown that imputation and especially denoising techniques can effectively eliminate the technical variability present in the given single-cell datasets. This helps reveal hidden biological information and improves global downstream analysis, such as dimensionality reduction. However, single-cell data are subject to many other applications that specifically investigate the functions and associations of individual ions. Therefore, it is important that the employed methods not only recover general patterns of information but also capture the key biological features and their relationships correctly.

A commonly studied detail of single-cell data are driving features of the separation of conditions, in the case of metabolomics/lipidomics, differentially abundant metabolites (DAMs). Analyses of multi-condition omics datasets commonly include pairwise or one-vs-all comparisons of the conditions. For simplicity, the second contrasting method was used for this step. Ideally, the DAMs of different conditions are conserved by imputation or denoising and recovered by these methods at higher dropout rates.

As a preliminary analysis, I compared the pure numbers of identified significant DAMs (minimum fold change of ± 2, adjusted p-value cutoff 0.05) across imputation methods applied to the seahorse dataset. Although the absolute numbers of significantly varying ions differed greatly between conditions and datasets, the respective abilities of imputation or denoising to preserve or recover these numbers served as a consistent metric. Without additional simulation of dropouts, most denoising methods allowed identification of slightly less up- and down-regulated ions compared to the original data (MAGIC and DCA on average 85-93% of ctrl, Figure 8a). Only ALRA enabled the detection of considerably more ions than the original data (55% more than ctrl on average). In contrast, imputation methods collectively failed to guarantee comparable amounts of significant DAMs (18-34% of ctrl). Among them, SVD imputation showed consistently the best performance (44% of ctrl).

With additional simulated dropouts, imputation methods had to not only preserve but recover the biological patterns of the baseline data to enable the detection of the original DAMs. The dropout simulation without imputation reduced the average number of detected DAMs by over 50% (Figure 8b). When applied to the highly sparse data matrix, only ALRA enabled the detection of more DAMs than the baseline (17% more than baseline), while DCA (dca_nb), SVD imputation and MAGIC (t1) still recovered more significant ions than no imputation (78%, 62% and 52% of baseline, respectively). In contrast, the application of all other imputation methods resulted in the detection of little to no significant DAMs. This is

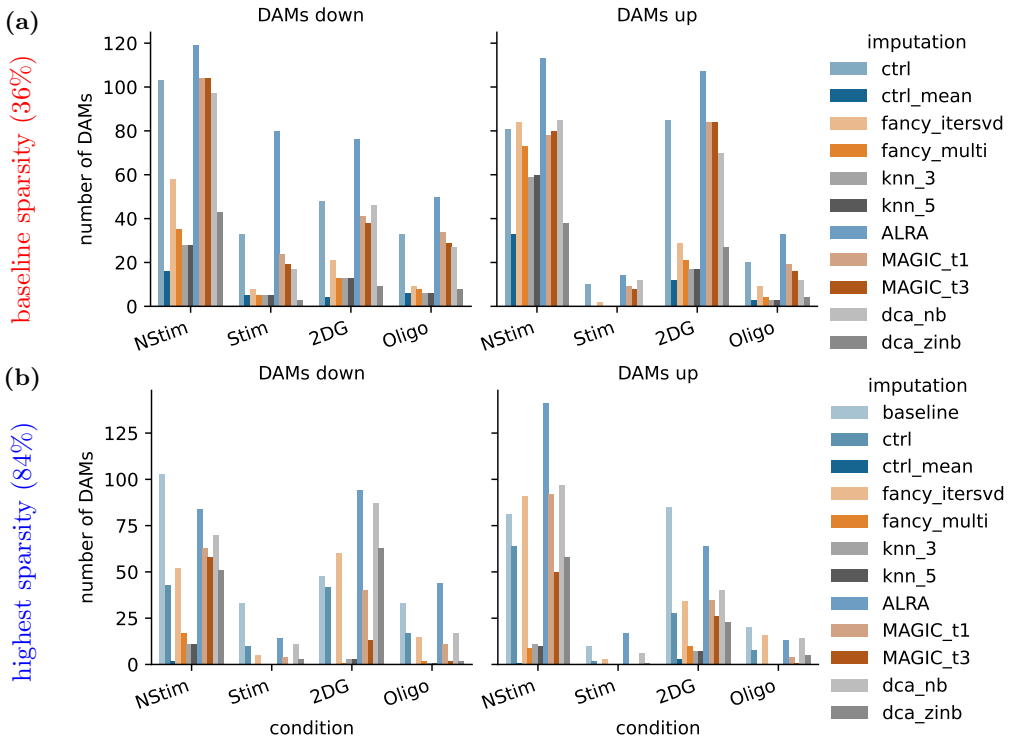

Figure 8: Numbers of significant DAMs detected after imputation of the seahorse dataset. **(a)** Without simulation of additional dropouts and after application of varying imputation/denoising methods, each condition in the dataset was contrasted against the rest of the cells and differentially abundant metabolites (DAMs) are identified at fold changes greater than ± 2 and adjusted p-value < 0.05. The numbers of significant ions without imputation (`ctrl`) and after applying different methods are visualized separately for down-regulated and up-regulated compounds. **(b)** The same way, DAMs were detected in the dataset after the simulation of the highest level of missing values. Here, `baseline` refers to the original data before simulation (same as `ctrl` in (a)) and `ctrl` to the dropout-simulated but not imputed data. Methods are separated in imputation (top: `ctrl_mean` - `knn_5`) and denoising (bottom: `ALRA` - `dca_zinb`).

in line with the previous observations that denoising methods best retained the patterns of the data at low sparsity and that SVD imputation performed well on the data at simulated high sparsity. However, ALRA seemed to be an exception as it consistently increased the number of detected DAMs.

In addition to the mere number of significant differentially abundant ions, their corresponding effects should be conserved after imputation. This included the relative order of ion fold changes and top differential ions identified. To enable a comprehensive comparison despite the varying numbers of significant DAMs across methods, I analyzed the calculated fold changes of the down- and up-regulated ions with respect to the different conditions. As a metric of preservation of the metabolite differences after imputation, I compared the ranks of the corresponding log-fold changes using Kendall's rank correlation (see Supplementary Figure S.11). Without additional simulation of missing values, MAGIC denoising exhibited the highest accordance with the original data across datasets (Figure 9a). Despite its increased yield of significant DAMs, ALRA showed a slightly lower agreement. Again, SVD imputation

performed slightly better than the other imputation methods. To assess the performance at high levels of sparsity across datasets, I selected the respective highest or second highest level of missing value simulation to assemble a set with homogeneous dropout rates between 84 - 89%.

Surprisingly, here the control of no imputation (`ctrl`) showed overall the highest agreement with the original data despite the high level of sparsity. Yet, SVD imputation enabled comparable recovery of log-fold change ranks as the best denoising method MAGIC (Figure 9b). The high performance of the control at high sparsity could be connected to the biased simulation approach: As the simulated dropouts affected mostly low intensities, the higher intensities that likely drive the inter-condition differences were largely preserved.

Of note, the agreements between DAM ranks were more consistent across datasets for the baseline sparsity than for the simulated high sparsity datasets. This is likely owing to the fact that the datasets were subject to slightly different simulated dropout rates. It further indicates that the exact level of sparsity is an important driver of the loss of biological patterns.

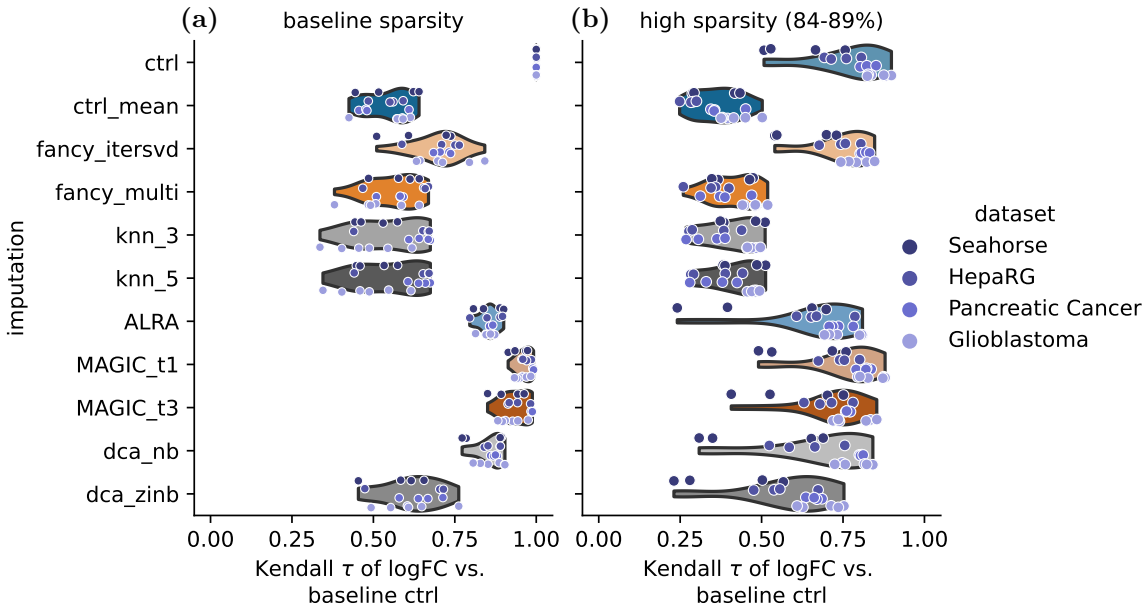

Figure 9: Recovery of ion log fold changes through imputation/denoising. **(a)** For each condition across datasets, log-fold changes of all ions were calculated and ranked from most up-regulated to most down-regulated. This ranking for different imputed data matrices is compared to the original data before simulation and imputation. This baseline data equals the ctrl in the left panel. **(b)** On the right side, a high dropout rate was selected for each dataset and the log-fold change ranks with/without imputation were compared to the ground truth data at baseline dropout rates. A visualization of the rank comparisons is shown in Supplementary Figure S.11.

In a practical analysis, usually, only the most prominent features in a differential analysis are closely investigated. Thus, especially those features should be preserved after imputation or denoising. To assess the methods with respect to this aspect, I selected the top 20 significant up- and down-regulated DAMs and calculated how well this set from the baseline data was recovered after direct imputation or after simulation of high sparsity followed by impu-

tation. Similarly to the overall agreement of the log-fold change ranks, MAGIC best retained the top differentially abundant features from the baseline data (70% and 60%, for $t = 1$ and $t = 3$ respectively) while all other methods performed poorly (Figure 10a). However, at high dropout rates, none of the methods recovered more than 18% of the top differentially abundant ions on average, and mean imputation showed no overlap at all (Figure 10b). Again, SVD imputation performed best under this condition and matched the recovery rate of the non-imputed control data.

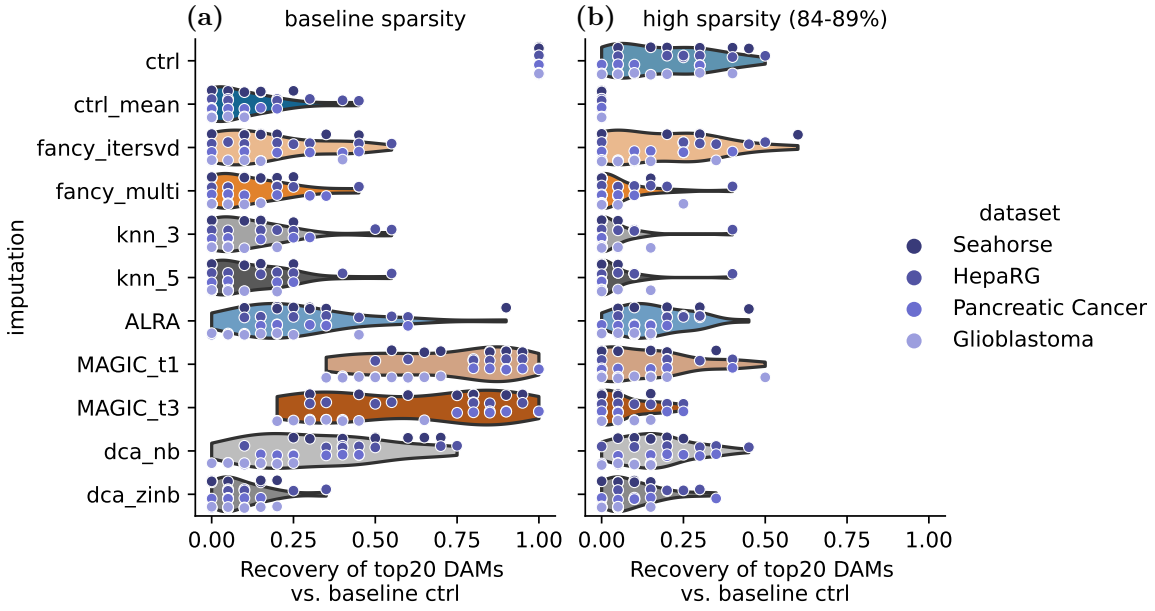

Figure 10: Recovery of top DAMs after imputation/denoising. **(a)** For each condition across datasets, the top 20 significant up- and down-regulated differentially abundant metabolites (DAMs) ($|FC| > 2$ and $P_{adj} < 0.05$) were selected. These sets were then compared between the ground truth data matrix and all imputation/denoising methods employed (up/down separately). The fraction of ions from the baseline data, that were also found after applying the respective methods is visualized on the x-axis. **(b)** The right plot compares the top DAMs recovered by imputation/denoising after extensive simulation of missing values.

Taken together, these results suggest that especially denoising methods like MAGIC retained a lot of the quantitative biological patterns in the data. In contrast, most of the imputation methods led to the loss of significant features that drive the separation of conditions. Again, SVD imputation stood out at high levels of sparsity, where it best recovered the general order and most of the top significant differential ions. Although differential analysis does not seem to be the most promising application of data imputation or denoising, this analysis provided some general insights into the impact of the respective methods on the biological information encoded in the given datasets.

3.7 Imputation and denoising highlight biologically relevant ion-ion correlations

A more common application of imputation and denoising in omics studies is correlation-based network inference. The associated methods establish compound networks based on co-abundance to identify modules of molecules that might be co-produced, co-regulated, or co-transported. While gene correlation network analysis has been developed and used for almost two decades [83], the concept has not been adopted to single-cell metabolomics and lipidomics so far. In this regard, one serious obstacle is the high sparsity of metabolomics and lipidomics data which impedes the calculation of common correlation metrics. This can be easily observed when looking at the distribution of Pearson correlation coefficients (scaled to edge weights between 0 and 1) for the control condition of the seahorse dataset. Already at baseline sparsity, almost all ion pairs had an edge weight around 0.5 which translates to a Pearson correlation of 0 and the lack of a positive or negative relationship (Figure 11a). This effect is even greater after the simulation of missing values. To reduce the influence of zeros on pairwise ion correlations, there is an alternative approach to imputation. This method calculates correlations selectively, disregarding cells in which one or both of the compared ions were not detected. The procedure broadens the distribution of edge weights to include higher positive and negative correlations at the cost of a smaller sample size in each ion pair. On average, correlations calculated in this way were based on only 59% and 16% of all cells under baseline and high sparsity conditions, respectively. In the latter case, the accumulation of extreme weights indicates that they were solely based on a few cells. Interestingly, the distribution of edge weights based on nonzero intensities at baseline sparsity was shifted towards higher values, indicating that true ion relationships might have been rather positive. This was reflected by the selected imputation but not by the denoising methods. The latter exhibited a greater dispersion, thus more ion pairs with a strong positive or negative correlation. These observations also applied to the control condition of the glioblastoma dataset (Figure 11b). Notably, correlations were analyzed separately for conditions, because some ion-ion correlations across conditions followed non-linear relationships or deviated from within-condition associations (compare Figure 5a).

For both effects, increased dispersion and shift of the distributions, it is not straightforward to determine the desired effect size. For example, ALRA produced strong positive pairwise ion correlations after simulation of high sparsity (Figure 11a, red asterisk), confirming previous observations that the method led to overall increased ion-ion correlations (see Supplementary Figure S.9). This can be attributed to the approximation rank k, which ALRA determines empirically during runtime and uses to compress the data. For this run of the seahorse dataset, the determined k decreased to a value of 5 compared to 11 for the glioblastoma dataset. As a result, an artificial yet pronounced linear dependence was introduced to the data, which became observable through increased positive correlations. Nevertheless, other imputation and denoising methods successfully increased the dynamic range of ion-ion relationships and thus facilitated the generation of well-differentiated correlation networks.

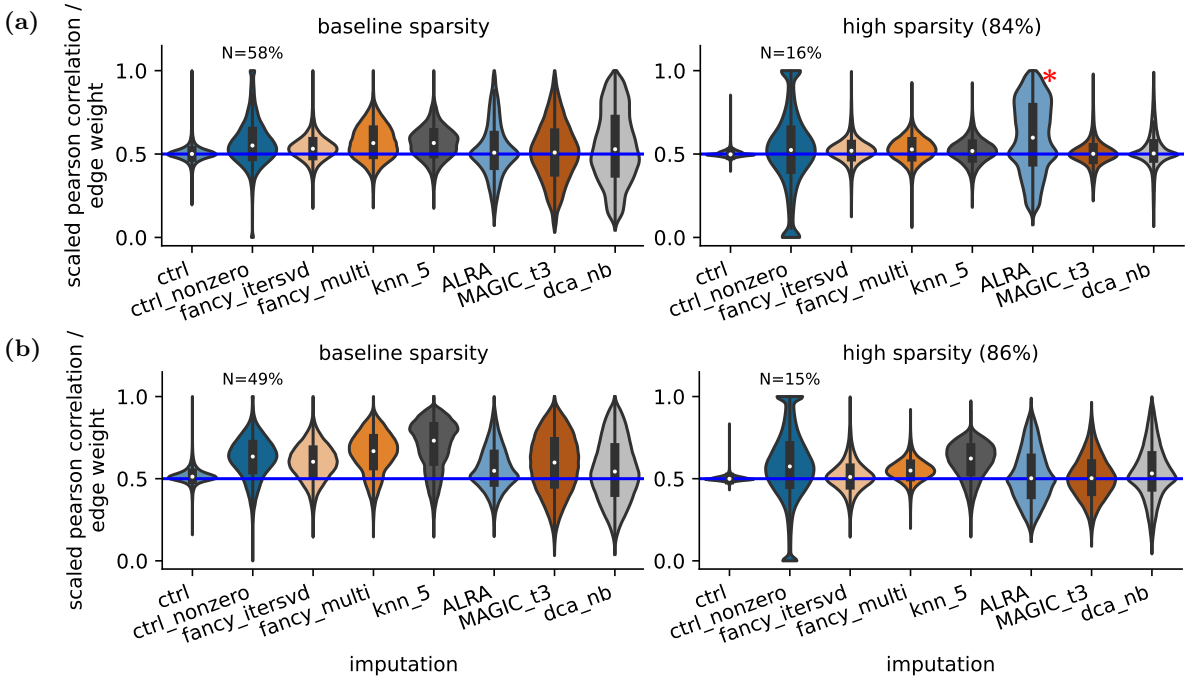

Figure 11: Impact of imputation/denoising on ion-ion correlations. This figure shows the pairwise ion-ion correlations (Pearson) calculated for cells from the respective control conditions (NStim, Naive_WT) of the **(a)** seahorse and **(b)** glioblastoma datasets. Each imputation/denoising method yields a different distribution of correlation coefficients. ctrl_nonzero serves as a simple control, ignoring all zero intensities from the non-imputed data (ctrl) when calculating the pairwise correlations. In this regard, N denotes the average fraction of cells available for the calculation after removing zero intensities. The red asterisk (*) denotes likely artificial positive correlations due to a very low rank approximation.

The simplest ion correlation networks are based directly on the scaled correlation coefficients between pairs of ions. The resulting networks are fully connected and highly co-abundant metabolites are connected by edges with high weights. To establish whether a correlation network of the glioblastoma conditions reflects biological relationships, I compared them with a prior knowledge network of lipid enzymatic reactions. This approach is driven by the assumption that two lipids connected through a biological reaction should generally exhibit a higher agreement in their abundance than molecules that are not directly connected. A lipid reaction network including the most important lipid conversions (elongation, desaturation, hydroxylation/oxidation) was generated by linex2 based on the ions present in the glioblastoma dataset [79]. This prior knowledge network was used to classify the edges of the correlation network into two groups: Pairs of lipids connected by a reaction and pairs of lipids that were not connected by a reaction. Indeed, for the network of the Naive_WT condition based on non-imputed intensities, the median edge weight of ion pairs connected by a reaction was slightly higher than without a reaction (ctrl: 2.4%, ctrl_nonzero: 4.4%, Figure 12a). Notably, this was associated with an absolute increase in the underlying correlation instead of a reversal of the effect, as edges without biological evidence already had a positive average correlation (Supplementary Figure S.12, applies to all methods/condition

networks). However, compared to the non-imputed intensities, many imputation and denoising methods improved the prominence of edges with biological reactions to a larger extent (fancy_itersvd: 7.5%, ALRA: 11.2%, MAGIC: 11.7%, dca_nb: 19.6%), even after simulation of additional missing values. When evaluating the different imputation methods on all conditions of the glioblastoma dataset, it became obvious that the selected methods highlighted biologically relevant network edges consistently (Figure 12b): While the raw intensities without imputation only led to modest prominence of biologically relevant edges, MAGIC, DCA, ALRA, and SVD imputation highlighted these ion pairs on average by 10%. Moreover, this effect was highly significant (adjusted $P = 0.0034$, Wilcoxon's rank sum test, sample size of 6 conditions) for these methods, in contrast to MICE and kNN imputation as well as the correlation approach only using nonzero intensities. These results show not only that the generated correlation networks for the glioblastoma dataset reflected the topology of true biological reactions to some extent, but also that imputation or denoising have the potential to significantly clarify the agreement of the obtained correlations with prior knowledge.

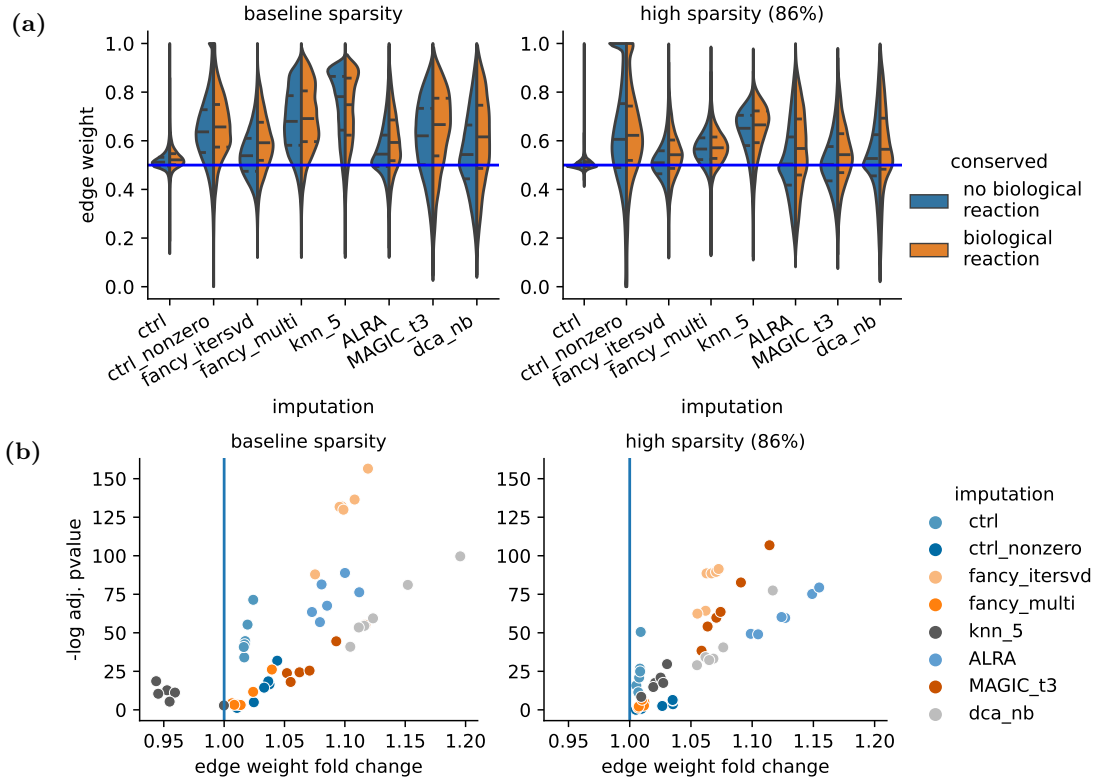

Figure 12: Biological relevance of lipid-lipid co-abundance from the glioblastoma dataset. **(a)** For the ion correlation network of Naive_WT cells from the glioblastoma dataset, edge weights were classified if they can be associated with a biological reaction (lipid reaction network generated using linex2) or not. Generally, the biologically relevant edges have higher weight distributions than the rest. The half violins show median and quartiles as horizontal lines. Apart from knn_5, all differences are highly significant (adjusted $P < 10^{-7}$) **(b)** For all condition networks and imputation methods, the comparison between edges with and without biological reaction is shown as a volcano plot. A higher fold change (x-axis) corresponds to a stronger prominence of biologically relevant edges. Adjusted p-values of Wilcoxon's rank sum test are plotted on the y-axis.

3.8 Batch-balanced MAGIC removes batch effects and missing values

With the relatively limited throughput of this novel single-cell technology, larger single-cell metabolomics and lipidomics studies usually require the acquisition of multiple distinct batches. Moreover, in the future, separate datasets will be integrated to establish atlas-level data collections. Usually, whenever MS data acquired on different occasions is combined, batch effects introduce serious technical variability and hence impede the biological analysis of the data. To circumvent these issues in other omics studies, a host of batch correction and integration methods have been introduced that integrate the data on different levels: The simplest method ComBat shifts and scales the intensity distributions of different batches to generate an integrated data matrix. However, this approach is not necessarily suitable to cope with the zero-inflated single-cell metabolomics and lipidomics data. In contrast, BBKNN creates a batch-corrected neighborhood graph that can be used for an integrated visualization, for instance in UMAP space. This method has been among the popular approaches for single-cell data, but its output is unsuitable for further downstream applications of the data. Thus, there is an unmatched demand for a unified approach to tackle data sparsity, technical noise, and batch effects all at once.

To address these challenges, I combined the existing tools for batch integration and denoising to create a new method: Batch-balanced MAGIC (bbMAGIC) uses the batch-integrated neighborhood graph generated by BBKNN to drive the iterative data smoothing process of MAGIC. That way, cells are denoised based on information from neighboring cells in all separate batches such that inter-batch differences in populations are reduced through the smoothing process. The two methods fit perfectly together since the smoothing process of MAGIC is based on a neighborhood graph that does not account for batch differences. This first step of the method can be readily replaced by its batch-balanced counterpart as implemented in BBKNN. To showcase this combined method to overcome data sparsity and batch effects in comparison to ComBat, I selected the pancreatic cancer dataset as it suffers from the given limitations.

Again, a UMAP embedding can give a comprehensive overview of the overall structure of the dataset. For the original (non-integrated, non-denoised) ion intensities, batch artifacts are clearly visible: Cells separate into clusters not only by conditions but also by batches (Figure 13, `ctrl` panel). In particular, the third batch forms clusters distinct from the others. In principle, these batch effects could be reduced by both ComBat and BBKNN, reestablishing cell clusters according to the biological conditions. While the former method only brought the batches closer together, the latter led to a true integration (Figure 13b, `combat`, `bbknn` panels). In comparison, running MAGIC alone preserved the population structure of the data, but did not reduce the impact of batch artifacts. To combine the batch-integration and denoising effects, both ComBat and BBKNN were then coupled with MAGIC. Since ComBat produced an integrated matrix of ion intensities, MAGIC could be applied directly to this data. In contrast, the connection of BBKNN and MAGIC was achieved at the level of

a neighborhood graph, through batch-balanced MAGIC (bbMAGIC). Employed after both batch correction methods, the denoising preserved the respective batch integration effects of both ComBat and BBKNN. Again, ComBat+MAGIC only brought batches closer together while bbMAGIC led to proper intermixing between them. These first insights indicate that

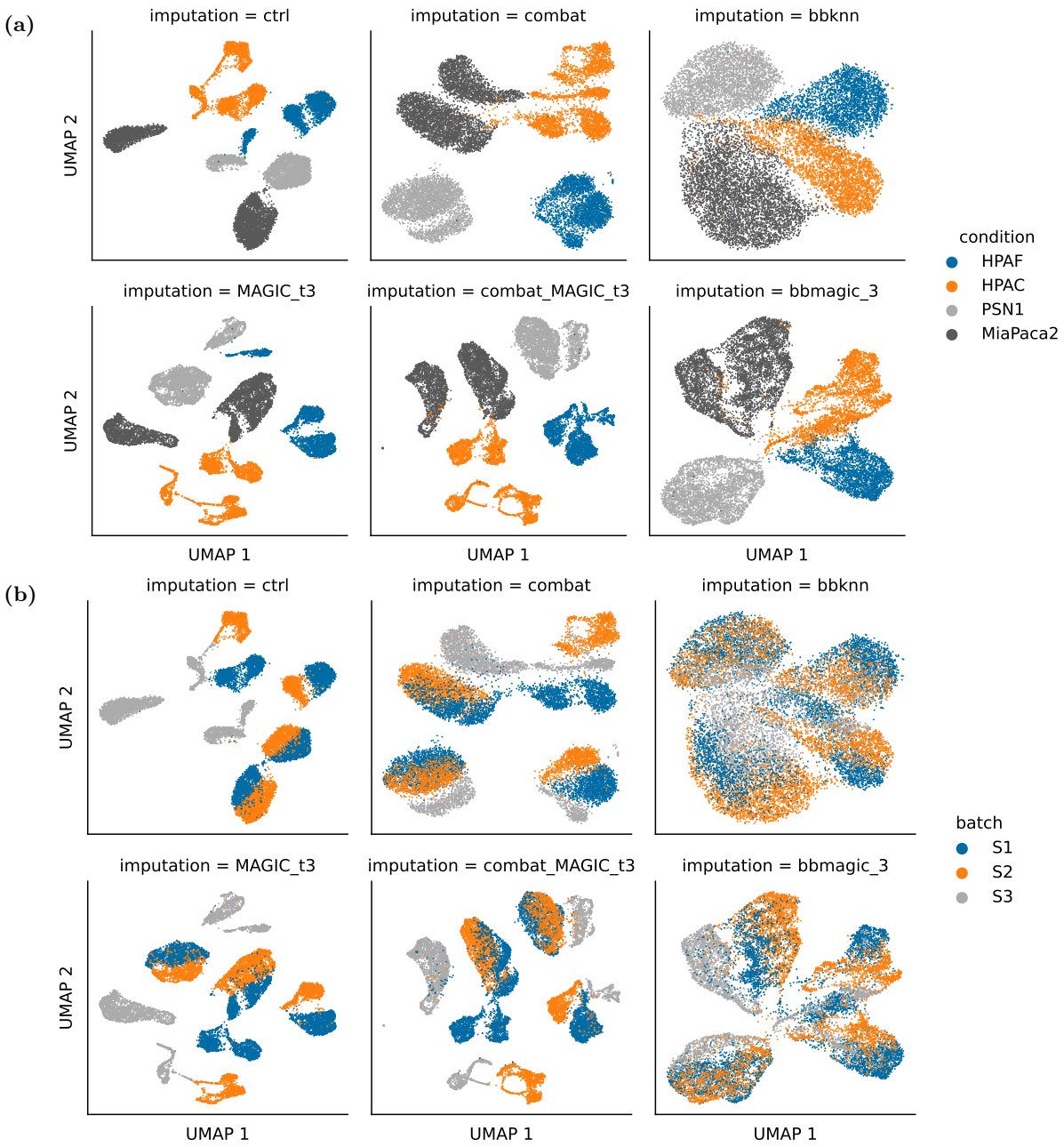

Figure 13: Combined batch integration and denoising on the pancreatic cancer dataset. UMAP embedding of the dataset (filtered according to Table S.1, without simulation of missing values) before/after data imputation. The top row shows different processing regimens without denoising, namely the embedding of the original intensities (`ctrl`) and after the application of two batch correction methods. In the bottom row, these methods (or the control) are complemented by MAGIC denoising: For `combat_MAGIC_t3`, the batch-corrected intensities were denoised while for `bbmagic_3` the batch-integrated neighborhood graph from BBKNN was supplied directly to MAGIC. Cells are colored by **(a)** conditions and **(b)** batches, respectively.

combining batch correction and denoising may aid in recovering the overall structure of a dataset.

Since both reviewed tools generated batch-corrected and denoised ion intensity matrices, the question arose of how they impacted the distributions of ion intensities and whether the processed distributions were suitable for further downstream analyses. For this purpose, I selected three representative ions with stronger batch differences that were annotated with phosphatidylcholine (PC), lyso-phosphatidylcholine (lyso PC), and glycosyl-ceramide (GlcCer) species, respectively. Analogous to the UMAP embedding, the individual raw ion distributions exhibited clear batch differences: In particular, the third batch (Figure 14a) deviated strongly from the others, in terms of the mean but also in the fraction of missing values. By design, ComBat shifts and scales the ion distributions of different batches to align them with respect to their mean and standard deviation. Usually, this works well for low-sparsity data where batches mostly differ in their location and scale. However, since the examined lipidomics data had many missing values, the batch correction of ComBat produced negative values and strongly increases zero intensities (not shown). This state was only slightly modified by the following denoising step, leading to strong modifications of the ion distributions of the third batch that included a serious shift of missing values by an order of magnitude (Figure 14b). In contrast, the batch-balanced neighborhood graph of bbMAGIC smoothed out batch differences such that the ion distributions across batches moved closer together and now occurred in the same interval (Figure 14c). Furthermore, as MAGIC-based smoothing intensifies successively with higher iterations, the degree of integration could be controlled manually: Whereas the first iteration of smoothing only brought the separate batch clusters closer together, the fifth iteration led to strong intermixing of the batches and greater con-

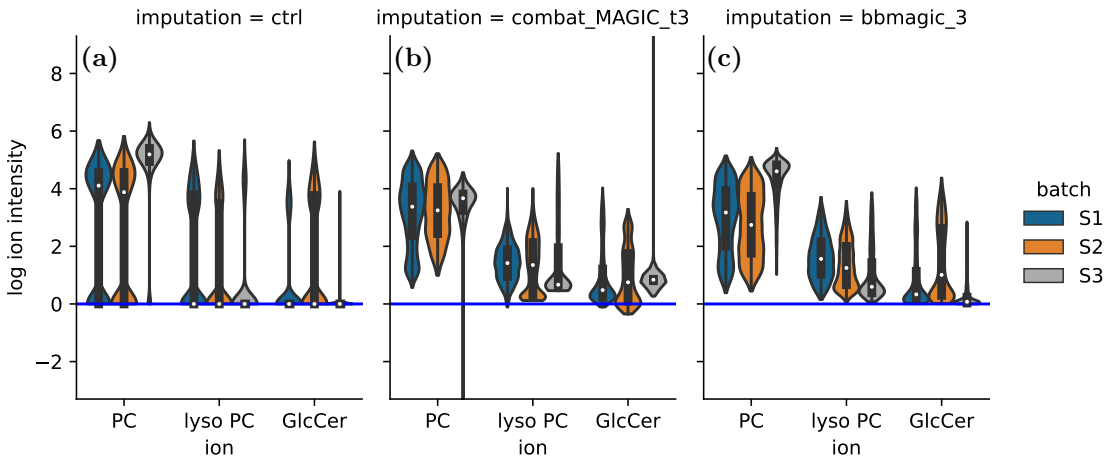

Figure 14: Effects of combined batch effect correction and data imputation on selected ions. Compared to the non-imputed data in **(a)**, the transformative effects of **(b)** ComBat+MAGIC and **(c)** bbMAGIC are shown with three exemplary ions: a phosphatidylcholine (PC, $C_{44}H_{84}NO_8P$+Na), a lyso-phosphatidylcholine (lyso PC, $C_{26}H_{48}NO_7P$+H) and a glycosyl-ceramide (GlcCer, $C_{47}H_{79}NO_8$+Na). Some violins extend beyond the plotting area, indicating that the intensity distributions generated by ComBat+MAGIC range from negative to very high values. Cells are colored according to batches.

vergence in the distributions (see Supplementary Figure S.13). This suggests that bbMAGIC may be a promising solution for handling both batch effects and data sparsity and could thus aid emerging efforts to establish single-cell metabolomics and lipidomics atlases.

The combined tool bbMAGIC is available as a python library from GitHub (mariusrklein/bbMAGIC) and can be easily integrated with an existing scanpy-based analysis workflow.

Chapter 4

Discussion

Single-cell metabolomics and lipidomics have emerged as powerful technologies for studying cellular heterogeneity and elucidating metabolic processes in the context of health and disease. However, these cutting-edge methods still face several challenges that impede their full potential. The limitations include data sparsity, technical noise, and especially batch effects, which can introduce significant biases and hinder the accurate interpretation of results. Fortunately, advancements in other scientific domains and omics levels have seen the development of effective tools to overcome similar obstacles. This represents an opportunity to adapt those methods to single-cell metabolomics and lipidomics analysis, enabling and enhancing downstream analyses, such as cell-type and lineage analysis, correlation-based network inference, pathway enrichment, and biomarker discovery. Therefore, this work aimed to employ various metabolomics and lipidomics datasets to identify suitable approaches for imputation and denoising, thereby increasing the applicability of novel single-cell datasets. Finally, I aimed to provide a comprehensive framework for extending this comparative analysis, paving the way for future advancements in single-cell metabolomics and lipidomics research.

In this work, I systematically applied a range of imputation and denoising methods to single-cell metabolomics and lipidomics datasets, quantitatively evaluating their performance in terms of information recovery, cell population separation, and preservation of crucial biological features. Through this comprehensive analysis, I successfully characterized the different underlying approaches and assessed their suitability for datasets with varying prerequisites with regard to patterns of missingness and the baseline level of sparsity. Furthermore, I validated the investigated methods in a probable application scenario, namely correlation-based network inference. Leveraging prior knowledge of lipid reactions, I confirmed their ability to unveil concealed biological information within the datasets. My findings provide valuable insights into the efficacy of different imputation and denoising techniques, offering the field a clearer overview of the most appropriate strategies for enhancing data quality and extracting meaningful biological insights from single-cell metabolomics and lipidomics experiments.

Systematic filtering balances the trade-off between data sparsity and information loss

Early in the investigation, I observed that filtering, despite being an often underestimated and underutilized step, played a crucial role in producing cleaner single-cell data, both to generate a ground truth and for general analysis purposes. With previous filtering strategies that relied on arbitrary cutoffs, if thresholds were set too high, studies faced the risk of unevenly diminishing cell populations and thus introducing systematic bias. In contrast, low thresholds would eliminate the effect of filtering to reduce data sparsity. Using the presented systematic approach, I was able to balance this trade-off in order to salvage the maximum amount of biological information while effectively removing the most corrupted features. In general, a closer investigation of the cell distributions (given the unique number of ions) stratified by conditions has proven useful to guide this decision. In bulk metabolomics, some strategies are commonly employed that extend this simple filtering approach. For instance, one method considers the features in individual conditions, retaining ions that are only abundant in one of the conditions. This accounts for the possibility that certain ions are not measured under specific circumstances [55]. Implementing such filtering techniques would further increase the chance of retaining those biologically highly relevant but elusive metabolites. However, even without that extension, filtering in this work not only mitigated sparsity and thus improved the quality of downstream results. It also enabled the inference of a low-sparsity ground truth for the following benchmark.

Despite these efforts, a truly noise-less ground truth of metabolite or lipid intensities could not be established based on the given noisy data. In particular, rigorous filtering reduced the impact of missing values but did not remove technical noise in the measured intensities. This reduced confidence in the ground truth, especially at higher baseline dropout rates (e.g. glioblastoma dataset), making it more difficult to conduct a comprehensive comparison of data recovery methods. To address this limitation, a more extensive solution would involve a complete simulation of data. To my knowledge, this has not been done in single-cell MS-based omics yet, and doing so would fall beyond the scope of this research. However, both completely theoretical models and ML-based simulators trained on real data have been applied successfully in single-cell transcriptomics and are now being translated to other modalities [84, 85]. With the accelerating progress in deep learning, especially the latter approach has gained increased momentum [86]. Such simulation studies would complement the present work in providing more accurate insights into the performance of different imputation and denoising methods, in particular regarding technical noise in intensity measurements.

New specialized methods for MS-based data hint at an evolution of the field

It is important to note that in this study, dropouts were simulated as the largest driving pattern missing not at random (MNAR), while in reality, this can be overlaid with other patterns, such as missing at random (MAR) and missing completely at random (MCAR). Furthermore, each algorithm makes different assumptions about the nature of the missing

data and will perform optimally only if confronted with its assumed pattern. Even though most of the used methods assumed one of the simpler types of missingness, they were collectively applied to data with MNAR characteristics. For example, despite its design for MCAR, fixed value imputation by the feature mean was used to determine whether the simplest possible methods already served as a satisfactory solution. Of the more sophisticated methods, DCA made the strongest assumptions, modeling the data like counts with a discrete negative binomial distribution. Although DCA generally performed well, this indicates that there is still great potential to adjust imputation and denoising methods to the particular characteristics of single-cell MS-based data.

For instance, a new method for proteomics was published recently that tackled both MAR and MNAR sparsity of MS-based data, matching the early observations from this work [87]. Given the shared mass spectrometry background, it may be more applicable to single-cell metabolomics and lipidomics than methods derived from transcriptomics. In addition, many ML-based methods, such as DCA, focus their assumptions on statistical distributions for data modeling and the nature of feature relationships (e.g. linear or non-linear). Another example is scVI, an emerging framework for modeling and analyzing single-cell RNA-sequencing data using variational autoencoders [88]. Instead of imputation, the authors aimed to probabilistically estimate the cellular mechanisms underlying observed gene expression patterns, taking into account missing values, batch effects, and other confounding factors. When using such a machine learning framework, exploration of alternative distributions and other small adjustments are feasible and have the potential to greatly improve the performance with respect to the present data.

Yet another approach to analyzing sparse data has been proposed recently in the context of bulk proteomics: As an alternative to imputation, the authors modeled and estimated the detection probability of peptides based on their intensities using a logit-linear function [53]. On the one hand, their results confirmed again the observations from this and other work that missingness is driven by at random and not at random (MAR and MNAR) patterns. On the other hand, the knowledge of detection probabilities was successfully incorporated into a likelihood-based analysis of differential expression, significantly outperforming other imputation approaches. However, it should be noted that the ion intensities in this work were processed by the SpaceM method, which limits the applicability of the approach in this particular context.

The desired biological application should drive the choice of imputation and denoising methods

Out of the imputation and denoising methods included in this study, no single approach excelled in all circumstances and applications. Instead, due to the variety of underlying mechanisms different methods adapted best to the associated requirements of various use cases. A common early analysis concerns the separation of cell types, perturbations, or cell states. Beyond quality control, this part of an investigation can help uncover general heterogeneity and cell type transitions in a cell population. For the underlying dimensionality

reduction tasks, overall variability in the data, in this case, the metabolic profiles, drove the separation of cell subpopulations instead of individual features. Here, the abundance of missing values, represented by zeros, already introduces serious technical variability that obscures the biological variability necessary to distinguish between conditions or cell types. Thus, even a simple method such as fixed value imputation could reduce this technical variability by replacing zeros with values from within the intensity distributions. That way, mean and minimum imputation bear the potential for successful application in this early part of an analysis.

In order to identify the driving features for the separation of cell subpopulations, researchers commonly use differential expression analysis. In my research, I have applied these well-established methods to identify differentially abundant metabolites between conditions to evaluate the effectiveness of imputation and denoising methods in restoring biologically relevant patterns. However, the applicability of denoising to differential expression analysis has been questioned, as highlighted by the authors of the MAGIC framework [66]. Their concerns regarded the risk of overestimating statistical significance when analyzing denoised data. Therefore, beyond evaluation, simple statistical differential expression analysis should not be performed on denoised data. Instead, methods specifically designed for sparse MS-based intensity data, for example from more established proteomics research, could be a more suitable alternative to approach differential expression analysis in single-cell metabolomics and lipidomics [89, 90].

In transcriptomics investigations, another important objective is to distill the scattered biological information from high-dimensional datasets into concise gene correlation or regulatory networks. Here, novel imputation and denoising methods have raised hopes for increasing the modest success in extracting clear relationships from the noisy single-cell RNA-sequencing data. Although the methods were recommended by their developers to improve gene correlations [66, 68], critical voices stressed the risk of introducing artificial false-positive correlations and urged that results should be interpreted cautiously [91, 59]. Network inference was also adopted in bulk metabolomics, mostly based on Pearson/Spearman correlation as an association metric [92]. However, in contrast to the present findings on lipid networks, a metabolomics study found that certain neighboring amino acids showed weak correlations, whereas distant pairs demonstrated robust relationships [92]. This was attributed to the quasi-steady-state nature of metabolism, which cancels out associations between substrates and products of the same enzymes and instead promotes indirect system-wide correlations [93]. Such considerations may be less applicable to lipids, as their structural relevance to cells and accumulation justifies a more intricate regulation of their conversion and abundance. Nevertheless, this underscores the need for further validation of the presented lipid correlation networks beyond the association of compounds with a biological reaction.

The main limitation of current correlation-based network inference approaches is the challenge of determining appropriate hard thresholds for retaining meaningful relationships from a complete correlation matrix. Traditional solutions commonly applied an arbitrary cutoff of the correlation coefficient or p-value, which not necessarily reconstructed biologically rel-

evant patterns. In contrast, a promising approach from bulk omics used prior knowledge of biochemical networks to determine the correlation cutoff in a data-driven manner: The networks resulting from different thresholds were compared to a corresponding prior knowledge reaction graph and the threshold with the highest overlap was used to subset the network edges [94]. This method robustly enriched meaningful biological relationships in correlation networks and may also be suitable for single-cell datasets.

Another notable method in transcriptomic network inference is weighted gene correlation network analysis (WGCNA), where DCA has demonstrated good performance in recovering correlations under medium-sparsity conditions [95]. Instead of applying thresholds manually, this approach shapes networks according to the assumption that the network topology is scale-free with few hub nodes and many features with low connectivity. Whether this theory applies to many biological and natural networks has been controversially discussed, especially with respect to metabolic networks it has often been rejected [96, 97]. It was also not observed in the datasets examined in this work (data not shown).

However, one of the datasets allowed for the combined assessment of missing values with batch effects. Specifically, I observed that the pancreatic cancer dataset exhibited batch differences that were to some extent driven by varying sparsity of individual ions. This finding was consistent with previous observations in bulk MS-based metabolomics studies that both intensity means and detection rates varied between batches [44]. Considered as MAR, these batch effects align nicely with the less pronounced negative relationship between mean ion intensities and dropout rates observed early in this study. Since SVD imputation was designed specifically to tackle MAR, it is not surprising that the method outperformed the others in recovering the condition-based clusters within this dataset. However, I have also shed light on dedicated methods for dealing with batch effects. bbMAGIC combines the neighborhood-based batch integration method BBKNN with the smoothing-based denoising approach MAGIC. These two widely used tools complement each other perfectly: The batch-balanced neighborhood graph generated by BBKNN is predominantly used for visualization purposes, limiting the applicability of the method in downstream analysis. In turn, the iterative smoothing approach of vanilla MAGIC includes the calculation of a neighborhood graph of similar cells, that does not take batch differences into account. Consequently, feeding the batch-balanced neighborhood graph of BBKNN to MAGIC's smoothing process optimally exploits the respective strengths of these methods. Although I have presented general proof of this concept, a more detailed investigation of combined batch integration and denoising goes beyond the scope of this work. Thus, further research is needed to systematically evaluate different approaches to this problem.

Conclusions

Taking all facets of the evaluation together, some of the employed imputation and denoising methods stood out from the crowd: After strict filtering, MAGIC and DCA denoising performed best at preserving the overall structure and clarifying the clustering of the given datasets. In contrast, iterative SVD imputation excelled at recovering both clustering and

information from highly sparse data matrices. Accordingly, these methods dominated the evaluation in terms of recovering biological information. Although each dataset presents different requirements concerning noise and sparsity, the present work provides a detailed overview of the underlying mechanisms of imputation and denoising and their impact on sparse single-cell metabolomics and lipidomics data.

In addition, this work gives actionable insights to guide future analysis workflows. Sparsity and noise have been shown to occur in heterogeneous patterns. To mitigate them while preserving biological information, the structure of a dataset should first be dissected regarding conditions, batches, and other confounding factors. Filtering out sparse cells and ions has the potential to greatly reduce data sparsity, but thresholds should be chosen systematically to maximize its effect. The various technical influences in mass spectrometry typically result in a mixture of missingness patterns, in particular, measurement sensitivity bias and batch effects drive missing not at random (MNAR) and missing at random (MAR), respectively. Still, different datasets and even different applications may warrant the use of specific imputation or denoising methods. For dimensionality reduction and visualization purposes, simple fixed value estimation may be sufficient to promote biological variability in the data. In contrast, further downstream analyses like network inference can require more sophisticated methods to specifically highlight biological relationships. However, recent developments have focused on specific analysis approaches for sparse data, rather than processing it to enable analysis with traditional methods. These novel methods have great potential to complement data imputation and denoising in extracting biological information from noisy and sparse data. Although the field of single-cell metabolomics and lipidomics is still in its infancy, these insights contribute to an initial framework for the beginning standardization of data processing and analysis.

Taken together, my findings provide valuable lessons about the challenges of data sparsity and noise. On this basis, I have identified approaches and prioritized methods to overcome these challenges and make better use of emerging single-cell metabolomics and lipidomics data. However, it is important to recognize the limitations associated with the assumptions that certain methods make and with the narrow applicability of the presented ground truth estimation. Newly acquired datasets should be carefully examined to characterize the driving pattern of missingness, as it was shown to be an important determinant of the success of imputation [98]. In turn, only the correct pairing of these will allow for sound conclusions from various downstream analyses. Moreover, future research should aim to address the current limitations, extend the relevance of this work with simulations, and explore new approaches, leveraging the latest insights from proteomics research. In this way, the power of single-cell metabolomics data can be harnessed further, contributing to more comprehensive insights into cellular metabolism in the context of health and disease.

Bibliography

[1] NCD Risk Factor Collaboration (NCD-RisC). Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19·2 million participants. *The Lancet* **387**, 1377–1396 (2016). URL https://doi.org/10.1016/s0140-6736(16)30054-x.

[2] Kleinman, N., Abouzaid, S., Andersen, L., Wang, Z. & Powers, A. Cohort analysis assessing medical and nonmedical cost associated with obesity in the workplace. *Journal of Occupational & Environmental Medicine* **56**, 161–170 (2014). URL https://doi.org/10.1097/jom.0000000000000099.

[3] Organisation for Economic Co-Operation and Development (OECD). The prevention of lifestyle-related chronic diseases (2008). URL https://doi.org/10.1787/243180781313.

[4] Saltiel, A. R. & Kahn, C. R. Insulin signalling and the regulation of glucose and lipid metabolism. *Nature* **414**, 799–806 (2001). URL https://doi.org/10.1038/414799a.

[5] Marchesini, G. Nonalcoholic fatty liver, steatohepatitis, and the metabolic syndrome. *Hepatology* **37**, 917–923 (2003). URL https://doi.org/10.1053/jhep.2003.50161.

[6] Hotamisligil, G. S. Inflammation and metabolic disorders. *Nature* **444**, 860–867 (2006). URL https://doi.org/10.1038/nature05485.

[7] Goldberg, I. J., Trent, C. M. & Schulze, P. C. Lipid metabolism and toxicity in the heart. *Cell Metabolism* **15**, 805–812 (2012). URL https://doi.org/10.1016/j.cmet.2012.04.006.

[8] Mattson, M. P. & Arumugam, T. V. Hallmarks of brain aging: Adaptive and pathological modification by metabolic states. *Cell Metabolism* **27**, 1176–1199 (2018). URL https://doi.org/10.1016/j.cmet.2018.05.011.

[9] Wang, Y.-P., Li, J.-T., Qu, J., Yin, M. & Lei, Q.-Y. Metabolite sensing and signaling in cancer. *The Journal of Biological Chemistry* (2020). URL http://dx.doi.org/10.1074/jbc.REV119.007624.

[10] Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics* **16**, 85–97 (2015). URL https://doi.org/10.1038/nrg3868.

[11] Johnson, C. H., Ivanisevic, J. & Siuzdak, G. Metabolomics: beyond biomarkers and towards mechanisms. *Nature Reviews Molecular Cell Biology* **17**, 451–459 (2016). URL https://doi.org/10.1038/nrm.2016.25.

[12] Collins, F. S. & Varmus, H. A new initiative on precision medicine. *New England Journal of Medicine* **372**, 793–795 (2015). URL https://doi.org/10.1056/nejmp1500523.

[13] Wellen, K. E. & Thompson, C. B. A two-way street: reciprocal regulation of metabolism and signalling. *Nature Reviews Molecular Cell Biology* **13**, 270–276 (2012). URL https://doi.org/10.1038/nrm3305.

[14] Baker, S. A. & Rutter, J. Metabolites as signalling molecules. *Nature Reviews Molecular Cell*

Biology **24**, 355–374 (2023). URL https://doi.org/10.1038/s41580-022-00572-w.

[15] Osborne, T. F. Sterol regulatory element-binding proteins (SREBPs): Key regulators of nutritional homeostasis and insulin action. *Journal of Biological Chemistry* **275**, 32379–32382 (2000). URL https://doi.org/10.1074/jbc.r000017200.

[16] Iizuka, K., Bruick, R. K., Liang, G., Horton, J. D. & Uyeda, K. Deficiency of carbohydrate response element-binding protein (ChREBP) reduces lipogenesis as well as glycolysis. *Proceedings of the National Academy of Sciences* **101**, 7281–7286 (2004). URL https://doi.org/10.1073/pnas.0401516101.

[17] Venable, M. E., Lee, J. Y., Smyth, M. J., Bielawska, A. & Obeid, L. M. Role of ceramide in cellular senescence. *Journal of Biological Chemistry* **270**, 30701–30708 (1995). URL https://doi.org/10.1074/jbc.270.51.30701.

[18] Hla, T., Lee, M.-J., Ancellin, N., Paik, J. H. & Kluk, M. J. Lysophospholipids–receptor revelations. *Science* **294**, 1875–1878 (2001). URL https://doi.org/10.1126/science.1065323.

[19] Hannun, Y. A. & Obeid, L. M. Sphingolipids and their metabolism in physiology and disease. *Nature Reviews Molecular Cell Biology* **19**, 175–191 (2017). URL https://doi.org/10.1038/nrm.2017.107.

[20] Mentch, S. J. *et al.* Histone methylation dynamics and gene regulation occur through the sensing of one-carbon metabolism. *Cell Metabolism* **22**, 861–873 (2015). URL https://doi.org/10.1016/j.cmet.2015.08.024.

[21] Wellen, K. E. *et al.* ATP-citrate lyase links cellular metabolism to histone acetylation. *Science* **324**, 1076–1080 (2009). URL http://dx.doi.org/10.1126/science.1164097.

[22] Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64–71 (1989). URL https://doi.org/10.1126/science.2675315.

[23] Kaddurah-Daouk, R., Kristal, B. S. & Weinshilboum, R. M. Metabolomics: A global biochemical approach to drug response and disease. *Annual Review of Pharmacology and Toxicology* **48**, 653–683 (2008). URL https://doi.org/10.1146/annurev.pharmtox.48.113006.094715.

[24] Schrimpe-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D. & McLean, J. A. Untargeted metabolomics strategies—challenges and emerging directions. *Journal of the American Society for Mass Spectrometry* **27**, 1897–1905 (2016). URL https://doi.org/10.1007/s13361-016-1469-y.

[25] Brodbelt, J. S. Ion activation methods for peptides and proteins. *Analytical Chemistry* **88**, 30–51 (2015). URL https://doi.org/10.1021/acs.analchem.5b04563.

[26] Berry, K. A. Z. *et al.* MALDI imaging of lipid biochemistry in tissues by mass spectrometry. *Chemical Reviews* **111**, 6491–6512 (2011). URL https://doi.org/10.1021/cr200280p.

[27] Miura, D., Fujimura, Y., Tachibana, H. & Wariishi, H. Highly sensitive matrix-assisted laser desorption ionization-mass spectrometry for high-throughput metabolic profiling. *Analytical Chemistry* **82**, 498–504 (2009). URL https://doi.org/10.1021/ac901083a.

[28] Liigand, P. *et al.* Think negative: Finding the best electrospray ionization/MS mode for your analyte. *Analytical Chemistry* **89**, 5665–5668 (2017). URL https://doi.org/10.1021/acs.analchem.7b00096.

[29] Steckel, A. & Schlosser, G. An organic chemist's guide to electrospray mass spectrometric structure elucidation. *Molecules* **24**, 611 (2019). URL https://doi.org/10.3390/molecules24030611.

[30] Schiller, J. *et al.* Lipid analysis by matrix-assisted laser desorption and ionization mass spectrometry: A methodological approach. *Analytical Biochemistry* **267**, 46–56 (1999). URL

https://doi.org/10.1006/abio.1998.3001.

[31] Knochenmuss, R. Ion formation mechanisms in UV-MALDI. *The Analyst* **131**, 966 (2006). URL https://doi.org/10.1039/b605646f.

[32] Caprioli, R. M., Farmer, T. B. & Gile, J. Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. *Analytical Chemistry* **69**, 4751–4760 (1997). URL https://doi.org/10.1021/ac970888i.

[33] Sighinolfi, G., Clark, S., Blanc, L., Cota, D. & Rhourri-Frih, B. Mass spectrometry imaging of mice brain lipid profile changes over time under high fat diet. *Scientific Reports* **11** (2021). URL https://doi.org/10.1038/s41598-021-97201-x.

[34] Tong, Y., Gao, W.-Q. & Liu, Y. Metabolic heterogeneity in cancer: An overview and therapeutic implications. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* **1874**, 188421 (2020). URL https://doi.org/10.1016/j.bbcan.2020.188421.

[35] Martinez-Outschoorn, U. E. *et al.* Oxidative stress in cancer associated fibroblasts drives tumor-stroma co-evolution: A new paradigm for understanding tumor metabolism, the field effect and genomic instability in cancer cells. *Cell Cycle* **9**, 3256–3276 (2010). URL http://dx.doi.org/10.4161/cc.9.16.12553.

[36] Zavalin, A., Yang, J., Hayden, K., Vestal, M. & Caprioli, R. M. Tissue protein imaging at 1 μm laser spot diameter for high spatial resolution and high imaging speed using transmission geometry MALDI TOF MS. *Analytical and Bioanalytical Chemistry* **407**, 2337–2342 (2015). URL https://doi.org/10.1007/s00216-015-8532-6.

[37] Prentice, B. M. *et al.* Imaging mass spectrometry enables molecular profiling of mouse and human pancreatic tissue. *Diabetologia* **62**, 1036–1047 (2019). URL https://doi.org/10.1007/s00125-019-4855-8.

[38] Rappez, L. *et al.* SpaceM reveals metabolic states of single cells. *Nature Methods* **18**, 799–805 (2021). URL https://doi.org/10.1038/s41592-021-01198-0.

[39] Hartmann, F. J. *et al.* Single-cell metabolic profiling of human cytotoxic t cells. *Nature Biotechnology* **39**, 186–197 (2020). URL https://doi.org/10.1038/s41587-020-0651-8.

[40] Kumar, A. & Misra, B. B. Challenges and opportunities in cancer metabolomics. *PROTEOMICS* **19**, 1900042 (2019). URL https://doi.org/10.1002/pmic.201900042.

[41] Ghosh, T., Philtron, D., Zhang, W., Kechris, K. & Ghosh, D. Reproducibility of mass spectrometry based metabolomics data. *BMC Bioinformatics* **22** (2021). URL https://doi.org/10.1186/s12859-021-04336-9.

[42] Lippa, K. A. *et al.* Reference materials for MS-based untargeted metabolomics and lipidomics: a review by the metabolomics quality assurance and quality control consortium (mQACC). *Metabolomics* **18** (2022). URL https://doi.org/10.1007/s11306-021-01848-6.

[43] Wehrens, R. *et al.* Improved batch correction in untargeted MS-based metabolomics. *Metabolomics* **12** (2016). URL https://doi.org/10.1007/s11306-016-1015-8.

[44] Hrydziuszko, O. & Viant, M. R. Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. *Metabolomics* **8**, 161–174 (2011). URL https://doi.org/10.1007/s11306-011-0366-4.

[45] Tu, T., Sauter, A. D., Sauter, A. D. & Gross, M. L. Improving the signal intensity and sensitivity of MALDI mass spectrometry by using nanoliter spots deposited by induction-based fluidics. *Journal of the American Society for Mass Spectrometry* **19**, 1086–1090 (2008). URL https://doi.org/10.1016/j.jasms.2008.03.017.

[46] Ellis, S. R., Bruinen, A. L. & Heeren, R. M. A. A critical evaluation of the current state-of-the-art in quantitative imaging mass spectrometry. *Analytical and Bioanalytical Chemistry* **406**,

1275–1289 (2013). URL https://doi.org/10.1007/s00216-013-7478-9.

[47] Perry, W. J. *et al.* Uncovering matrix effects on lipid analyses in MALDI imaging mass spectrometry experiments. *Journal of Mass Spectrometry* **55**, e4491 (2020). URL https://doi.org/10.1002/jms.4491.

[48] Palmer, A. *et al.* FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nature Methods* **14**, 57–60 (2016). URL https://doi.org/10.1038/nmeth.4072.

[49] Webb-Robertson, B.-J. M. *et al.* Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *Journal of Proteome Research* **14**, 1993–2001 (2015). URL https://doi.org/10.1021/pr501138h.

[50] Southam, A. D., Payne, T. G., Cooper, H. J., Arvanitis, T. N. & Viant, M. R. Dynamic range and mass accuracy of wide-scan direct infusion nanoelectrospray fourier transform ion cyclotron resonance mass spectrometry-based metabolomics increased by the spectral stitching method. *Analytical Chemistry* **79**, 4595–4602 (2007). URL https://doi.org/10.1021/ac062446p.

[51] Rubin, D. B. INFERENCE AND MISSING DATA. *ETS Research Bulletin Series* **1975**, i–19 (1975). URL https://doi.org/10.1002/j.2333-8504.1975.tb01053.x.

[52] Lin, D. *et al.* An integrative imputation method based on multi-omics datasets. *BMC Bioinformatics* **17** (2016). URL https://doi.org/10.1186/s12859-016-1122-6.

[53] Li, M. & Smyth, G. K. Neither random nor censored: estimating intensity-dependent probabilities for missing values in label-free proteomics. *Bioinformatics* **39** (2023). URL https://doi.org/10.1093/bioinformatics/btad200.

[54] Do, K. T. *et al.* Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics* **14** (2018). URL https://doi.org/10.1007/s11306-018-1420-2.

[55] Armitage, E. G., Godzien, J., Alonso-Herranz, V., López-Gonzálvez, Á. & Barbas, C. Missing value imputation strategies for metabolomics data. *ELECTROPHORESIS* **36**, 3050–3060 (2015). URL https://doi.org/10.1002/elps.201500352.

[56] Xia, J., Psychogios, N., Young, N. & Wishart, D. S. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Research* **37**, W652–W660 (2009). URL https://doi.org/10.1093/nar/gkp356.

[57] Walczak, B. & Massart, D. Dealing with missing data. *Chemometrics and Intelligent Laboratory Systems* **58**, 15–27 (2001). URL https://doi.org/10.1016/s0169-7439(01)00131-9.

[58] Little, R. & Rubin, D. *Statistical Analysis with Missing Data, Third Edition* (Wiley, 2019). URL https://doi.org/10.1002/9781119482260.

[59] Hou, W., Ji, Z., Ji, H. & Hicks, S. C. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biology* **21** (2020). URL https://doi.org/10.1186/s13059-020-02132-x.

[60] Wei, R. *et al.* Missing value imputation approach for mass spectrometry-based metabolomics data. *Scientific Reports* **8** (2018). URL https://doi.org/10.1038/s41598-017-19120-0.

[61] Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001). URL https://doi.org/10.1093/bioinformatics/17.6.520.

[62] van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* **45** (2011). URL https://doi.org/10.18637/jss.v045.i03.

[63] Agarwal, D., Wang, J. & Zhang, N. R. Data denoising and post-denoising corrections in single cell RNA sequencing. *Statistical Science* **35** (2020). URL https://doi.org/10.1214/19-sts7560.

[64] Patruno, L. *et al.* A review of computational strategies for denoising and imputation of single-cell transcriptomic data. *Briefings in Bioinformatics* (2020). URL https://doi.org/10.1093/bib/bbaa222.

[65] Qiu, P. Embracing the dropouts in single-cell RNA-seq analysis. *Nature Communications* **11** (2020). URL https://doi.org/10.1038/s41467-020-14976-9.

[66] van Dijk, D. *et al.* Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729.e27 (2018). URL https://doi.org/10.1016/j.cell.2018.05.061.

[67] Linderman, G. C. *et al.* Zero-preserving imputation of single-cell RNA-seq data. *Nature Communications* **13** (2022). URL https://doi.org/10.1038/s41467-021-27729-z.

[68] Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications* **10** (2019). URL https://doi.org/10.1038/s41467-018-07931-2.

[69] Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nature Communications* **9** (2018). URL https://doi.org/10.1038/s41467-018-03405-7.

[70] Mongia, A., Sengupta, D. & Majumdar, A. McImpute: Matrix completion based imputation for single cell RNA-seq data. *Frontiers in Genetics* **10** (2019). URL https://doi.org/10.3389/fgene.2019.00009.

[71] Huang, M. *et al.* SAVER: gene expression recovery for single-cell RNA sequencing. *Nature Methods* **15**, 539–542 (2018). URL https://doi.org/10.1038/s41592-018-0033-z.

[72] Wang, J. *et al.* Data denoising with transfer learning in single-cell transcriptomics. *Nature Methods* **16**, 875–878 (2019). URL https://doi.org/10.1038/s41592-019-0537-1.

[73] Wishart, D. S. *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research* **46**, D608–D617 (2017). URL https://doi.org/10.1093/nar/gkx1089.

[74] Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* **8**, 118–127 (2006). URL https://doi.org/10.1093/biostatistics/kxj037.

[75] Polański, K. *et al.* BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964–965 (2019). URL https://doi.org/10.1093/bioinformatics/btz625.

[76] Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65 (1987). URL https://doi.org/10.1016/0377-0427(87)90125-7.

[77] Calinski, T. & Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods* **3**, 1–27 (1974). URL https://doi.org/10.1080/03610927408827101.

[78] Davies, D. L. & Bouldin, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-1**, 224–227 (1979). URL https://doi.org/10.1109/tpami.1979.4766909.

[79] Rose, T. D. *et al.* Lipid network and moiety analysis for revealing enzymatic dysregulation and mechanistic alterations from lipidomics data. *Briefings in Bioinformatics* **24** (2023). URL https://doi.org/10.1093/bib/bbac572.

[80] Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19** (2018). URL https://doi.org/10.1186/s13059-017-1382-0.

[81] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[82] Liu, M. & Dongre, A. Proper imputation of missing values in proteomics datasets for differential

expression analysis. *Briefings in Bioinformatics* **22** (2020). URL https://doi.org/10.1093/bib/bbaa112.

[83] Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology* **4** (2005). URL https://doi.org/10.2202/1544-6115.1128.

[84] Cao, Y., Yang, P. & Yang, J. Y. H. A benchmark study of simulation methods for single-cell RNA sequencing data (2021). URL https://doi.org/10.1101/2021.06.01.446157.

[85] Song, D. *et al.* scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nature Biotechnology* (2023). URL https://doi.org/10.1038/s41587-023-01772-1.

[86] Li, R., Li, L., Xu, Y. & Yang, J. Machine learning meets omics: applications and perspectives. *Briefings in Bioinformatics* **23** (2021). URL https://doi.org/10.1093/bib/bbab460.

[87] Hediyeh-zadeh, S., Webb, A. I. & Davis, M. J. MsImpute: Estimation of missing peptide intensity data in label-free quantitative mass spectrometry. *Molecular & Cellular Proteomics* 100558 (2023). URL https://doi.org/10.1016/j.mcpro.2023.100558.

[88] Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature Methods* **15**, 1053–1058 (2018). URL https://doi.org/10.1038/s41592-018-0229-2.

[89] Choi, M. *et al.* MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* **30**, 2524–2526 (2014). URL https://doi.org/10.1093/bioinformatics/btu305.

[90] Zhang, X. *et al.* Proteome-wide identification of ubiquitin interactions using UbIA-MS. *Nature Protocols* **13**, 530–550 (2018). URL https://doi.org/10.1038/nprot.2017.147.

[91] Breda, J., Zavolan, M. & van Nimwegen, E. Bayesian inference of the gene expression states of single cells from scRNA-seq data (2019). URL https://doi.org/10.1101/2019.12.28.889956.

[92] Camacho, D., de la Fuente, A. & Mendes, P. The origin of correlations in metabolomics data. *Metabolomics* **1**, 53–63 (2005). URL https://doi.org/10.1007/s11306-005-1107-3.

[93] Lee, J. M., Gianchandani, E. P., Eddy, J. A. & Papin, J. A. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS Computational Biology* **4**, e1000086 (2008). URL https://doi.org/10.1371/journal.pcbi.1000086.

[94] Benedetti, E. *et al.* A strategy to incorporate prior knowledge into correlation network cutoff selection. *Nature Communications* **11** (2020). URL https://doi.org/10.1038/s41467-020-18675-3.

[95] Steinheuer, L. M., Canzler, S. & Hackermüller, J. Benchmarking scRNA-seq imputation tools with respect to network inference highlights deficits in performance at high levels of sparsity (2021). URL https://doi.org/10.1101/2021.04.02.438193.

[96] Lusczek, E. R., Lexcen, D. R., Witowski, N. E., Mulier, K. E. & Beilman, G. Urinary metabolic network analysis in trauma, hemorrhagic shock, and resuscitation. *Metabolomics* **9**, 223–235 (2012). URL https://doi.org/10.1007/s11306-012-0441-5.

[97] Smith, H. B., Kim, H. & Walker, S. I. Scarcity of scale-free topology is universal across biochemical networks. *Scientific Reports* **11** (2021). URL https://doi.org/10.1038/s41598-021-85903-1.

[98] Garciarena, U. & Santana, R. An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. *Expert Systems with Applications* **89**, 52–65 (2017). URL https://doi.org/10.1016/j.eswa.2017.07.026.

Appendix A

Supplementary material

Table S.1: Parameters of ion/cell filtering. This table lists the characteristics of the included datasets before and after filtering with the selected thresholds. Incidently, for three of the datasets, close to 350 features are sufficient to maintain cellular population structure.

	seahorse	**glioblastoma**	**pancreatic cancer**	**HepaRG**
Raw dropout rate	71.3%	79.0%	86.6%	71.1%
Raw size (cells x ions)	9.908 x 966	10.848 x 1.503	15.685 x 1.667	29.738 x 740
Filter unique ions/cells	20% / 10%	5% / 5%	5% / 20%	20% / 20%
Filtered dropout rate	35.8%	62.6%	49.8%	39.2%
Filtered size (cells x ions)	7.926 x 348	10.716 x 835	15.593 x 353	27.229 x 314

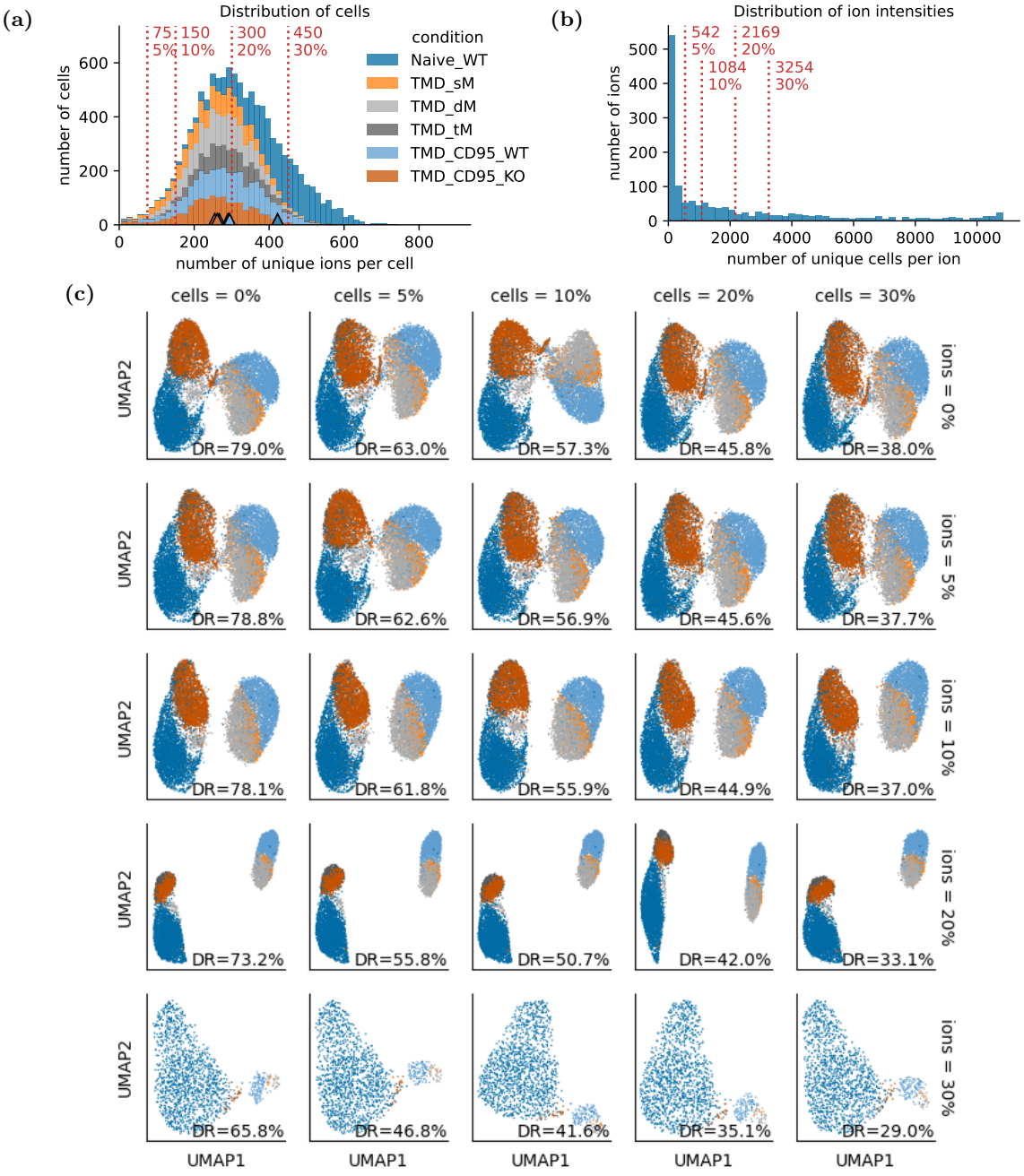

Figure S.1: Systematic filtering on the glioblastoma dataset. **(a)** Systematic thresholds remove cells with fewer unique ions than specified in red (absolute number and fraction). Cells are colored by their affiliation to conditions. Colored triangles at the bottom indicate the respective mean number of unique ions per condition. **(b)** Analogous, ions are removed if detected in less cells than specified in different thresholds. **(c)** Population structure and dropout rate (DR) of the dataset after different filtering thresholds. Rows correspond to (a), columns to (b).

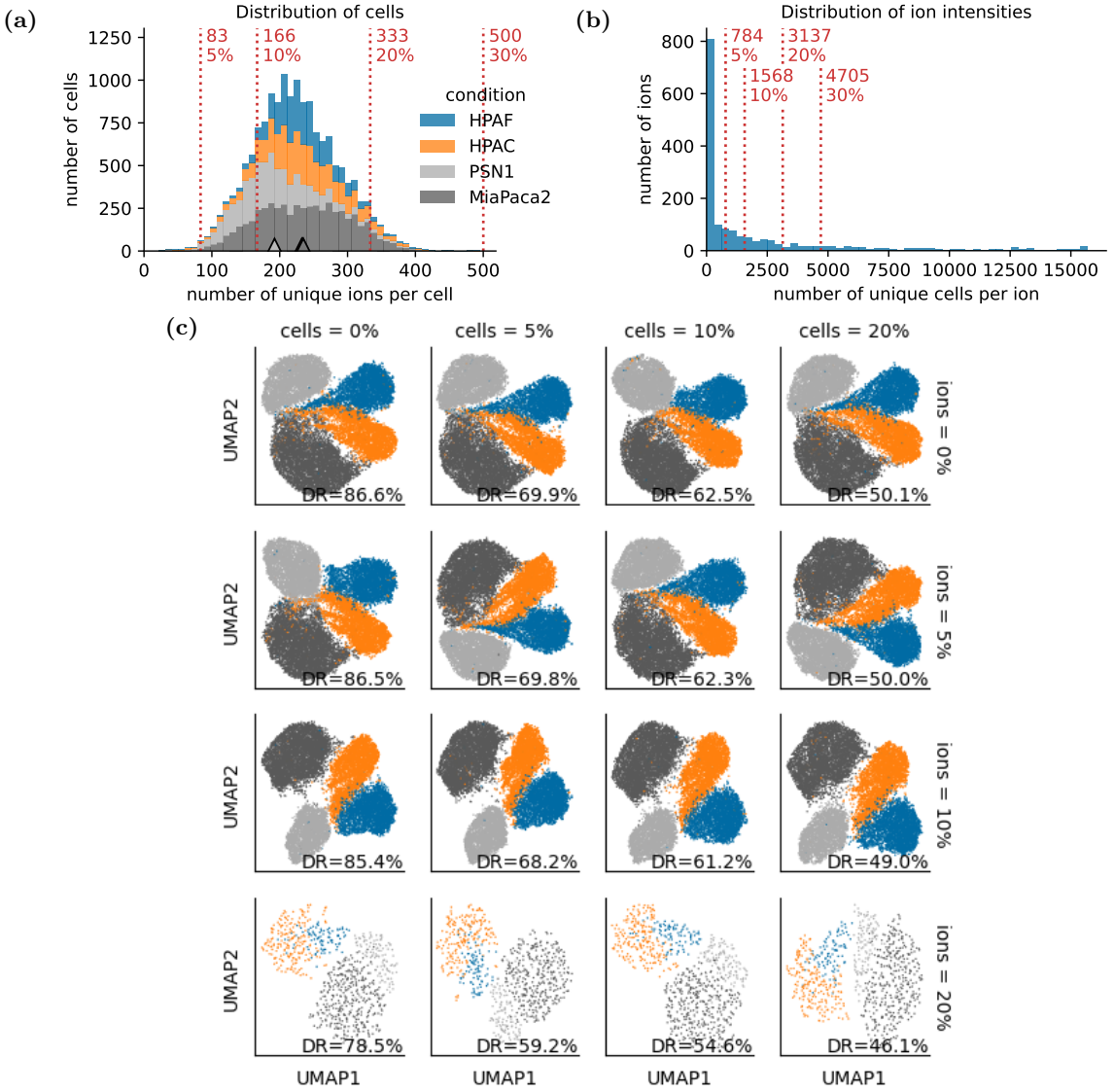

Figure S.2: Systematic filtering on the pancreatic cancer dataset. (a) Systematic thresholds remove cells with fewer unique ions than specified in red (absolute number and fraction). Cells are colored by their affiliation to conditions. Colored triangles at the bottom indicate the respective mean number of unique ions per condition. **(b)** Analogous, ions are removed if detected in less cells than specified in different thresholds. **(c)** Population structure and dropout rate (DR) of the dataset after different filtering thresholds. To mitigate batch effects, the embedding was integrated using `BBKNN`. Rows correspond to (a), columns to (b). Cell and ion thresholds of 30% removed all ions and cells, respectively, thus they are not shown.

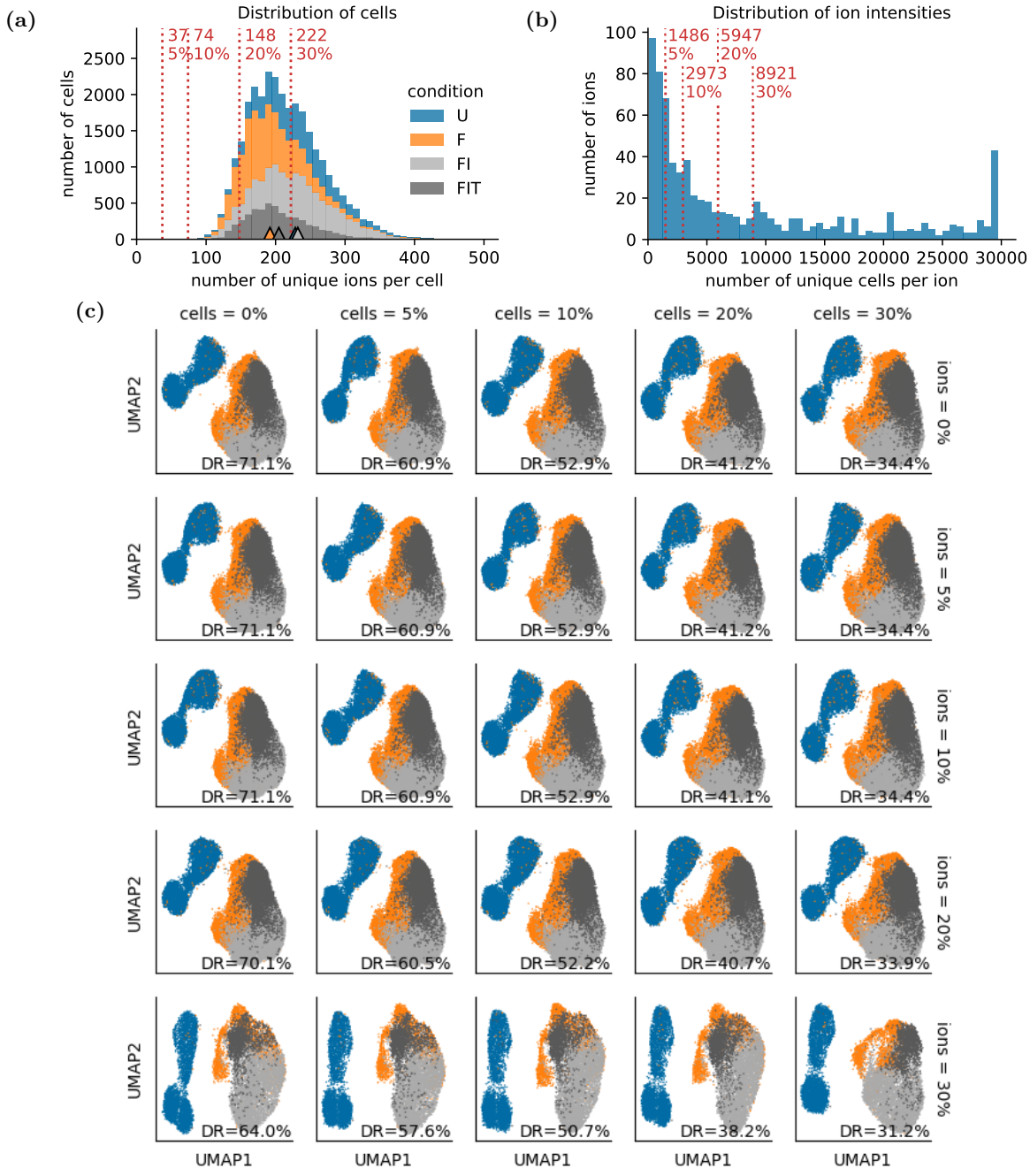

Figure S.3: Systematic filtering on the HepaRG dataset. (a) Systematic thresholds remove cells with fewer unique ions than specified in red (absolute number and fraction). Cells are colored by their affiliation to conditions. Colored triangles at the bottom indicate the respective mean number of unique ions per condition. **(b)** Analogous, ions are removed if detected in less cells than specified in different thresholds. **(c)** Population structure and dropout rate (DR) of the dataset after different filtering thresholds. Rows correspond to (a), columns to (b).

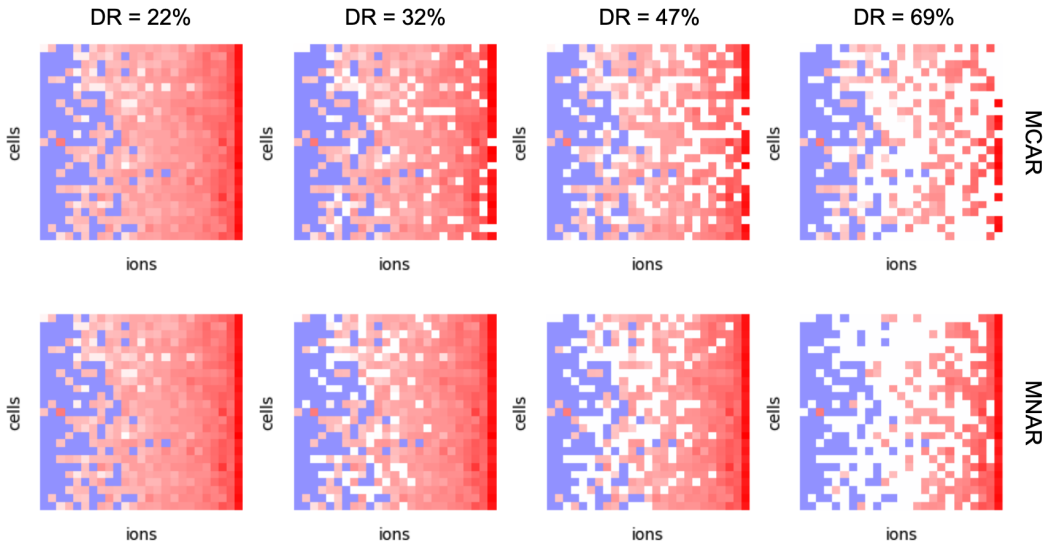

Figure S.4: Visualization of different dropout simulation mechanisms. The figure shows the different effects of simulating dropouts completely at random (MCAR) and not at random (MNAR) in a dataset with preexisting missing values. Every square represents a data matrix with cells in rows and ions in columns, where columns are ordered by the mean ion intensity from smallest to largest. The four panels in each row show increasing levels of simulated missing values, starting from the baseline dropout rate of the data matrix, 22%. Red pixels symbolize detected intensities, higher opacity indicating higher values. Blue pixels are missing values from the acquisition while white pixels stand for missing values by simulation. Clearly, simulation of MCAR affects all values similarly, regardless of the intensities. In contrast, MNAR simulation removes high intensities to a much smaller extent than low intensities.

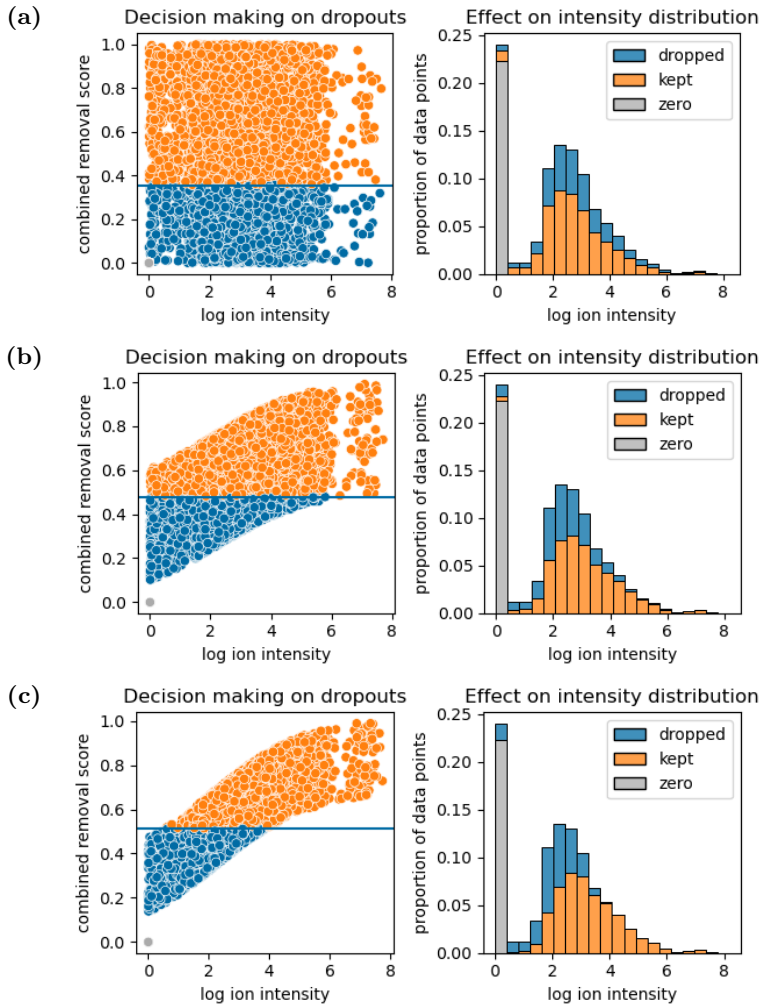

Figure S.5: Simulation process of dropouts. This figure visualizes the process of simulating a dropout rate (DR) of 0.5 on a reduced version of the seahorse dataset. The left plots show the combination of random and deterministic parts to a combined removal decision score (blue horizontal line) and how that score is related to the corresponding log ion intensities. This scatterplot only displays a random subsample of 10k datapoints. The right plots show how the dropout simulation affects the distribution of all ion intensities in the dataset. **(a)** Intensity values are removed purely based on a random process. **(b)** The original intensity values take the same importance as the random values (1:1). **(c)** Original intensities are twice as important as random values (2:1).

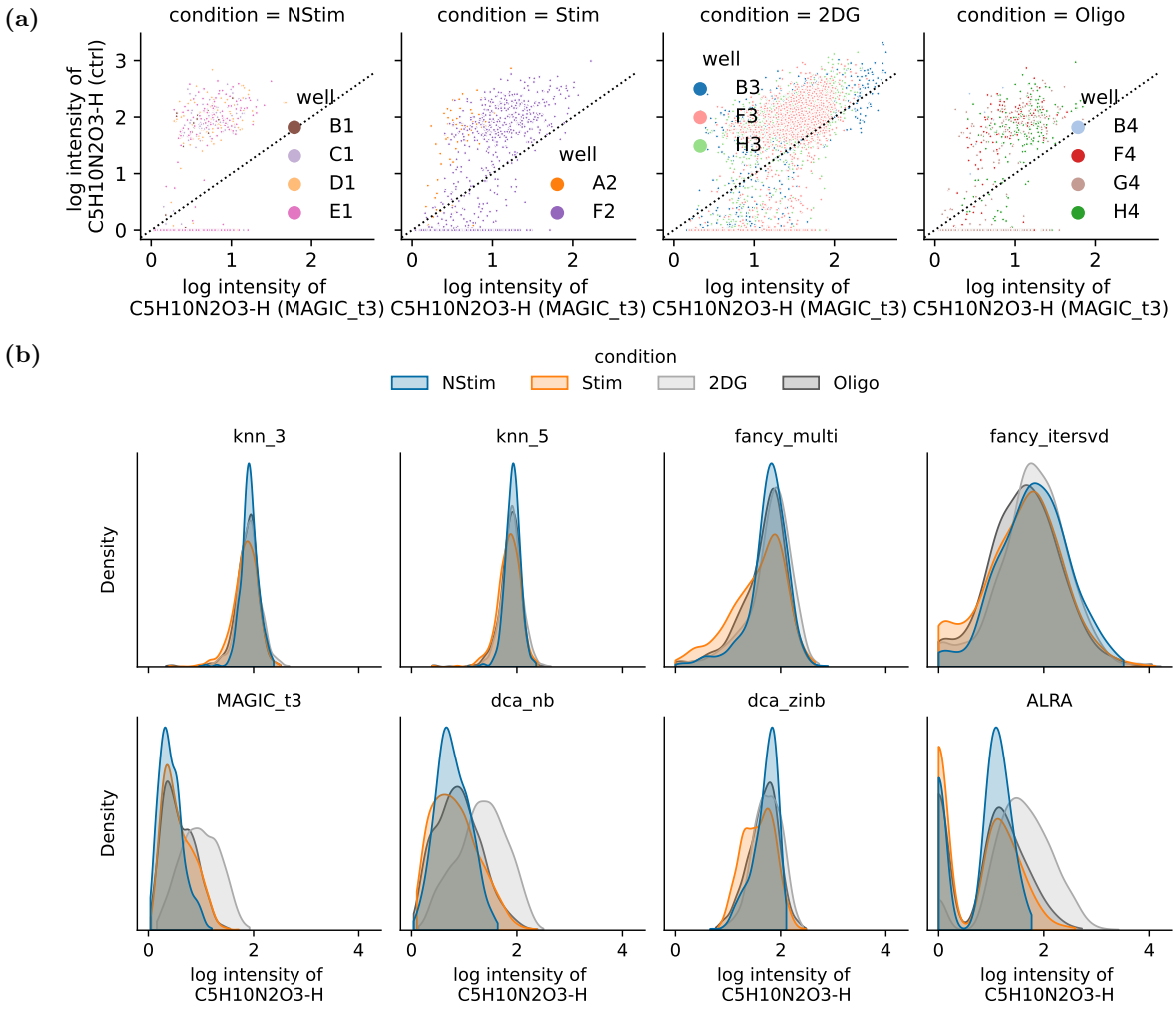

Figure S.6: Differential denoising effects in the seahorse dataset. **(a)** The `MAGIC_t3` panel from Figure 5a is separated by conditions and colored by individual wells to show that denoising variability occurs only on the level of conditions, not on the level of replicates. **(b)** Distribution of the ion intensities introduced by the given methods to replace missing values. While the distributions for different conditions overlap for kNN imputation, they are shifted under the different denoising methods.

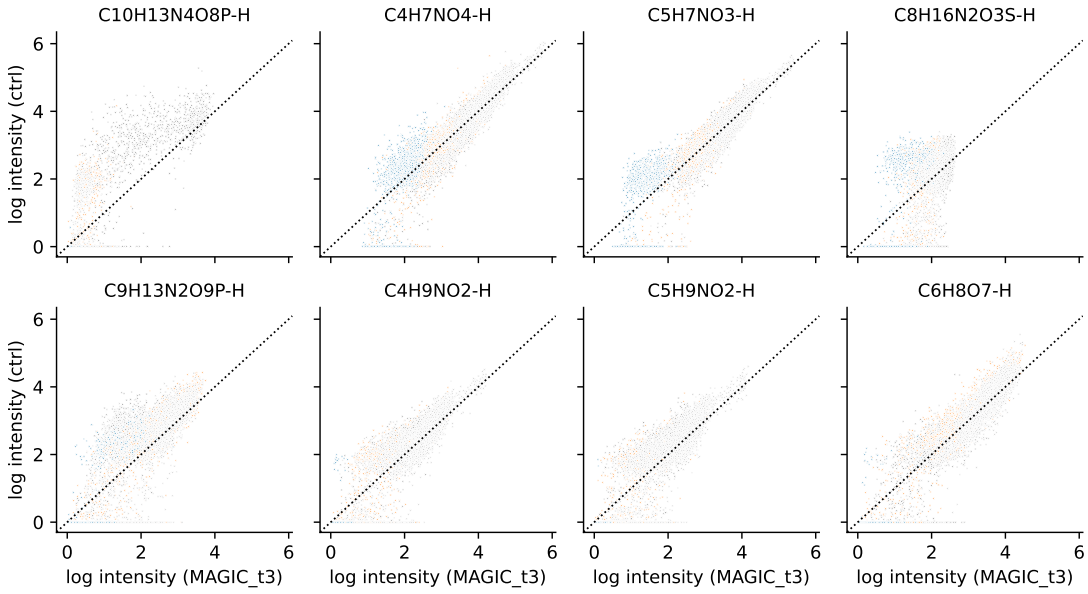

Figure S.7: Differential denoising of other ions in the seahorse dataset. Other ions show varying effects of MAGIC denoising on cells from different conditions of the seahorse dataset. Intensities below the black identity lines are increased by MAGIC, intensities above it are decreased. DCA and ALRA exhibit similar results (not shown). The colors from (b) apply. Possible annotations of the top row: Inosinic acid, aspartic acid, L-pyroglutamic acid, Met-Ala/Cys-Val dipeptide; bottom row: uridine mono-phosphate, γ-aminobutyric acid (GABA), proline, citric acid/isocitric acid

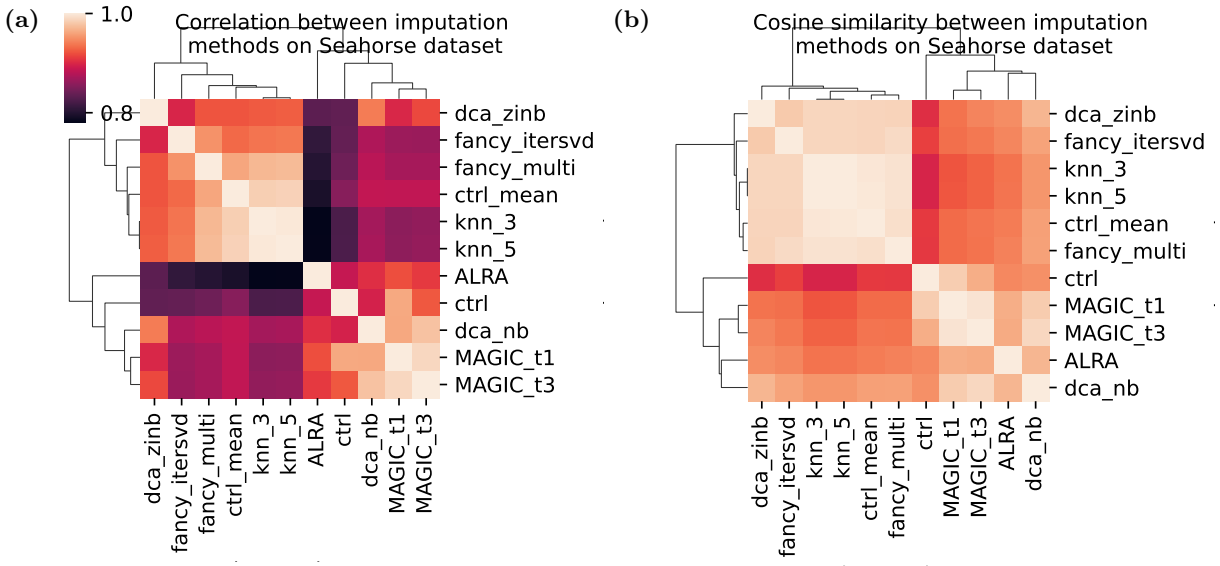

Figure S.8: Overall similarities between imputed data matrices (seahorse). **(a)** All ion intensities before and after imputation (no simulation of dropouts) are converted into vectors to calculate pairwise Pearson correlations. The resulting heatmap shows which methods create similar data. Hierarchical clustering assigns imputation and denoising methods to separate clusters. Only `dca_zinb` is an exception, clustering with otherwise imputation methods. In contrast, the non-imputed data (`ctrl`) is co-localized with denoising methods. **(b)** Analogous to (a), pairwise cosine similarity is used to visualize the similarity of data after imputation/denoising. Although less pronounced, the same separation can be observed. Both heatmaps use the same color scale, determined by the smallest correlation/similarity.

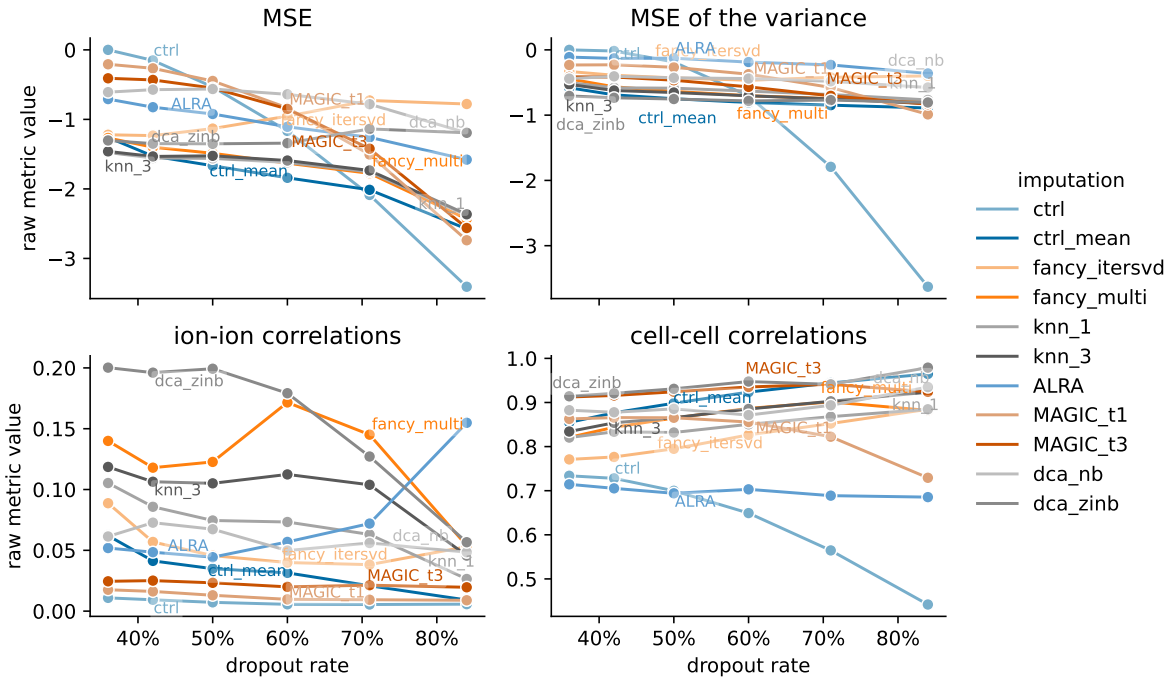

Figure S.9: Overview of performance metrics on the seahorse dataset. After simulating different levels of missing values in an MNAR pattern, the dataset is processed with a variety of imputation/denoising methods. The `ctrl` curve represents a baseline of only simulation. Some metrics are inverted for easier comparison. Apart from ion-ion correlations, all metrics are maximized in the ranking. mean square error (MSE) measures the deviation in the individual intensities and MSE of variance the change in ion variability. The correlation-based scores average the pairwise ion and cell correlations to quantify the linear dependance in the imputed data matrices

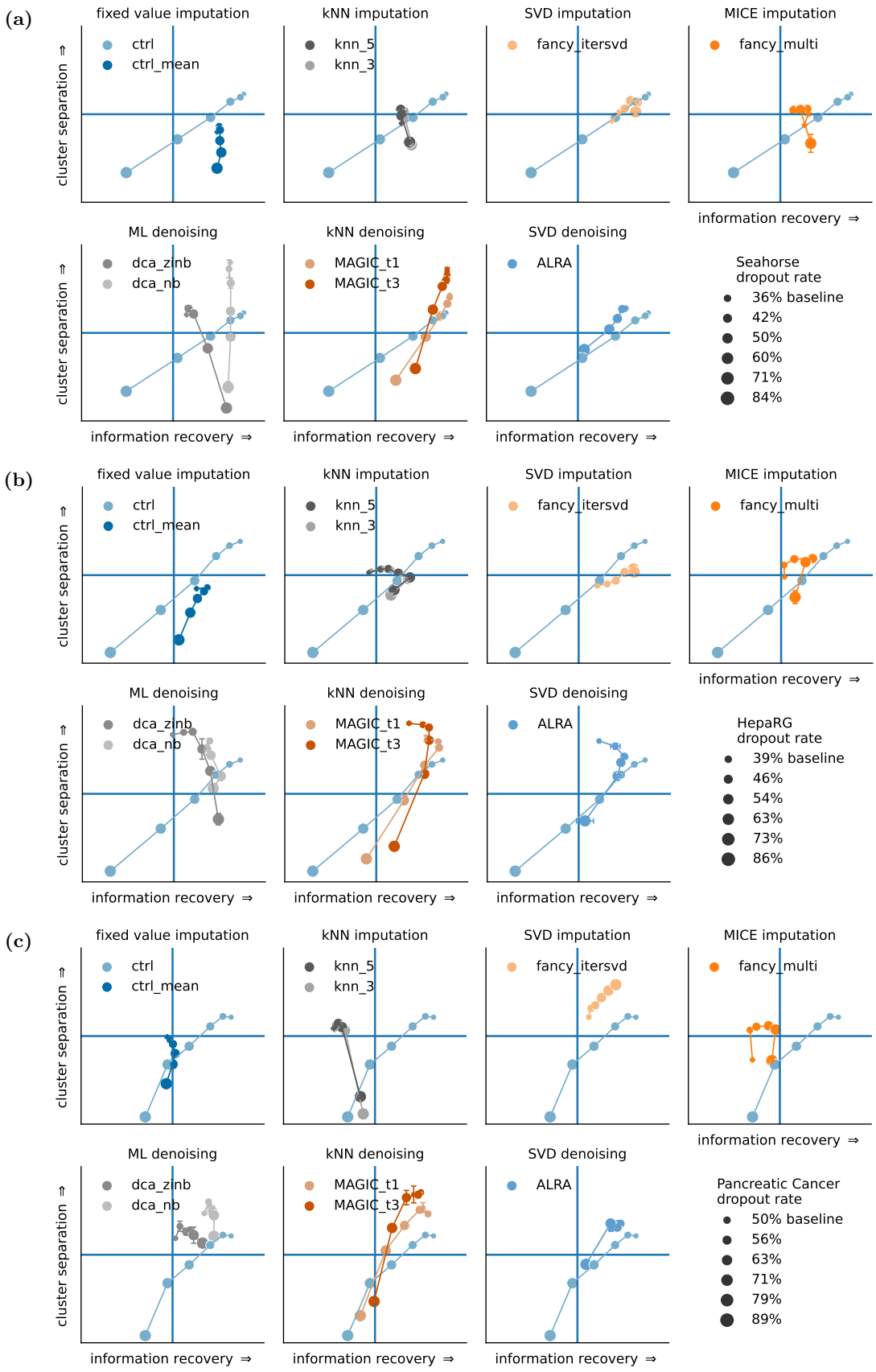

Figure S.10: Performance of imputation/denoising on the other datasets. Analogous to Figure 7, this Figure shows the 2x2 performance matrices for the **(a)** seahorse, **(b)** HepaRG, and **(c)** pancreatic cancer dataset.

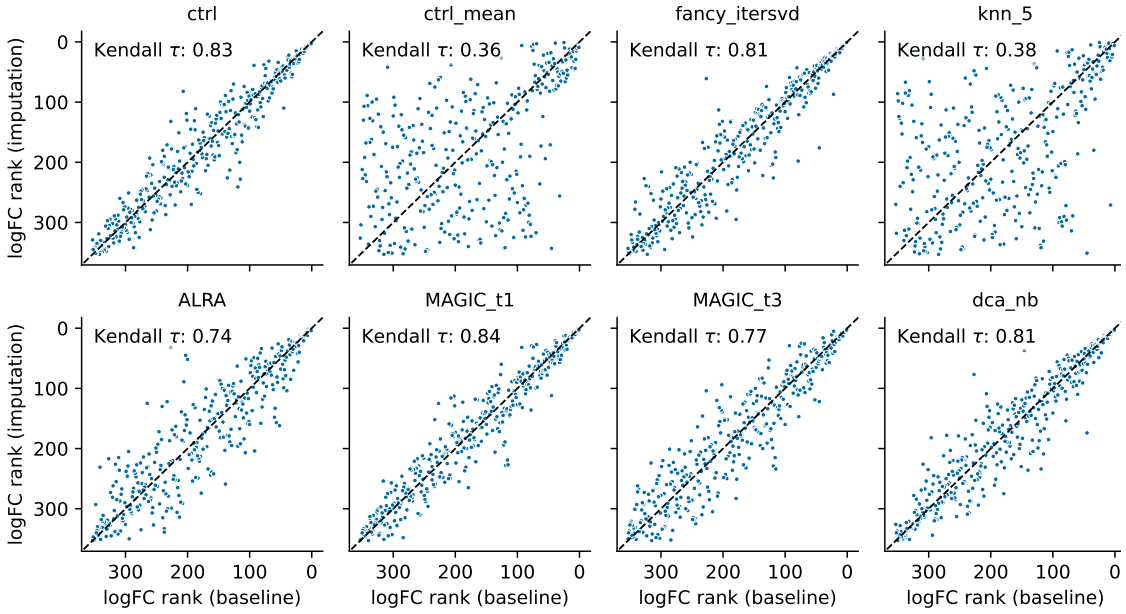

Figure S.11: Exemplary relationships of ranked ion log-fold changes. Shown are the ranks of ions by log-fold change of mean intensity in the HPAF condition compared to the three other conditions of the pancreatic cancer dataset (with simulated dropouts). Compared are different imputation/denoising methods (including ctrl: only simulation, no imputation) to the baseline data before simulation of missing values. A black dashed identity line marks a perfect agreement of the fold changes. A rank close to 1 symbolizes that an ion is highly upregulated, while ions with very high ranks are strongly downregulated in the HPAF condition.

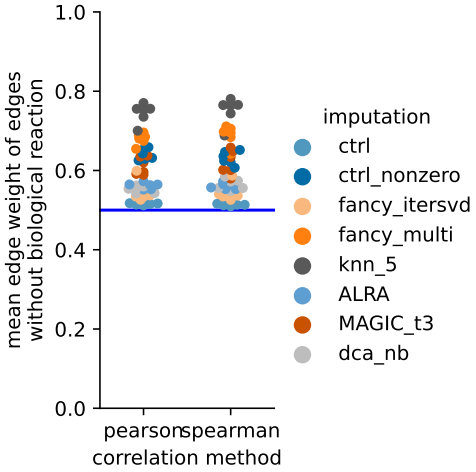

Figure S.12: Effects of imputation/denoising on ion correlations in the glioblastoma dataset. This figure shows that across imputation and denoising methods, the distributions of ion-ion correlations without a biological reaction remain centered around a positive correlation.

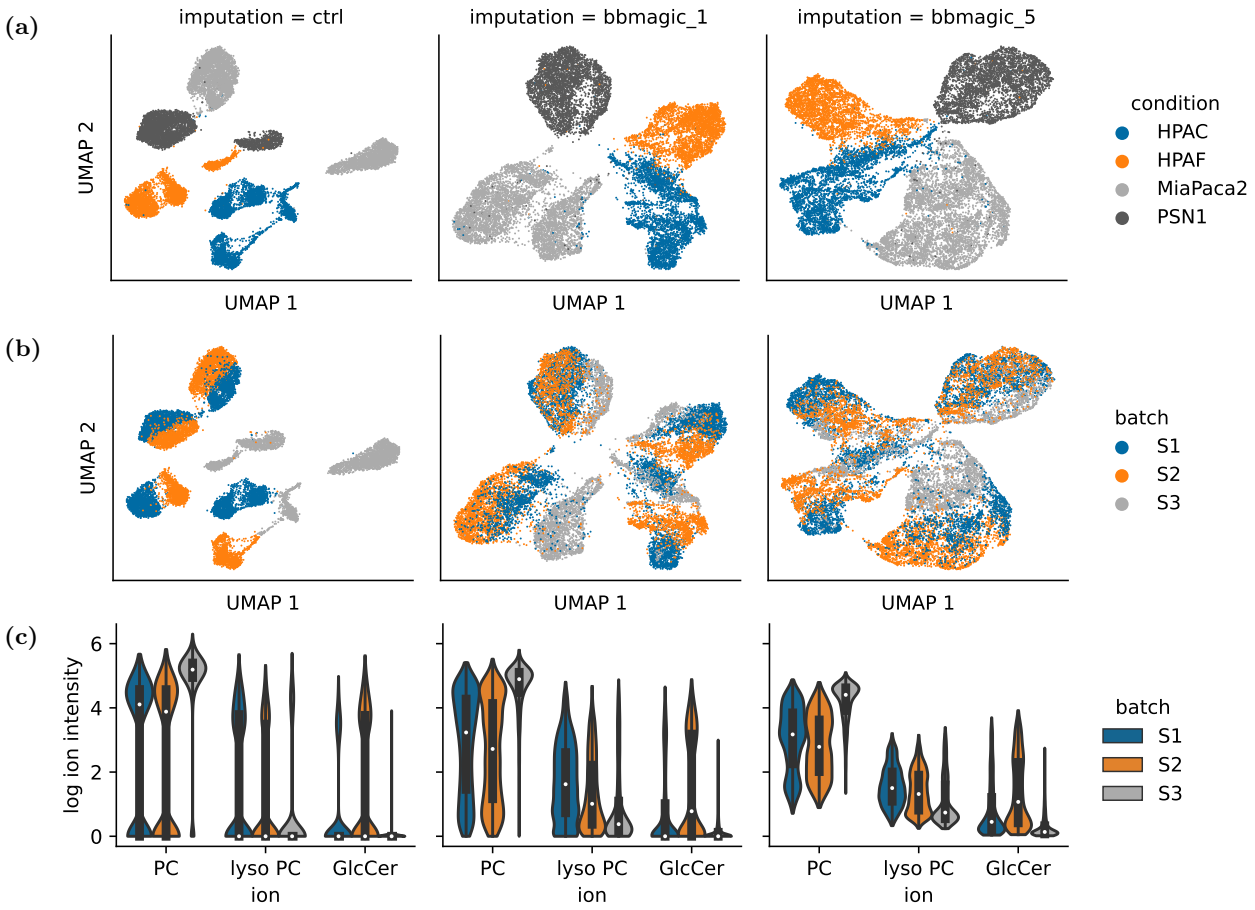

Figure S.13: Increasing effect of bbMAGIC on the pancreatic cancer dataset. Using the batch-balanced neighborhood graph, MAGIC is applied iteratively to a dataset. Thus, the degree of batch correction can be controlled manually. The effects of a low (middle column) and high iteration value (right column) are shown with respect to **(a)** separation of conditions, **(b)** integration of batches and, **(c)** convergence of distributions. The same ions are shown as in Figure 14.