Dissertation zur Erlangung des Doktorgrades

der Fakultät für Chemie und Pharmazie

der Ludwig-Maximilians-Universität München

Quantification algorithms for multiplexed data-independent acquisition data in shotgun proteomics

von

Dmitry Alexeev

aus

Krasnodar, Russia

2025

Erklärung

Diese Dissertation wurde im Sinne von § 7 der Promotionsordnung vom 28. November 2011 von Frau Prof. Dr. Johanna Klughammer betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, 15.10.2025

.....

Dmitry Alexeev

Dissertation eingereicht am 15.10.2025

1. Gutachter: Prof. Dr. Johanna Klughammer

2. Gutachter: Prof. Dr. Maria Robles

Mündliche Prüfung am 16.01.2026

SUMMARY

Proteomics is a rapidly growing research field. Recent advancements towards data-independent acquisition (DIA) are overcoming limitations of the data-dependent approaches (DDA), such as the limited amount of analyzed ions per run and the exclusion of fragmentation signals from the quantification process. Nevertheless, techniques developed at first to boost the DDA performance can be beneficial to the DIA workflow. As such, multiplexing techniques found their way into the newly emerging field of DIA proteomics. The concept of multiplexing implies the use of labels – be it isotopically labeled amino acids or chemical groups attached to the amino acids during sample preparation. By marking proteins from each sample via its unique label, researchers can analyze them simultaneously in a single run. While improving the throughput of the experiment, multiplexing can also boost the identification rate by transferring identifications between channels.

This study's goal is to provide a computational platform based on the established MaxDIA workflow to analyze multiplexed DIA proteomics samples termed MultiplexDIA. Moreover, the existing MaxLFQ algorithm was generalized to account for multiplexed signals for quantification. The results of this study highlight the algorithm's structure and its performance against the DIA-NN software on datasets with normal and single-cell-like amounts of proteomes in the samples. Different types of labels, as well as different modes of identification transfer between labels, were tested.

Our results suggest that MultiplexDIA is able to achieve similar levels of improvement as DIA-NN in identification rate, quantification accuracy, and false discovery rate (FDR) control when comparing results of the multiplexed and label-free analysis. The complete MS1 multiplex transfer proves to be the most resilient mode of identification transfer between labels, allowing for additional identifications without impairment of the quantification. The reworked MaxLFQ can normalize and quantify each multiplexed channel separately, thus retaining more capabilities for throughput. Both metabolic and chemical non-isobaric labels are suitable for the MultiplexDIA workflow, though the former show better performance.

Table of contents

Abbreviations

Abbreviation	Full name
CID	Collision-Induced Dissociation
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
FC	Fold-change
FDR	False discovery rate
HCD	Higher-Energy Collisional Dissociation
HPLC	High-performance liquid chromatography
ID	Identification
IM	Ion mobility
IQR	Inter-quantile range
LC-MS	Liquid Chromatography-Tandem Mass Spectrometry
LFQ	Label-free quantification
Log	Logarithm
MS1	Full-scan precursor ion spectrum
MS2 (MS/MS)	Fragmentation spectrum
mTRAQ	Mass differential tags for relative and absolute quantification
PASEF	Parallel Accumulation–Serial Fragmentation
ROC	Receiver operating characteristic
RT	Retention time
SILAC	Stable isotopic labeled amino acids
TIMS	Trapped ion mobility spectrometry
TOF	Time-of-flight analyzer
XIC	Extracted ion current

1 Introduction

1.1 Advancements in proteomics

Proteins play a central role in virtually all biological processes, acting as enzymes, structural components, signaling molecules, and regulators of cellular functions. Understanding their structure, function, and interactions is crucial for deciphering the molecular mechanisms underlying health and disease [1]. The evolution of protein research reached levels of high-throughput quantitative analysis of tens of thousands of proteins per sample due to the development of bottom-up 'shotgun-proteomics' – an approach utilizing a breakdown of proteins into peptides via enzymatic digestion, separation through high-performance liquid chromatography (HPLC), electrospray ionization, and subsequent detection of ions in the mass spectrometer [2]. Identified peptides are assembled into full protein sequences, hence the name 'bottom-up'. Another name for this approach would be Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS).

The common principle of all mass spectrometers used in proteomics is the separation of ions based on their mass-to-charge (m/z) ratio [3, 4]. The initial ion's m/z, retention time in the HPLC, and its intensity produce a peak in the MS1 spectrum. Normally, MS1 ions are further fragmented into smaller ions, resulting in MS2 spectra. There are several types of ion fragmentation implemented in proteomics, the most common being Higher-Energy Collisional Dissociation (HCD) and Collision-Induced Dissociation (CID) [5, 6]. Both are tuned to cleave the amide bond of the peptides to generate a series of b- and y-ions, where b-ions always contain the N-terminus of the original peptide and y-ions – the C-terminus. Every software that identifies and quantifies the mass spectrometry signal in proteomics analyzes MS1- and MS2-level spectra [7]. However, the structure of that signal will vary drastically, depending on the type of mass spectrometer and the data acquisition mode used.

1.2 Mass spectrometers

1.2.1 Orbitrap analyzer

The Orbitrap Q Exactive mass spectrometer model, used to acquire one of the datasets in the scope of this thesis, consists of the following major structural elements (Figure 1.1) [8, 9]:

1. RF-lens (S-lens): a stacked ring ion guide, which focuses ions into an ion beam and increases the sensitivity of the incoming signal;
2. Inject flatapole and bent flatapole: an ion transmission part that filters out any uncharged molecules that cannot follow the bent beam;
3. Quadrupole: a four-rod structure able to filter out ions of particular m/z (resonant ions) by applying a fluctuating electric field to the rods;
4. C-trap: an ion storage device, allowing for the accumulation of ions before injecting them into the analyzer. It facilitates the use of a pulse-operating Orbitrap analyzer alongside a continuous source of ions like electrospray;
5. Orbitrap mass analyzer;
6. HCD fragmentation cell: a quadrupole, tuned for collision dissociation of ions.

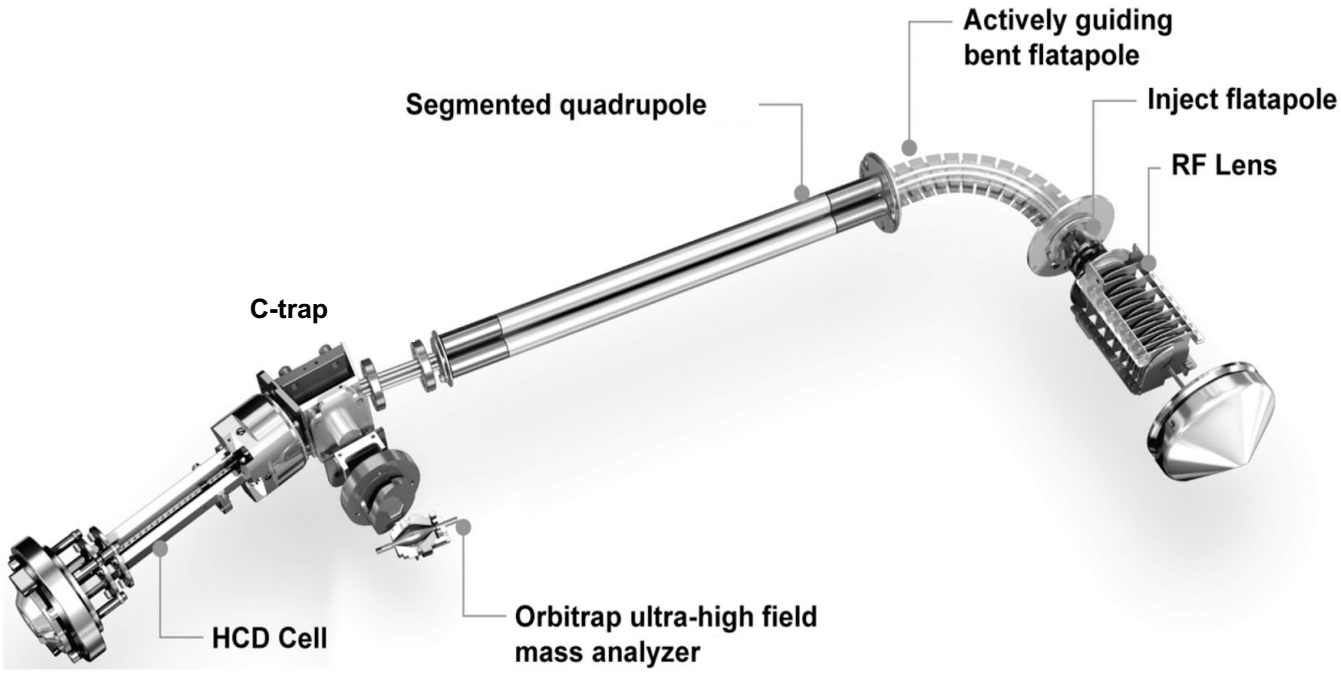
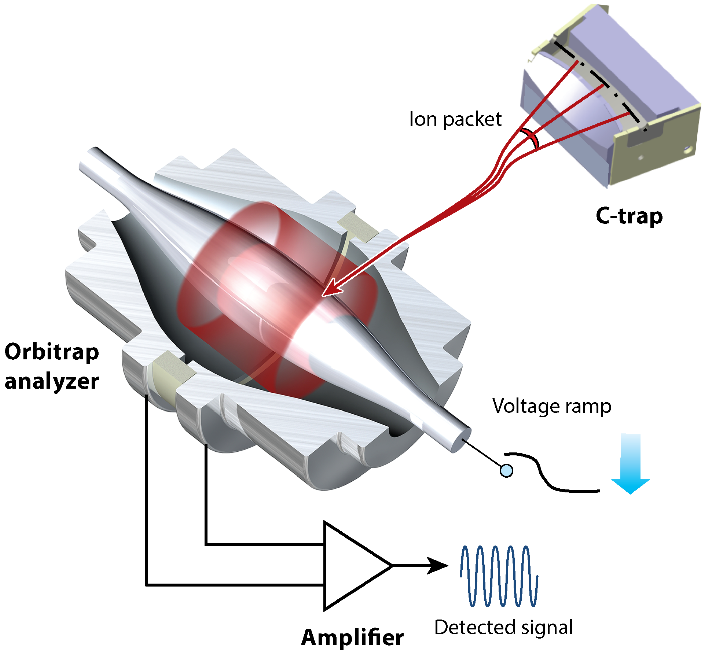


Figure 1.1: Scheme of the Q Exactive HF instrument. Electrospray-ionized molecules enter the instrument through the focusing RF-lens. Uncharged and contaminated species are filtered through the system of flatapoles. In the full-scan MS1 mode, all ions pass an inactive quadrupole into the C-trap and Orbitrap analyzer. In the MS2 mode, the quadrupole filters a specific m/z range of ions before sending them into the C-trap. Before entering the Orbitrap, ions are sent for fragmentation into the HCD cell. Adopted from [9].

The Orbitrap mass analyzer is a part of a Fourier transform family of mass analyzers and is commonly used in proteomics (Figure 1.2) [3]. It consists of the ion chamber, spindle-shaped central electrode, and two external electrodes, isolated from each other. The central electrode establishes an electrostatic field, making charged particles orbit the electrode when they enter the chamber at the moment of the voltage ramp. The lateral oscillation rate is dependent on the m/z of ions, and the current from coherently oscillating ions can be deconvoluted through the Fourier transformation into separate pairs of oscillation frequencies and amplitudes per ion. Ion's m/z is further derived from frequency and intensity – from the amplitude [10]. External electrodes implement two functions: establishment of the ion trapping field and oscillating current detection.



Eliuk S, Makarov A. 2015.
Annu. Rev. Anal. Chem. 8:61–80

Figure 1.2: A schematic cross-section of a C-trap and Orbitrap mass analyzer. Ions, entering the ion trapping field, demonstrate the oscillating frequency around the main electrode, dependent on the m/z of ions. External electrodes pick up the signal current of the ions and transmit it to the amplifier. Adopted from [3].

1.2.2 timsTOF analyzer

The timsTOF mass spectrometer has a linear structure with its parts positioned in the following order (Figure 1.3A) [4, 11]:

1. Trapped ion mobility spectrometry (TIMS) tunnel;
2. Quadrupole filter;
3. Collision cell;
4. Time-of-flight (TOF) analyzer.

The TIMS tunnel is an incremental part of the timsTOF machine, responsible for the separation of ions based on their ion mobility. Ions are aggregated in the tube during the desired ion accumulation time and further trapped by the electric fields of the entrance and exit funnels. Inside, a constant electric field is applied in the direction of the entrance funnel and an inert gas flow in the opposite direction. The gas is drifting the bigger peptides with more complex shapes towards the exit funnel, while smaller molecules remain in the area, close to where they were initially aggregated. After the stored ions are released by decreasing the electric field strength at the exit funnel, larger ions enter the quadrupole first, while smaller ions enter the last [4]. The inverse reduced ion mobility $1/K_0$ is a measure of friction frequency against the drift gas inside a TIMS tube and is directly proportional to the mass of the peptide [12].

In the full MS1 scan, ions pass the inactive quadrupole and collision cell to be detected by the TOF analyzer. The latter consists of an acceleration region with a homogeneous electric field and a field-free drift region [13]. The two-stage reflectron compensates for the initial differences in the kinetic energy, so that the time ions spend in the drift region is only dependent on their m/z. Lighter molecules with a bigger charge reach the highest speed in the drift area.

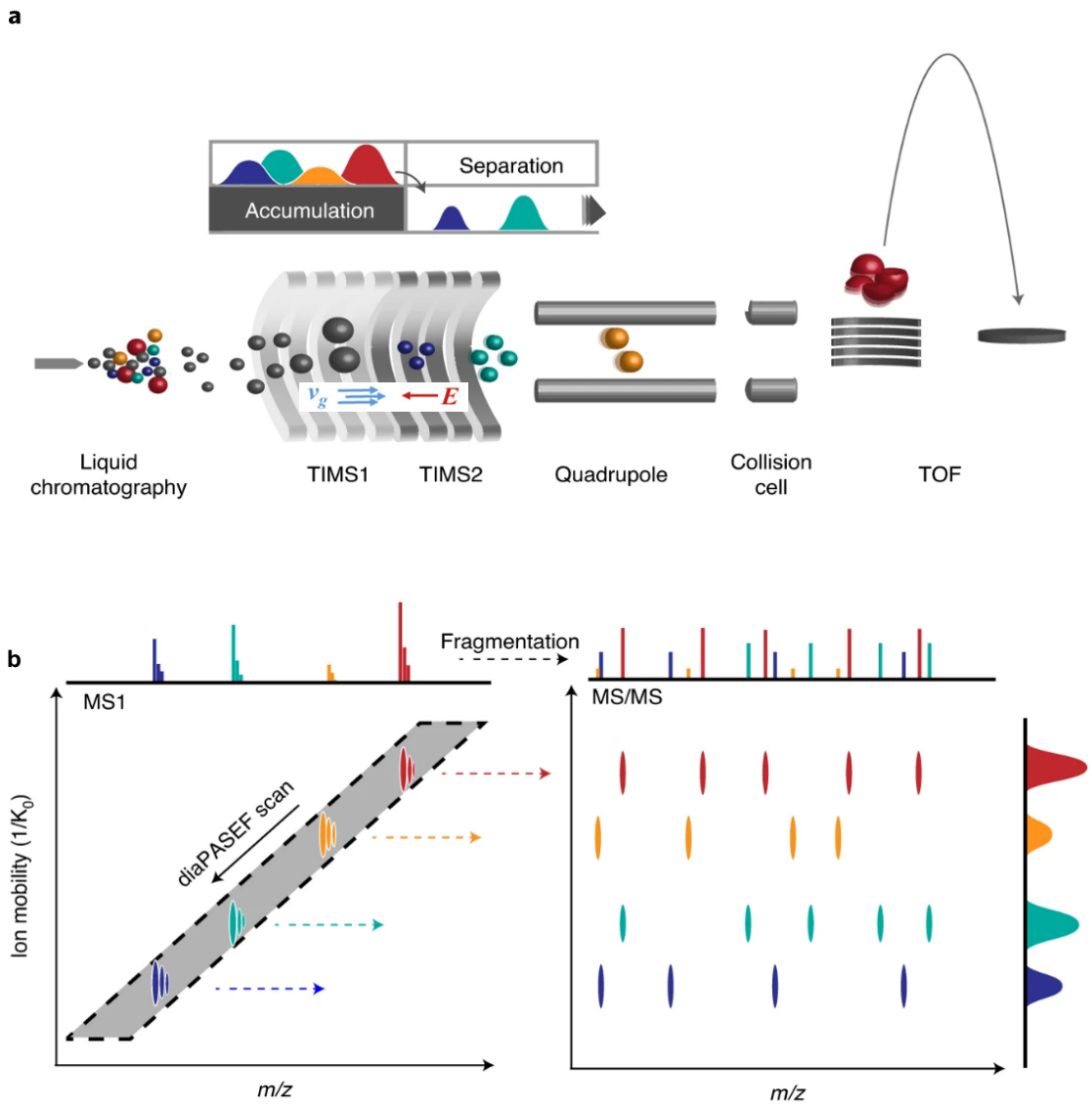


Figure 1.3: timsTOF mass spectrometry. a. Schematic representation of the ion path inside the timsTOF Pro mass spectrometer. Ions are separated in the TIMS tube under a constant electric field E and the drift gas velocity v_g. Larger ions exit first. The quadrupole and collision cell can be activated for MS2 acquisition. The TOF mass analyzer detects ions of different m/z based on their time of flight in the drift area. **b.** diaPASEF scan shifts isolation ranges from high to low m/z, resulting in multiple fragmentation events. Adopted from [11].

The TIMS 'elution' is operated in a pulse-like manner – ions, grouped by ion mobility, form a distinct 'mobility peak'. In other mass spectrometry setups, MS2 fragmentation and acquisition are done after filtering a single m/z range via a quadrupole [3, 9]. A distinct nature of the

mobility peaks and the fact that $1/K_0$ correlates with precursor mass allows for acquiring multiple MS2 scans per one full TIMS cycle by shifting quadrupole isolation windows from high to low m/z ranges alongside the TIMS elution (Figure 1.3B) [4, 11]. This method is termed 'parallel accumulation–serial fragmentation' (PASEF) and results in significantly higher acquisition speed and a better signal-to-noise ratio. Isolation windows of a quadrupole are stepped as a function of the TIMS ramp time (decrease in voltage of the exit funnel).

1.3 Data acquisition modes

The base data-dependent acquisition (DDA) mode was developed in conjunction with the aforementioned mass spectrometers [14, 15]. This approach relies on information from the full MS1 scan: m/z of the top N precursors with the highest intensities are selected by the quadrupole (Figure 1.4A) [16]. These precursors are individually isolated and fragmented in sequential MS2 scans. The result is clean spectra with a high signal-to-noise ratio due to the precise single precursor isolation. Prefix ions that contain the N-terminus are termed b-ions, and suffix ions containing the C-terminus are termed y-ions. The biggest drawback of this approach lies in the stochastic nature of sampling the most abundant precursors and leaving out the rest – such a strategy leads to irreproducible results and prevents measurements of low-abundant peptides. Moreover, to improve peptide coverage, DDA surveys the same precursor no more than twice, reducing the quantification precision, which benefits from multiple measurements (Figure 1.4C) [15]. A specific field, where DDA remained prevalent for a long time, was chemical labeling of samples for their simultaneous mass spectrometry analysis and subsequent relative quantification of precursors between labels [17, 18].

In data-independent acquisition (DIA), as the name states, the fragmentation does not rely on the information received from the full MS1 scan. Instead, during the MS2 acquisition, the quadrupole sequentially samples m/z ranges of pre-determined length over the whole initial MS1 scan (Figure 1.4B) [19]. The number and length of possible isolation ranges are dependent on the speed of each MS2 acquisition and the speed of HPLC [20]. DIA solves two fundamental problems of the previous DDA approach. Firstly, it systematically samples all the peptides for fragmentation, guaranteeing the presence of low-abundant molecular species in the MS2 spectra and surpassing the irreproducibility problem. Secondly, not only does it increase the

coverage of MS1-quantifiable signal by identifying the same precursors in multiple scans, but it also samples the same fragments after each MS2 cycle of the whole m/z range. That leads to both MS1 and MS2 signals obtaining elution profiles – distribution of intensities over several retention time points, which in turn can be used for quantification (Figure 1.4D) [15, 21].

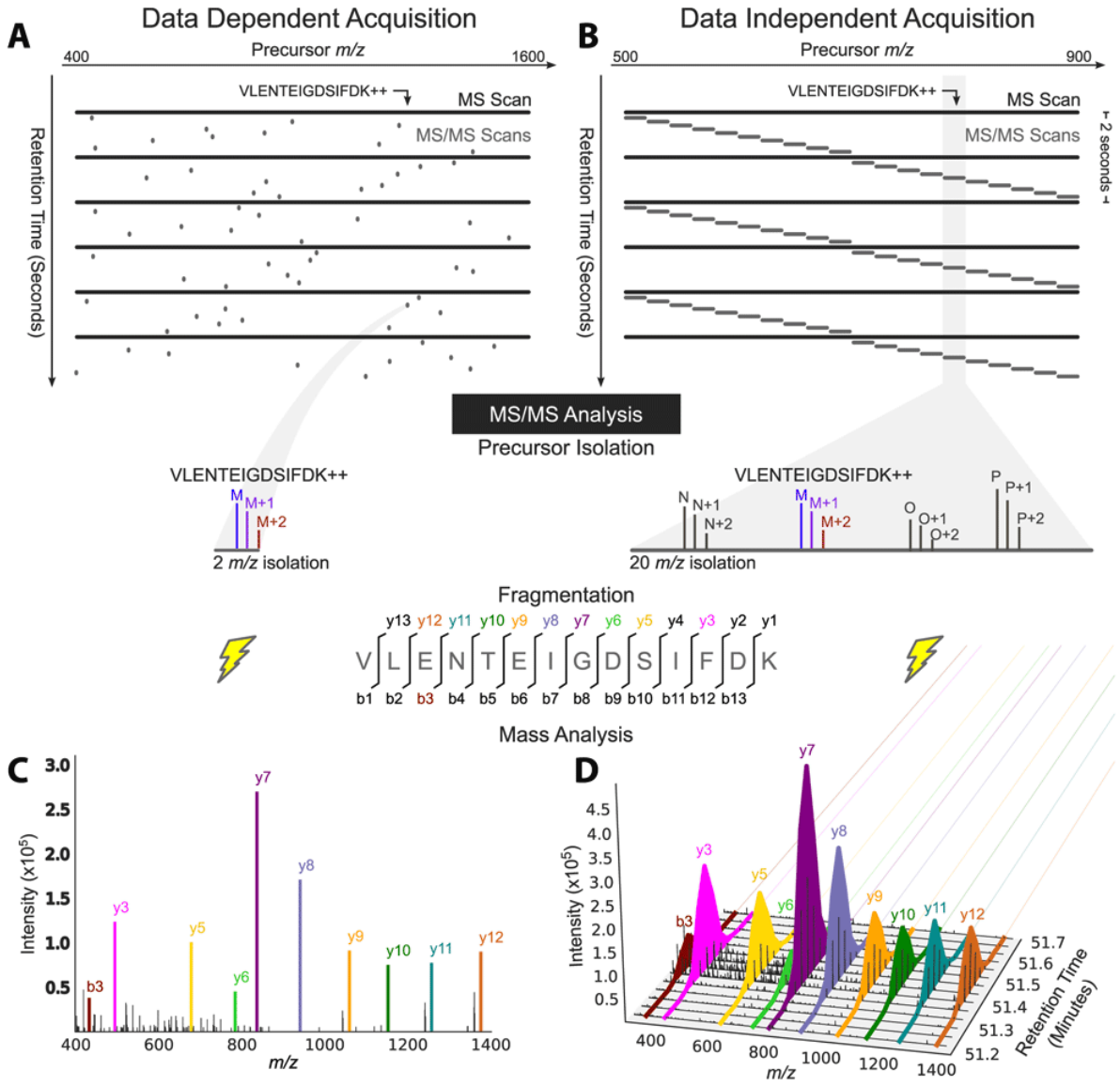


Figure 1.4: a. In DDA, MS/MS scans are collected using narrow isolation windows based on peptide precursors detected in an MS1 scan (peptide M). **b.** DIA employs wide isolation windows that are not targeted to specific peptide precursors. Instead, consecutive windows comprehensively cover a defined precursor m/z range, fragmenting multiple precursors

simultaneously within a single MS2 event (peptides N, M, O, P). **c.** DDA fragmentation information is captured within a single spectrum. **d.** In DIA, the same fragmentation information can be reconstructed into an elution profile and used for quantification. Adopted from [15].

1.4 Multiplexing in proteomics

The quantification in proteomics faces several intrinsic problems. The first one is run-to-run variability due to the differences in HPLC separation, ionization efficiency, and mass spectrometer performance [7, 22]. Altogether, this shifts the general distribution of protein intensities, making cross-sample comparisons inadequate. Another major problem is missing values – many low-abundant peptides are identified only in a subset of samples, hindering the quantification [23, 24].

To overcome those limitations, multiplexing techniques were developed during the DDA period of proteomics. If one combines multiple samples inside one LC-MS run using any sort of labeling, it removes the need to correct for retention time and m/z inconsistencies between runs and allows for direct intensity comparisons between labeling states [18, 25]. For the same reason of inter-comparability, identifications can be transferred between labels by incorporating the signal in the presumable m/z-RT frame of the undetected peptide, given that it was detected in another label – an algorithm termed ‘Re-Quantify’ in MaxQuant [22].

Another obvious advantage of multiplexing is throughput – the higher the number of possible multiplexing channels, the more samples can be analyzed simultaneously [26]. It is especially crucial in the applications for single-cell proteomics, where the number of samples grows proportionately to the number of cells taken for the analysis [27, 28].

Labels used in proteomics can be largely split into two categories (Figure 1.5). The first one is termed non-isobaric or MS1 precursor ion-based. These labels rely on introducing mass differences already at the MS1 level, with both MS1 precursor and MS2 fragment signals being shifted relative to each other by the mass difference of the labels used. This type of labeling

normally involves the use of stable isotopic forms of either amino acids or chemical groups [29].

The second type of labels is isobaric, MS2 reporter ion-based. Peptides are labeled with chemical groups of the same mass (hence the name 'isobaric'), but with different distributions of heavy isotopes in their structure. Those labels, known as tandem mass tags, contain a specified linker group optimized for cleavage during the MS/MS collision dissociation, resulting in tags of different masses detectable in the MS2 spectra, termed reporter ions [30].

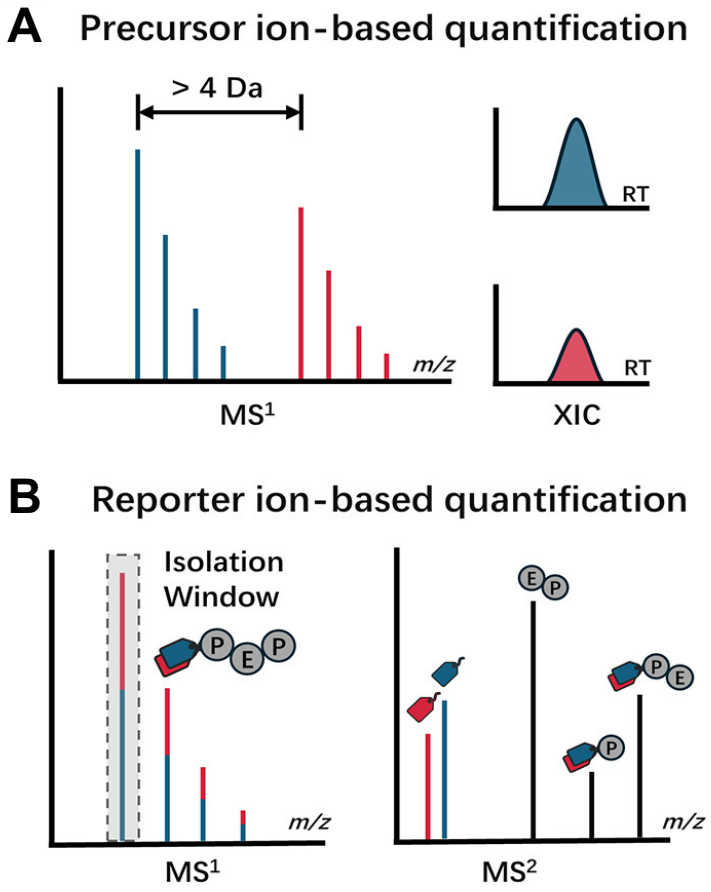


Figure 1.5: Schematic representation of labeling approaches in mass spectrometry. a. MS1 precursor ion-based quantification. Light and heavy precursor signals differ at least by 4 Da. Quantification is done by comparing MS1 elution profiles over retention time (extracted ion currents, XIC) between labels. **b.** MS2 reporter ion-based quantification. While all labeling states contribute to a single peak in the MS1 area, multiple mass tags are detected in the MS2 spectra after the fragmentation. Adapted from [29].

1.4.1 Non-isobaric labels

1.4.1.1 Stable isotopic labeled amino acids (SILAC)

SILAC utilizes amino acids with heavy isotopes of carbon and nitrogen incorporated into their structure. For example, $^{13}C_6{}^{15}N_4$-arginine or Arg10 is 10 Da heavier than its unlabeled counterpart. Arginine and lysine are used primarily, as they constitute a cleavage site for trypsin. Those amino acids are guaranteed to be included in every tryptic peptide, except the ones arising from the C-terminus of the protein [17].

Cell cultures are grown on different media: one containing unmodified arginine and lysine, and the other containing their isotopically labeled counterparts (Figure 1.6). Labels are incorporated metabolically in the protein structure during its synthesis, as all the amino acids are obtained from the media [16]. It is important to note that only cells that are unable to synthesize the amino acids used for labeling on their own can be used in such an experiment [31]. For example, all the mammalian cells are suitable, but only bacterial strains with defective arginine and lysine synthesis operons can be taken. Metabolic incorporation of labels reduces signal variability associated with sample preparation, which gives this method an advantage in quantification accuracy compared to other chemical types of labels [32].

As labeled amino acids are always present on the C-terminus of the peptide, MS1 precursor ions and MS2 fragment y-ions will be distinguishable between labels, while b-ions remain shared [28].

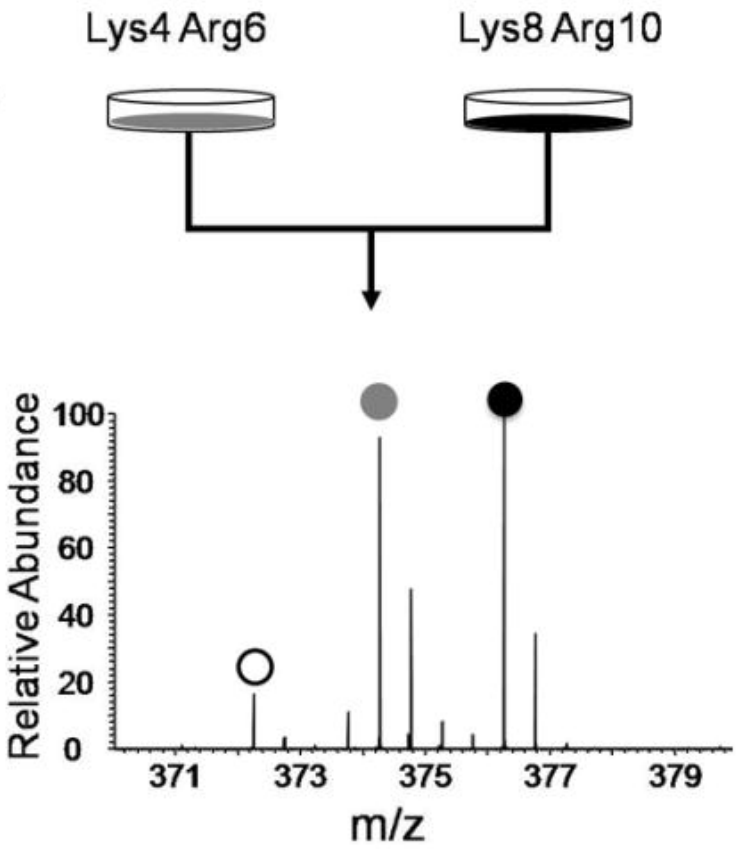


Figure 1.6: Schematic of multiplex SILAC labeling. Unlabeled, Lys4Arg6, and Lys8Arg10 cell cultures are mixed in the same run, resulting in three channels. Precursors and y-fragments demonstrate mass shifts of 0/4/8 for peptides with lysine at the C-terminus and 0/6/10 for peptides with arginine at the end. Adapted from [33].

1.4.1.2 Dimethylation

Dimethylation reaction is one of the chemical types of non-isobaric labeling. It is specific to the $-NH_2$ amine groups of N-termini and lysine residues (Figure 1.7). It involves an intermediate step of forming a Schiff base with formaldehyde CH_2O and subsequent reduction into a methyl group by cyanoborohydride $NaBH_3CN$ [34]. A light version of the label contains normal hydrogens, resulting in a mass shift of 28 Da per amino group. A heavy label incorporates deuterium ions instead of hydrogen, resulting in a mass shift of 36 Da per amino group. The use of tritium and ^{13}C increases the multiplexing capabilities to 5-plex [35].

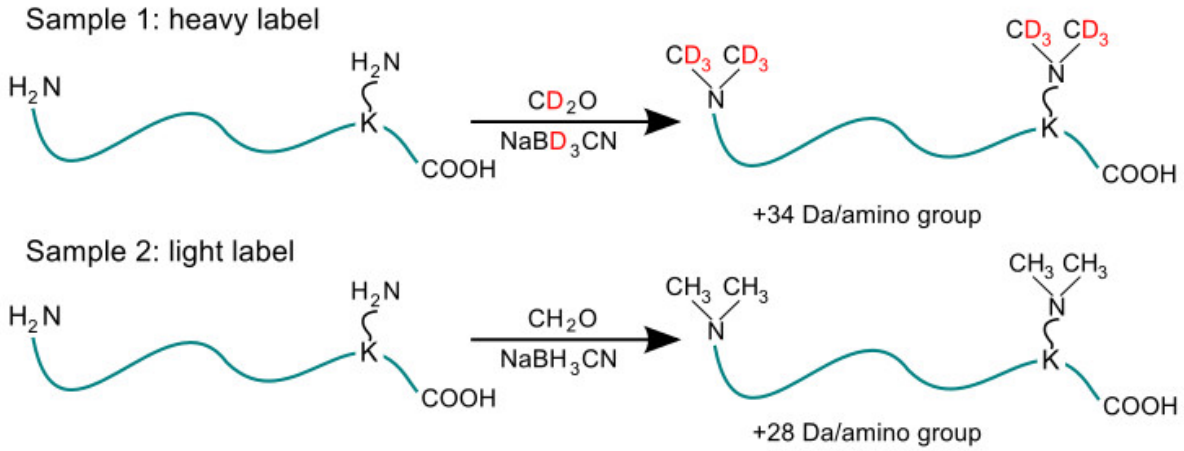


Figure 1.7: Schematics of dimethyl labeling. Free N-termini and lysine amine groups are methylated. The labeling capabilities are dependent on the inclusion of hydrogen and carbon isotopes.

1.4.1.3 Mass differential tags for relative and absolute quantification (mTRAQ)

mTRAQ is another type of non-isobaric chemical labeling done during sample processing. It also targets amine groups of N-termini and lysine residues via its peptide-reactive group (Figure 1.8) [36]. The mTRAQ Δ4 reagent has the identical structure and chemical composition as its isobaric counterpart iTRAQ Reagent 117 - 4plex [37]. The mTRAQ Δ0 has the same structure but lacks the stable isotopes (^{13}C, ^{15}N, ^{18}O), resulting in a molecular weight 4 Da lower than the Δ4 reagent. The mTRAQ Δ8 reagent has additional stable isotopes, resulting in a molecular weight 4 Da higher than the Δ4 reagent [38]. While the reporter group can be cleaved during the fragmentation, HCD/CID dissociation chambers are tuned to introduce one bond break at a time. Thus, when y- and b-ions are generated, the energy is spent to break the main series peptide bond and not the mTRAQ one [36].

MS1 precursor ions are always unique for amine-specific labels, as well as MS2 b-ions. y-ions are shared, if a peptide has an arginine at the end, and unique, when lysine is at the C-terminus [28].

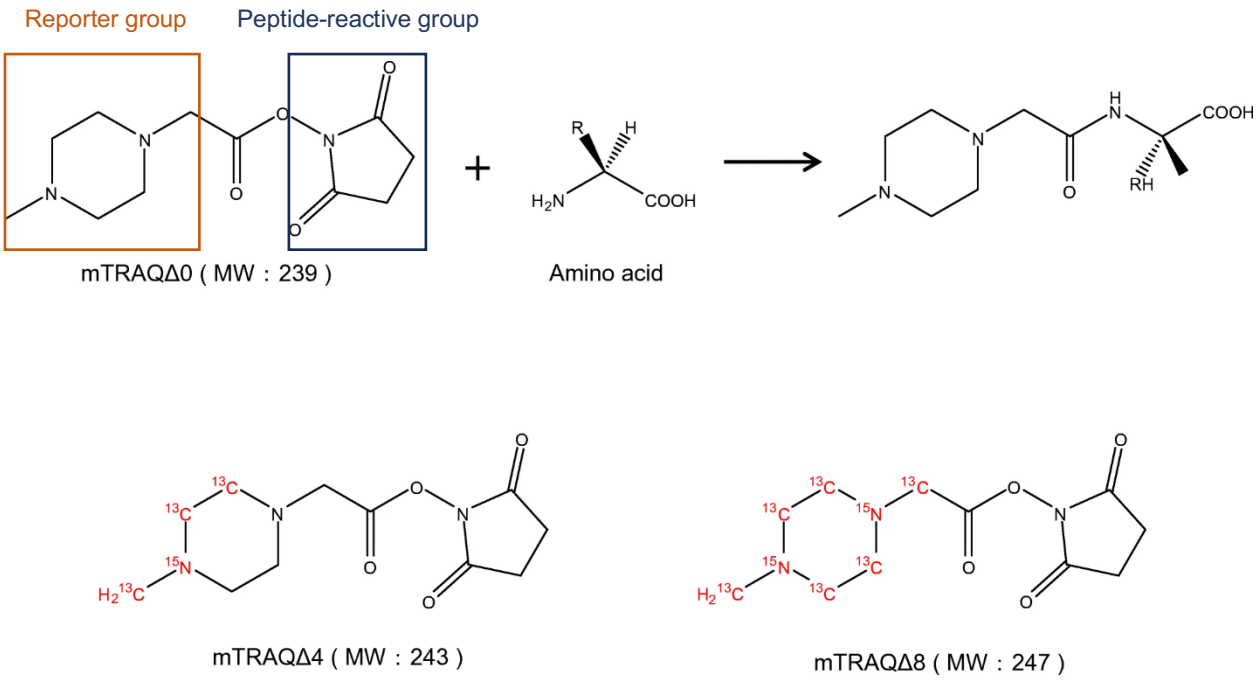


Figure 1.8: Schematics of mTRAQ labeling. mTRAQΔ4 and mTRAQΔ8 are stable isotopes of mTRAQΔ0. They bind to any free amine group of the peptide via its peptide-reactive group. Adapted from [38].

1.5 Quantification strategies for multiplexed samples

As mentioned before, multiplexing solves the problem of run-to-run variability due to sample processing and LS-MS performance changes. The most straightforward approach is to compare intensities of different channels inside a single run – this ensures minimal variances without applying normalization techniques and significantly reduces the number of missing values due to the 'Re-Quantify' algorithm [22, 32].

Large-scale experiments employ hundreds of samples to compare with each other. There are no labels with such multiplexing capabilities, but still, they can aid dramatically. One of the channels can be repurposed to be the reference channel with a constant known load of proteins. When comparing runs with each other, the reference channel serves as a normalization factor. E.g., if the experimental setup devotes a heavy channel to be the reference and a light one to contain biological samples, light-to-heavy intensity ratios (L/H) are compared between

samples. This way, differences between reference channels of different runs will account for run-to-run variability [18, 32, 39]. One can do an additional step and calculate the median intensity of a protein group in a reference channel across all runs. Multiplying the run-specific L/H ratio by this global reference intensity per protein group will directly yield normalized light channel intensity that can be compared between runs [31].

Traditionally, only the MS1 precursor ion signal is taken for quantification in the case of non-isobaric multiplexing approaches (precursor ion-based quantification). Isobaric techniques use reporter ion intensities produced during MS2 fragmentation for quantification (reporter ion-based quantification) [29]. Both approaches can utilize quantification inside a single run or reference channels.

1.6 Quantification strategies for label-free samples

To account for run-to-run variabilities and accurately compare intensities of label-free samples, the Label-free quantification algorithm, termed MaxLFQ, was developed [7]. It works under the assumption that most of the proteome between the compared samples doesn't experience drastic intensity fluctuations when subjected to case-control studies. This unchanged background proteome can be used to determine normalization factors. Another assumption is that retention time and m/z error are aligned between runs [22]. The algorithm can process both single-shot and fractionated samples.

The MaxLFQ algorithm starts with the calculation of normalization factors N_j (Figure 1.9). These factors determine a non-linear optimization model aimed at minimizing intensity changes across all peptides and samples. The total intensity of a peptide P in sample A is defined as the sum of its extracted ion currents (XICs or elution profiles) multiplied by normalization factors across all the fractions of sample A. Each sequence-charge-modification combination is treated separately. In DDA, XICs are only present for MS1 precursor signals. In DIA, fragments also possess their elution profiles and are subjected to normalization. By default, intensity at the RT of the XIC's apex is taken, but the whole volume of the XIC can be used as well. The sum of all squared logarithmic fold changes between all samples and summed

over all peptides H(N) is minimized using Levenberg–Marquardt optimization with respect to the normalization factors N_j.

A limitation of this normalization lies in the computation effort – it grows quadratically with the number of samples. The fast normalization option takes only a subset of fold changes into account to speed up the process.

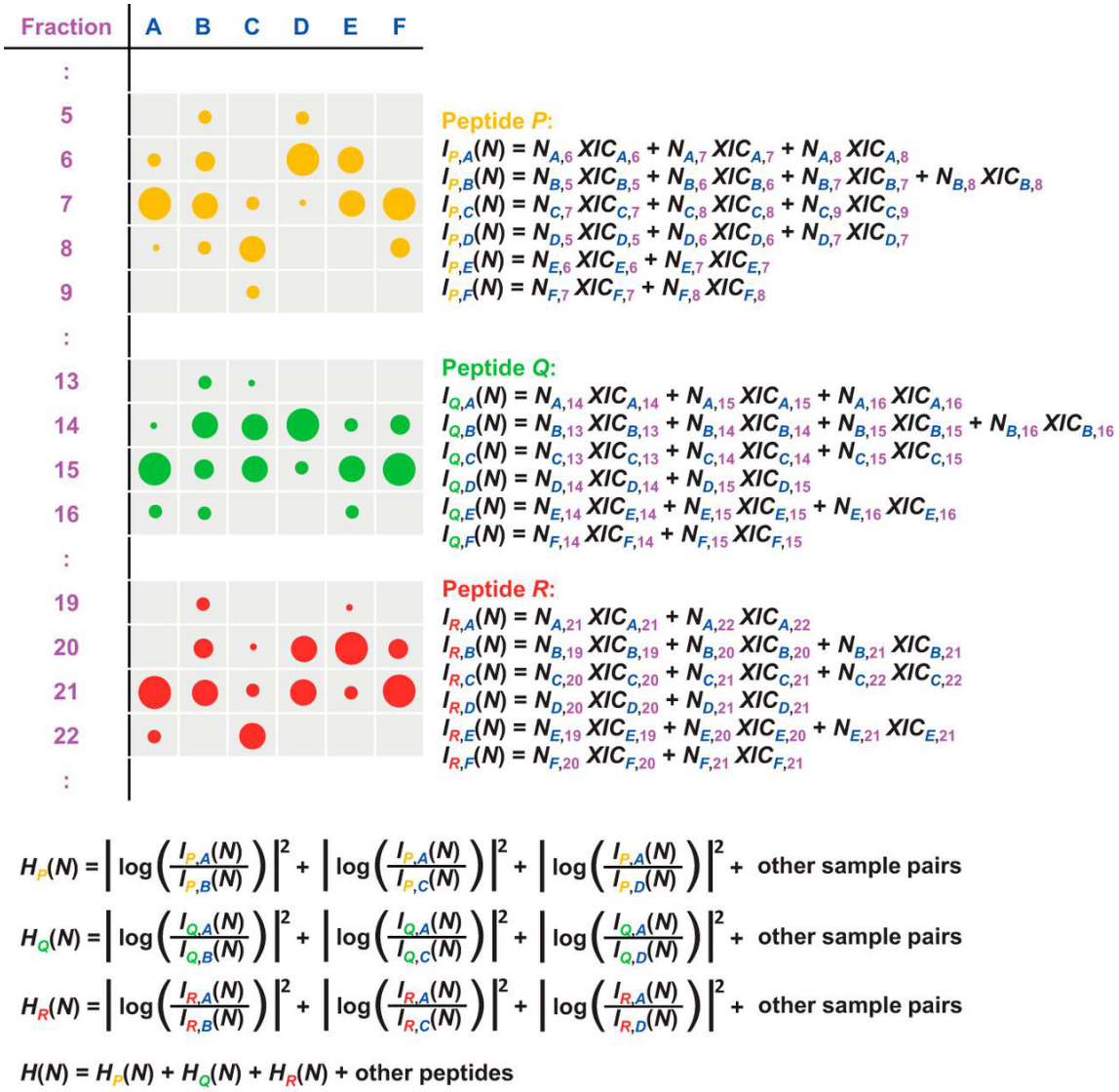
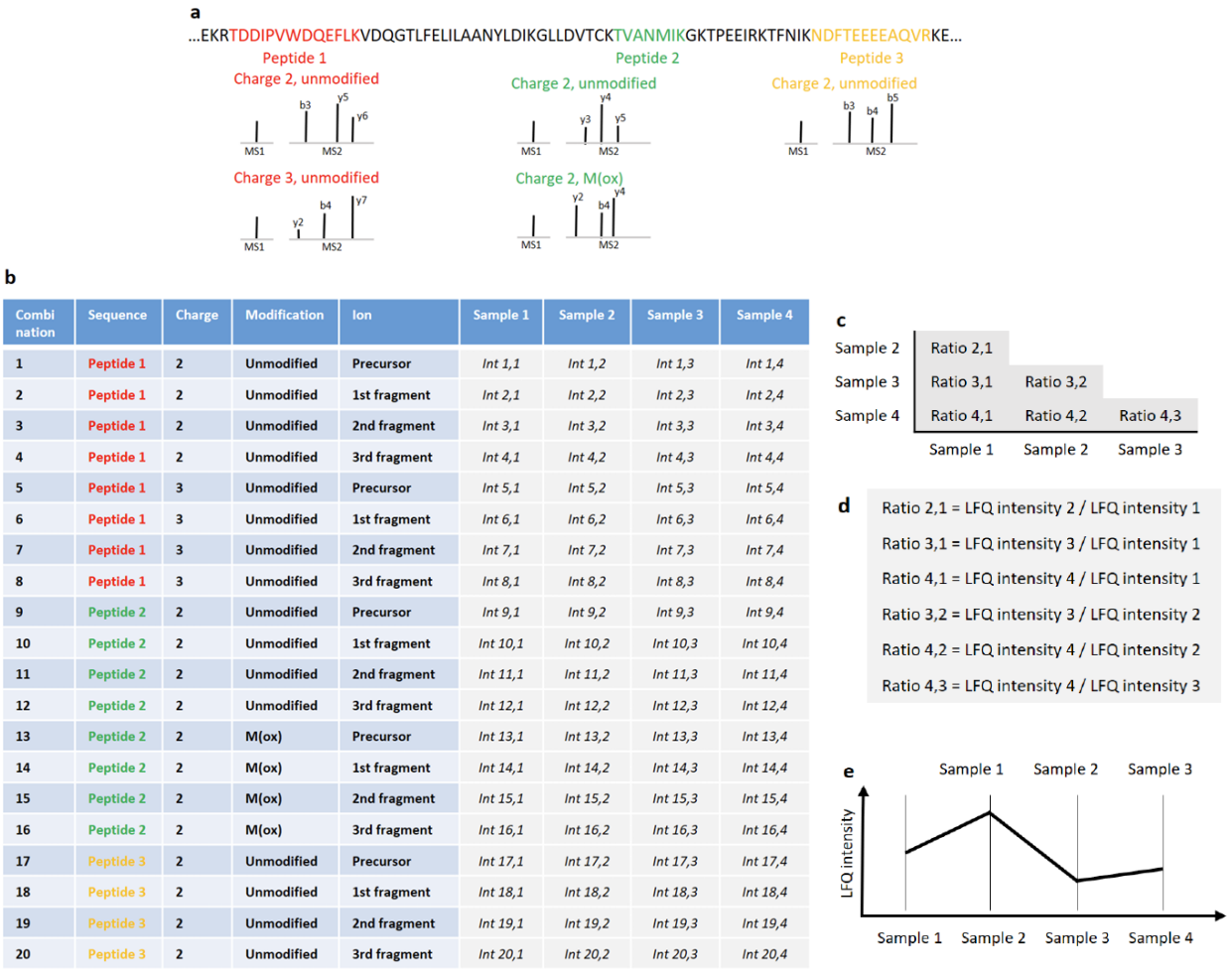


Figure 1.9: Schematics of H(N) calculation. Peptide intensity in each sample $I_j(N)$ is defined as the sum of its XICs multiplied by normalization factors. H(N) is the sum of the squared logarithmic changes in all samples (A, B, C, …) for all peptides (P, Q, R, …), and is minimized in relation to N_j. Adapted from [7].

The next step is protein group quantitation – for that, a matrix of MS1 precursor and top N MS2 fragment intensities is constructed across all samples for each sequence-charge-modification combination that constitutes a given protein group (Figure 1.10 a-b) [23]. By default, unique and razor peptides are used. The razor concept implies the use of non-unique peptides only in a protein group with the greatest number of underlying peptides. Ratio between samples is defined as the median of all the sequence-charge-modification intensity ratios between those samples (Figure 1.10c). If the ratios between samples are known, one can construct a system of equations, where each such known ratio corresponds to the ratio of LFQ intensities between the corresponding samples (Figure 1.10d). This system of equations is solved via the least-squares best fit, obtaining final LFQ intensities per sample (Figure 1.10e).



Combination	Sequence	Charge	Modification	Ion	Sample 1	Sample 2	Sample 3	Sample 4
1	Peptide 1	2	Unmodified	Precursor	*Int 1,1*	*Int 1,2*	*Int 1,3*	*Int 1,4*
2	Peptide 1	2	Unmodified	1st fragment	*Int 2,1*	*Int 2,2*	*Int 2,3*	*Int 2,4*
3	Peptide 1	2	Unmodified	2nd fragment	*Int 3,1*	*Int 3,2*	*Int 3,3*	*Int 3,4*
4	Peptide 1	2	Unmodified	3rd fragment	*Int 4,1*	*Int 4,2*	*Int 4,3*	*Int 4,4*
5	Peptide 1	3	Unmodified	Precursor	*Int 5,1*	*Int 5,2*	*Int 5,3*	*Int 5,4*
6	Peptide 1	3	Unmodified	1st fragment	*Int 6,1*	*Int 6,2*	*Int 6,3*	*Int 6,4*
7	Peptide 1	3	Unmodified	2nd fragment	*Int 7,1*	*Int 7,2*	*Int 7,3*	*Int 7,4*
8	Peptide 1	3	Unmodified	3rd fragment	*Int 8,1*	*Int 8,2*	*Int 8,3*	*Int 8,4*
9	Peptide 2	2	Unmodified	Precursor	*Int 9,1*	*Int 9,2*	*Int 9,3*	*Int 9,4*
10	Peptide 2	2	Unmodified	1st fragment	*Int 10,1*	*Int 10,2*	*Int 10,3*	*Int 10,4*
11	Peptide 2	2	Unmodified	2nd fragment	*Int 11,1*	*Int 11,2*	*Int 11,3*	*Int 11,4*
12	Peptide 2	2	Unmodified	3rd fragment	*Int 12,1*	*Int 12,2*	*Int 12,3*	*Int 12,4*
13	Peptide 2	2	M(ox)	Precursor	*Int 13,1*	*Int 13,2*	*Int 13,3*	*Int 13,4*
14	Peptide 2	2	M(ox)	1st fragment	*Int 14,1*	*Int 14,2*	*Int 14,3*	*Int 14,4*
15	Peptide 2	2	M(ox)	2nd fragment	*Int 15,1*	*Int 15,2*	*Int 15,3*	*Int 15,4*
16	Peptide 2	2	M(ox)	3rd fragment	*Int 16,1*	*Int 16,2*	*Int 16,3*	*Int 16,4*
17	Peptide 3	2	Unmodified	Precursor	*Int 17,1*	*Int 17,2*	*Int 17,3*	*Int 17,4*
18	Peptide 3	2	Unmodified	1st fragment	*Int 18,1*	*Int 18,2*	*Int 18,3*	*Int 18,4*
19	Peptide 3	2	Unmodified	2nd fragment	*Int 19,1*	*Int 19,2*	*Int 19,3*	*Int 19,4*
20	Peptide 3	2	Unmodified	3rd fragment	*Int 20,1*	*Int 20,2*	*Int 20,3*	*Int 20,4*

Figure 1.10: Schematics of MaxLFQ quantitation. a. All the detected sequence-charge-modification combinations are taken per protein group. **b-c.** A matrix of intensities across all

samples. Ratio 2,1 is determined as the median of per-row ratios between columns ‘Sample 1’ and ‘Sample 2’ in the matrix. **d-e.** LFQ intensities are acquired by solving the least-squares best fit on the system of equations with pre-determined sample ratios. Adapted from [23].

1.7 Advances of multiplexing into DIA

Most of the multiplexing techniques were developed when DDA was prevailing in proteomics. It partially resolved issues like missing values, quantification comparability, and throughput [29]. DIA allowed for systemic detection of MS1 precursors and MS2 fragments over retention time and ion mobility points, which significantly reduced the number of missing values compared to the plain DDA [23]. The MaxLFQ algorithm, in turn, enabled reliable label-free quantification [7]. Still, multiplexing found its way into the DIA field as a powerful approach to increase identifications in low-abundant samples by incorporating a carrier channel – it usually contains the same type of proteome as experiment channels, but with a hundred times higher load. This allows for easy detection of peptides in the carrier channel and transfer of those identifications to the experimental channels. Multiplexing also increases experimental throughput, which is crucial in applications like single-cell proteomics, where label-free approaches have to process each cell in a separate run [27, 28, 31].

As DIA aims to utilize both MS1 and MS2 fragment signals for quantification, a type of label, distinguishable on both of those levels, is required [28]. Non-isobaric labels fulfill that role and offer both metabolic (SILAC) and chemical (dimethylation, mTRAQ) types of marking peptides [17, 34, 38]. Thus appeared the need for appropriate computational algorithms to be developed and enable multiplexed DIA experiments.

1.7.1 DIA-NN multiplexing algorithm

DIA-NN software initiates its plexDIA workflow with the ‘Preliminary identification’ step, where all channels are searched separately using spectra constructed from the non-labeled library, treating labels as variable modifications (Figure 1.11A) [28]. For example, to specify the mTRAQΔ0, mTRAQΔ4, and mTRAQΔ8 labels, a fixed modification of mTRAQ 140.0949630177 is added to every library entry, and three channels have further mass shifts of

0, 4.0070994, and 8.0141988132. Each of the target queries has an associated decoy sequence, resulting from the ‘mutation’ of the original entry, while the last amino acid is swapped: either R → K or K → R. Additionally, a ‘decoy’ channel is created, referring to the library spectra, shifted by the non-present label – Δ12 in the case of mTRAQ. There is no conclusive information on whether DIA-NN has spectrum prediction models specific to different types of labels.

The q-values are calculated through the pre-established neural-network classifier of original and ‘mutated’ sequences, used analogously in the label-free DIA-NN workflow (Figure 1.11B) [40]. Even though the ‘decoy’ channel receives its own q-value, the information about its origin is not used during this round of the FDR estimation. At this point, DIA-NN estimates the best-scoring channel out of all the non-decoy channels. Its apex RT is assumed to pinpoint the correct retention time of the peptide. DIA-NN re-extracts the signals at this retention time for the other channels (including the decoy one), regardless of whether these have been successfully matched to some peak groups during the previous step. All identifications are subject to another round of regular neural-network classification, resulting in the ‘translated q-values’. As the last step of the FDR estimation, DIA-NN calculates ‘channel q-values’ that reflect the confidence in the precursors being present in specific channels (Figure 1.11C). The same neural-network classifier architecture is used, but the separation of the identifications is based on whether they are present in the ‘target’ or ‘decoy’ channel.

For quantification purposes, only MS1 precursors and MS2 unique fragments (e.g., b-ions for amine-specific labels) that reach sufficient quality are taken. There is no clear description of the quality criteria in the plexDIA publication, but presumably it relies on the sum of correlation scores for each fragment. DIA-NN calculates the ratios between different channels using the signal ratio of MS1-level precursors or the median of signal ratios of MS2-level selected fragments at the RT apex (Figure 1.11D). The ‘translated’ quantity of each channel except the best one is determined as the quantity of the best channel divided by the corresponding ratio.

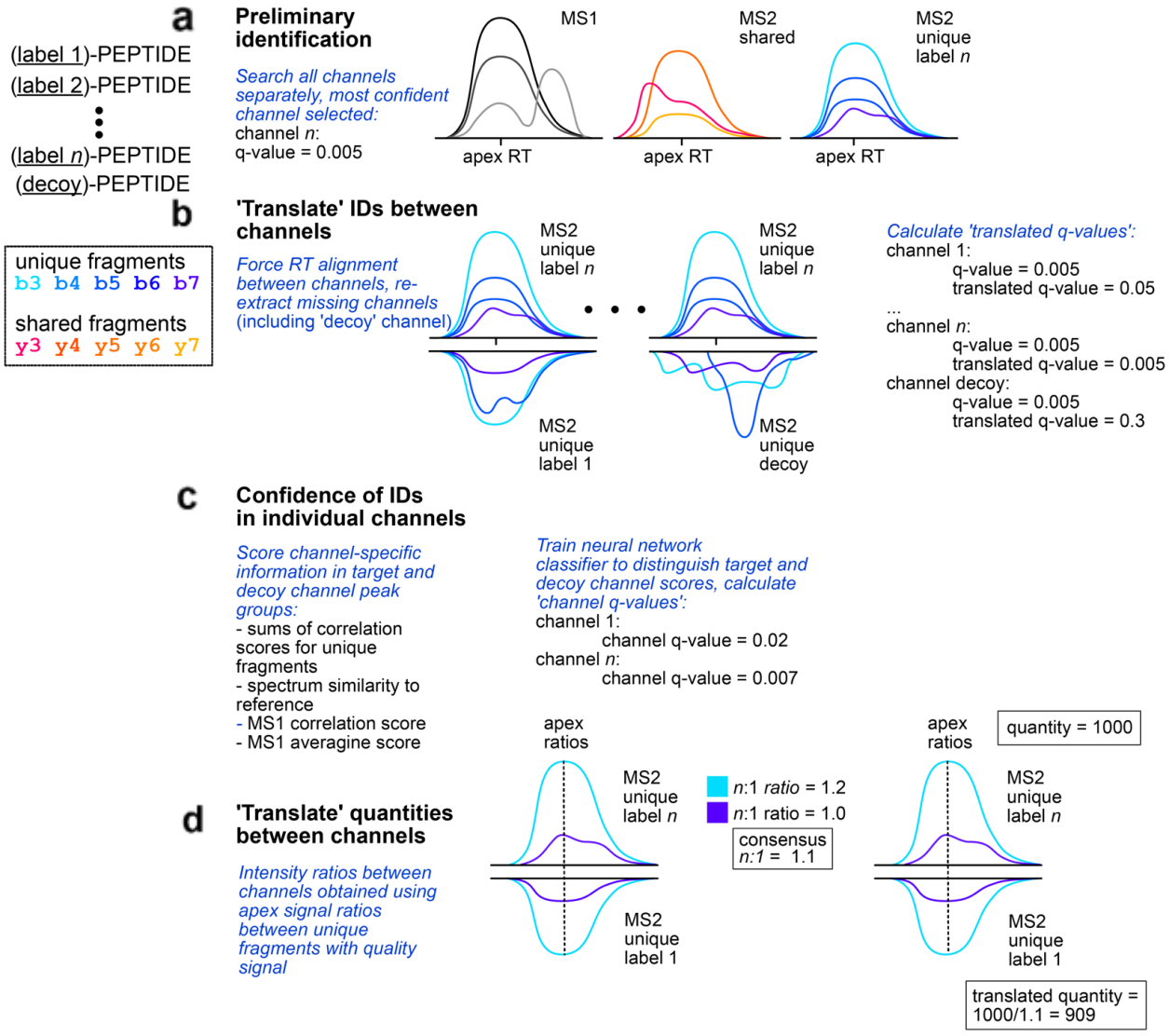


Figure 1.11: plexDIA data processing by DIA-NN. a. Preliminary identification is done using the multiplexed library entries. MS2 fragments, unique and shared between labels, are identified. A regular round of neural-network classification results in q-values, which in turn determine the best channel per sequence-charge-modification combination. **b.** MS1 and MS2 signals are re-extracted at the RT apex of the best identification, followed by another round of regular neural-network classification, resulting in 'translated q-values'. **c.** 'Channel q-values' are generated by using the same features and neural-network architecture as in the previous step, but distinguishing identifications in 'target' and 'decoy' channels (e.g., Δ12 for mTRAQ). **d.** DIA-NN calculates the ratios between different channels using the signal ratios for selected fragment ions at the elution apex. The 'translated' quantities are then calculated for all the

channels except the most confident one, by dividing the quantity in the latter by the respective ratio. Adapted from [28].

2 Research aims

The main goal of this project is to implement a multiplexing module termed MultiplexDIA into the MaxDIA workflow of MaxQuant and assess its performance in comparison with analogous software. The goal can be subdivided into the following tasks:

1 Development of the MultiplexDIA module inside MaxQuant;
2 Performance comparisons on the samples with normal levels of proteomes;
3 Performance comparisons on the samples with single-cell-like levels of proteomes;
4 Performance comparisons on different non-isobaric labels.

3 Materials and methods

3.1 SILAC benchmarking

3.1.1 SILAC bulk dataset description

Data for the identification and quantification benchmark of the MultiplexDIA performance was taken from the PXD052080 repository [31]. It was acquired through the timsTOF SCP instrument. *E. coli* and *H. sapiens* proteomes are mixed in predefined ratios ranging from 1:50 to 50:50 ng and constitute a label-free series (Figure 3.1). A heavy-labeled spike-in mixture of *E. coli* and *H. sapiens* proteomes, mixed in a 100:100 ratio, was added to the abovementioned samples to obtain a spiked series. Heavy stable isotope-encoded lysine and arginine (Lys8 and Arg10) were used. Thus, predefined ratios of the light channels across samples represent the ground truth in assessing the quantification accuracy of different software.

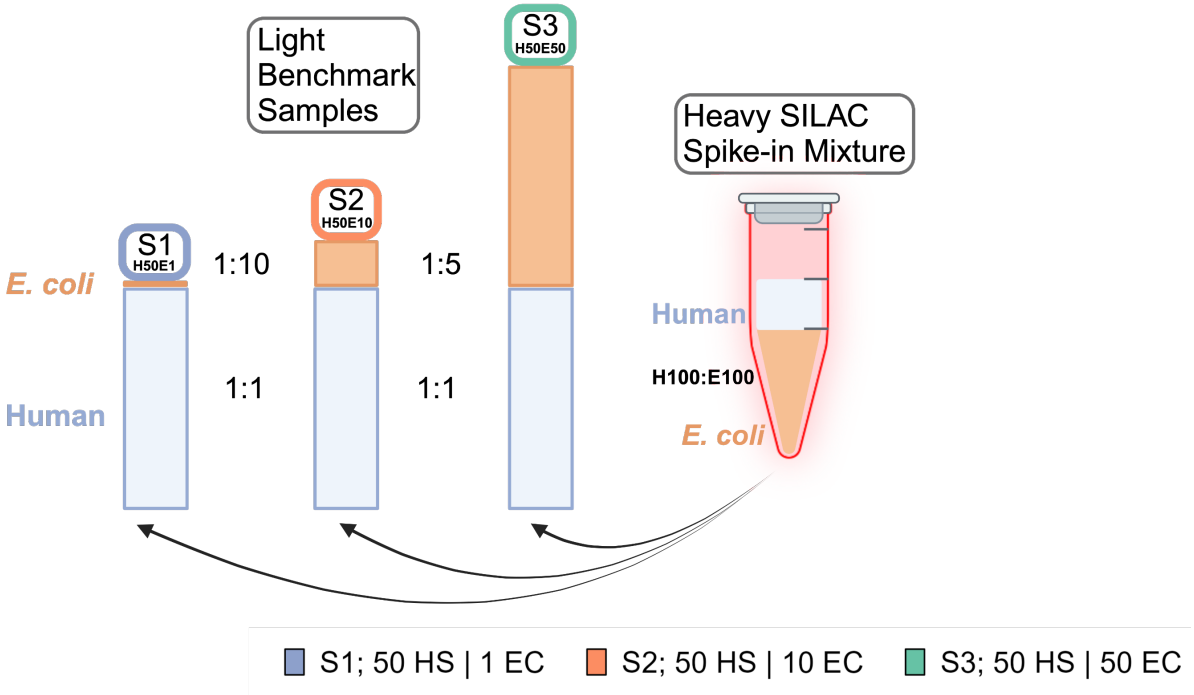


Figure 3.1: Experimental set-up of the bulk dataset. Label-free series with *E. coli* proteomes varying from 1 to 50 ng and with a constant *H. sapiens* background of 50 ng was spiked with Lys8/Arg10 heavy-labeled mixture of 100 ng *E. coli* to 100 ng *H. sapiens* proteomes. Adapted from [31].

3.1.2 SILAC single-cell-like dataset description

The next dataset was taken from the same PXD052080 repository and has a similar structure to the previous one: a series of unlabeled *E. coli* proteins, varying from 1 to 100 ng, with a constant *H. sapiens* background, was diluted in a 1:5 ratio, resulting in single-cell-like concentrations of proteins (series 15). It was then spiked with a 10 ng *E. coli* / 100 ng *H. sapiens* mixture, heavily labeled with Lys8 and Arg10 (Figure 3.2) [31].

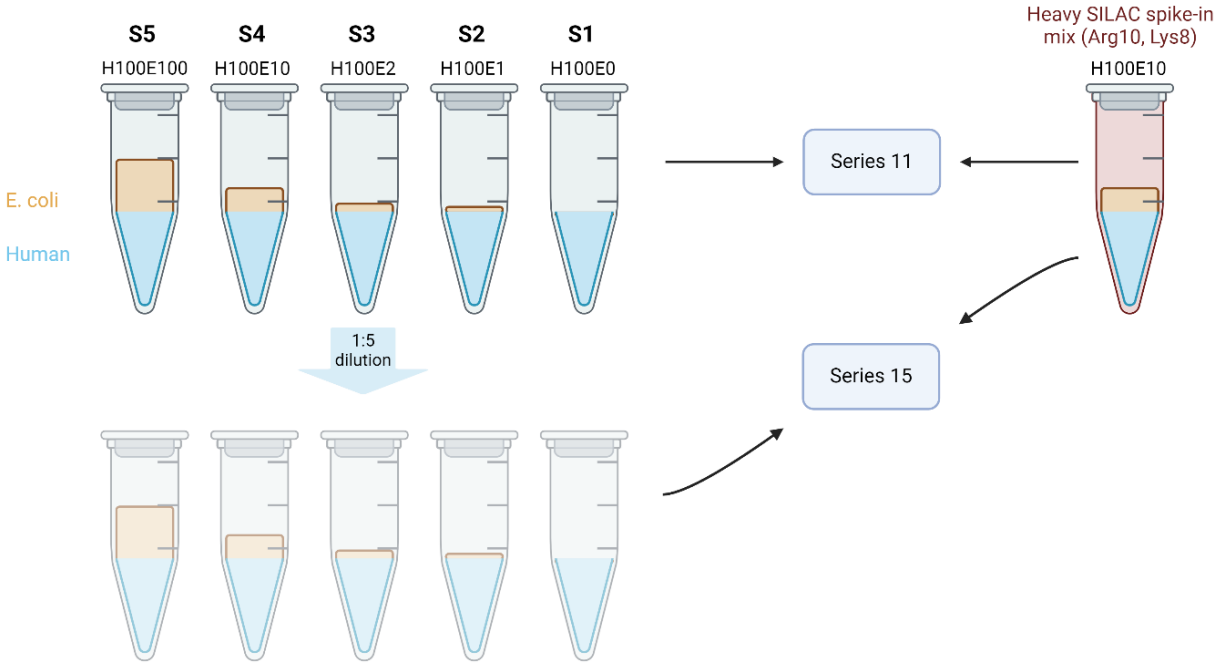


Figure 3.2: Experimental set-up of the single-cell-like dataset. Label-free series with *E. coli* proteomes varying from 1 to 100 ng and with a constant *H. sapiens* background of 100 ng was diluted in a 1:5 ratio and spiked with Lys8/Arg10 heavy-labeled mixture of 10 ng *E. coli* to 100 ng *H. sapiens* proteomes. Only series 15 was analyzed.

3.1.3 MaxQuant processing of SILAC data

Raw files were processed with MaxQuant v. 2.7.2.0. The run type was set to 'TIMS-MaxDIA' and the library type 'Predicted.' Multiplicity was set to 2, with the light channel left unmodified and the heavy channel configured to contain Lys8 and Arg10. For library generation, *E. coli* (UP000000625, accessed 25.01.2023) and *H. sapiens* (UP000005640, accessed 25.01.2023) were uploaded into the 'Sequences' tab, with contaminants incorporated into the prediction workflow by default. Label-free quantification (LFQ) was enabled using default parameters to normalize signals across all channels and samples. 'Split protein groups by protein ID' was turned on. To evaluate different transfer modes between labels, the 'DIA multiplex quant method' parameter has been switched between 'MS2 identified only', 'MS1 multiplex features', and 'Requantify'. All of the bulk data was analyzed using 'Intensity threshold MS1'

= 35, ‘Intensity threshold MS2’ = 10. Label-free single-cell-like data was analyzed using ‘Intensity threshold MS1’ = 4, ‘Intensity threshold MS2’ = 4. SILAC single-cell-like data was analyzed using ‘Intensity threshold MS1’ = 6, ‘Intensity threshold MS2’ = 6. The rest of the parameters remained at their default values.

3.1.4 Postprocessing of MaxQuant SILAC results

The *proteinGroups.txt* outputs from both label-free and spiked runs were filtered for contaminants and reverse sequences. The ‘LFQ intensity’ column was used in the label-free samples.

As MaxQuant uses the MaxLFQ algorithm to normalize the quantification signal across different samples and channels, there is no need to operate with L/H ratios of the protein – one can directly access the normalized light channel intensity of the protein by selecting the ‘LFQ intensity L’ column in the SILAC-spiked samples.

All samples except the lowest-diluted one (S3 for the bulk dataset and S5 for the single-cell-like) were corrected by the difference between the theoretical human ratio (1:1) and the observed median human log10 fold change. Median of across-sample log10 human protein group ratios between each sample and the lowest-diluted one was subtracted per replicate from the corresponding sample’s ‘LFQ intensity L’ (SILAC data) or ‘LFQ intensity’ (label-free data).

3.1.5 DIA-NN processing of SILAC data

DIA-NN v1.8.1 was used. FASTA files for *E. coli* (UP000000625, accessed 25.01.2023), *H. sapiens* (UP000005640, accessed 25.01.2023) and contaminants (Universal Contaminant Protein FASTA, accessed 25.01.2023) were concatenated and used for the generation of a predicted library using the following DIA-NN settings: --fasta-search --min-fr-mz 200 --max-fr-mz 1800 --met-excision --cut K∗,R∗ --missed-cleavages 1 --min-pep-len 7 --max-pep-len 30 --min-pr-mz 300 --max-pr-mz 1800 --min-pr-charge 1 --max-pr-charge 4 --unimod4 --reanalyze --relaxed-prot-inf --smart-profiling --peak-center --no-ifs-removal.

To generate smaller versions of the libraries, label-free raw files with the lowest dilution in each dataset were analyzed using the predicted library. In case of the bulk dataset, the label-free sample S3 was analyzed using default settings, except changing the library generation setting to “IDs, RT and IM profiling”. The same was done for the single-cell-like dataset using the label-free sample S5.

Bulk label-free raw files were processed with DIA-NN v.1.8.1 using the respective refined library and following settings: --qvalue 0.01 --reanalyse --relaxed-prot-inf --rt-profiling --peak-center --no-ifs-removal --report-lib-info --no-norm. Single-cell-like label-free raw files were processed in the same manner, but changing --rt-profiling to --smart-profiling.

Bulk SILAC raw files were processed with DIA-NN v.1.8.1 using the respective refined libraries and following settings: --qvalue 0.01 --relaxed-prot-inf --rt-profiling --peak-center --no-ifs-removal --fixed-mod SILAC,0.0,KR,label --lib-fixed-mod SILAC --channels SILAC,L,KR,0:0; SILAC,H,KR,8.014199:10.008269 --peak-translation --no-norm --no-maxlfq --original-mods --report-lib-info. Single-cell-like SILAC raw files were processed in the same manner, but changing --rt-profiling to --smart-profiling.

3.1.6 Postprocessing of DIA-NN SILAC results

LFQ reports were filtered to exclude contaminants and retain entries with Precursor.Charge > 1, Lib.PG.Q.Value < 0.01, and Lib.Q.Value < 0.01. The PG.MaxLFQ column was used as the quantitative output for each protein group (PG).

SILAC reports were filtered for contaminants and Precursor.Charge > 1, with additional thresholds applied: Global.PG.Q.Value < 0.01 and Channel.Q.Value < 0.03. These filters were applied to (a) both the light and heavy channels (basic filtering) and (b) only the heavy channel (‘intensity translation’ mode). In ‘intensity translation’ mode, if only the heavy channel passed the q-value filter, the corresponding light precursor was also retained.

Each protein group’s log10 L/H ratio in each replicate is defined as the median of the corresponding L/H ‘Ms1.Translated’ and ‘Precursor.Translated’ log10 ratios in this replicate. Next, the log10 global heavy intensity per protein group is defined by (1) summing up all heavy

precursor intensities for each protein group per sample, and (2) taking the median log10 of these summed intensities across all samples. Light intensity of the protein group in each sample is derived by multiplying (i.e., adding, in log space) the sample-specific L/H ratio by the global heavy intensity of the protein group.

All samples except the lowest-diluted one (S3 for the bulk dataset and S5 for the single-cell-like) were corrected by the difference between the theoretical human ratio (1:1) and the observed median human log10 fold change. Median of across-sample log10 human protein group ratios between each sample and the lowest-diluted one was subtracted per replicate from the corresponding sample's light channel intensity (SILAC data) or PG.MaxLFQ intensity (label-free data).

3.1.7 Common postprocessing of MaxQuant and DIA-NN SILAC results

In every result section, we operate with logarithmic fold changes (log10 for identification and quantification performance, log2 for FDR control) of *E. coli* protein groups from the light channel between different samples. If a protein group lacked even one of the four replicate ratios, it was discarded. The mean of log ratios across replicates was taken, otherwise.

Similarly to the intensity log fold change calculation, one can determine the significance level of the differences between four replicates of each condition using Student's t-test without multiple testing correction. p-value < 0.01 (-log10(p-value) > 2) and log2FC < -1 was considered significantly differentially abundant area in the volcano plot representation of the results. We know, as the ground truth, that unlabeled *E. coli* proteins differ between samples and *H. sapiens* proteins do not. Thus, one can classify E. coli data points in the significant area as True Positives and E. coli data points in the insignificant area as False Negatives. Alternatively, *H. sapiens* data points in the significant area are False Positives, and *H. sapiens* data points in the insignificant area are True Negatives.

The ROC-curve describes the precision (True Positives / (True Positives + False Positives)) against the recall (True Positives / (True Positives + False Negatives)) at a given threshold of the parameter. Since p-value and log2FC determine the significance of the observation, each

of those parameters can be used to generate a set of ROC-curves. Only proteins with negative fold changes were used. The focus was on the high-precision area in the range of 0.9 to 1.

3.2 mTRAQ benchmarking

3.2.1 mTRAQ dataset description

mTRAQ-labeled DIA data was taken from the MSV000089093 repository [28]. Data was acquired through the Q Exactive Orbitrap instrument. *E. coli*, *S. cerevisiae*, and *H. sapiens* proteomes are mixed in predefined ratios: 20/15/65% in sample A, 5/30/65% in sample B, and 30/5/65% in sample C (Figure 3.3). Each of those samples was analyzed in a label-free fashion. Samples A, B, and C were further chemically labeled by mTRAQΔ0, mTRAQΔ4, and mTRAQΔ8, respectively. Labeled samples were pooled and analyzed simultaneously in a multiplexed format. Thus, predefined ratios of different channels can be taken as a ground truth in assessing the quantification accuracy.

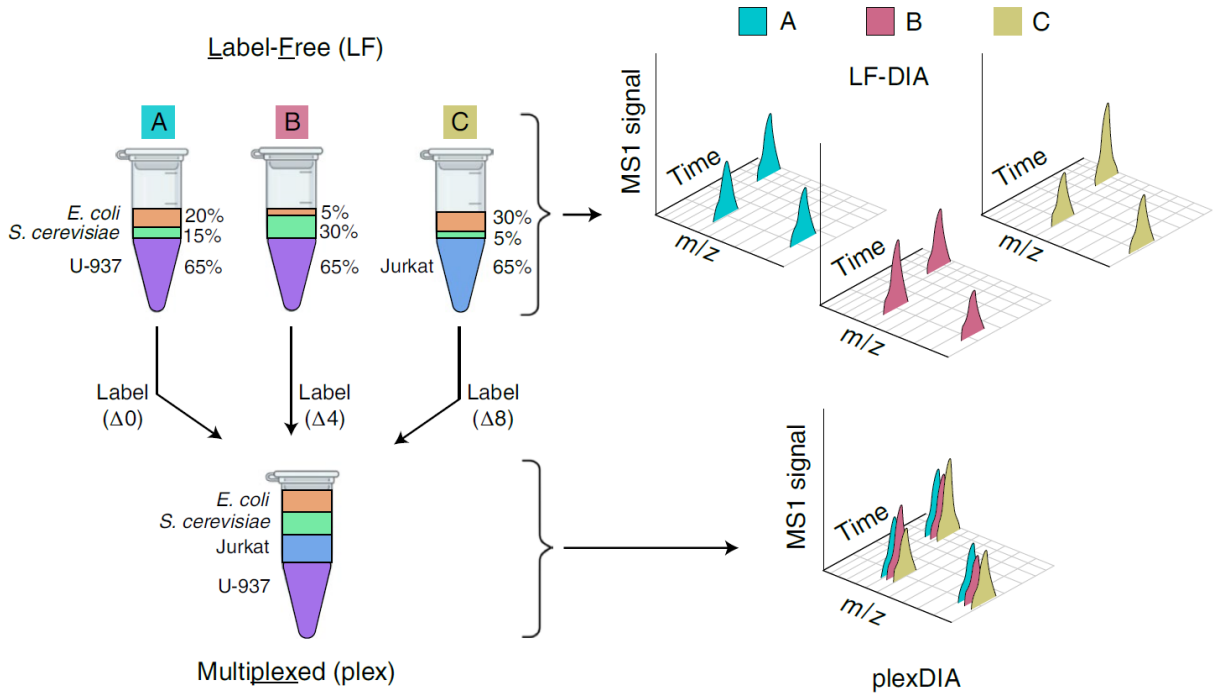


Figure 3.3: Experimental set-up of the mTRAQ dataset. Proteomes of *E. coli*, *S. cerevisiae*, and *H. sapiens* are mixed in pre-defined ratios in samples A, B, and C. They are analyzed in

parallel in a label-free mode, or pooled together using mTRAQΔ0, Δ4, and Δ8 labels. Adapted from [28].

3.2.2 MaxQuant processing of mTRAQ data

Raw files were processed with MaxQuant v. 2.7.2.0. The run type set to 'MaxDIA' and the library type 'Predicted.' Multiplicity was set to 3. Light labels: 'mTRAQ-Lys0' + 'mTRAQ-Nter0'; medium labels: 'mTRAQ-Lys4' + 'mTRAQ-Nter4'; heavy labels: 'mTRAQ-Lys8' + 'mTRAQ-Nter8'. For library generation, *E. coli* (UP000000625, accessed 01.02.2022), *S. cerevisiae* (UP000002311, accessed 01.02.2022), and *H. sapiens* (UP000005640, accessed 01.02.2022) were uploaded into the 'Sequences' tab, with contaminants incorporated into the prediction workflow by default. Label-free quantification (LFQ) was enabled using default parameters to normalize signals across all channels and samples. 'Split protein groups by protein ID' was turned on. The rest of the parameters remained at their default values.

3.2.3 Postprocessing of MaxQuant mTRAQ results

The *proteinGroups.txt* outputs from both label-free and spiked runs were filtered for contaminants and reverse sequences. The 'LFQ intensity' column was used in the label-free samples. Ratio of 'LFQ intensity L' and 'LFQ intensity M' represented the ratio of multiplexed samples A and B in the mTRAQ results. Human proteins — theoretically present at a 1:1 ratio — were used as a reference. A scaling factor was applied such that the median human protein ratio was centered at 1, resulting in a systematic adjustment of the ratios for the other species (*E. coli* and *S. cerevisiae*).

3.2.4 DIA-NN processing of mTRAQ data

DIA-NN v1.8.1 was used. Based on *E. coli* (UP000000625, accessed 01.02.2022), *S. cerevisiae* (UP000002311, accessed 01.02.2022), and *H. sapiens* (UP000005640, accessed 01.02.2022) fasta files, a predicted library was generated for label-free analysis using default settings. mTRAQ library was made similarly, with the addition of the following settings: –fixed-mod mTRAQ 140.0949630177, nK; –original-mods.

While analyzing both label-free and mTRAQ raw files, the following settings were applied: Library Generation was set to 'IDs, RT and IM Profiling', Quantification Strategy was set to 'Peak height', scan window = 1, Mass accuracy = 10 p.p.m. and MS1 accuracy = 5 p.p.m. 'Remove likely interferences', 'Use isotopologues', and 'MBR' were enabled. Additional settings for label-free: --original-mods --peak-translation --ms1-isotope-quant --report-lib-info. Additional settings for mTRAQ: --fixed-mod mTRAQ 140.0949630177, nK --channels mTRAQ, 0, nK, 0:0; mTRAQ, 4, nK, 4.0070994:4.0070994; mTRAQ, 8, nK, 8.0141988132:8.0141988132 --original-mods --peak-translation --ms1-isotope-quant --report-lib-info.

3.2.5 Postprocessing of DIA-NN mTRAQ results

All the reports were filtered for Lib.PG.Q.Value < 0.01, leaving out only protein groups confidently identified during the library search step. PG.MaxLFQ protein group intensities were taken in the label-free approach and Ms1.Area protein group intensities were taken from the mTRAQ results. These protein abundances were then used to compute protein ratios across samples. A scaling factor was applied such that the median human protein ratio was centered at 1, resulting in a systematic adjustment of the ratios for the other species (*E. coli* and *S. cerevisiae*).

3.2.6 Common postprocessing of MaxQuant and DIA-NN mTRAQ results

At least one valid replicate with abundance higher than zero is required to calculate the mean protein group intensity across replicates of either condition A or B. Log2 fold changes (log2FC) of *E. coli*, *S. cerevisiae*, and *H. sapiens* protein groups are represented as the log2 ratio of mean protein group intensities. While the x-axis of boxplots and scatterplots uses log2FC, the y-axis of the scatterplot demonstrates a log2 sum of means of A and B per protein group.

4 Results

4.1 MultiplexDIA mode in MaxQuant

While the standard LFQ approach relies on the sequential analysis of samples, the use of non-isobaric labels allows for parallel processing of different samples in the same run, owing to their distinguishable precursor and fragment ion peaks (Figure 4.1A). Labeling strategies vary in their specificity, targeting either amine groups or the C-terminus. When the isolation window exceeds the mass difference between labeling states, co-isolation of multiple labeled precursors may occur, leading to the generation of shared and unique fragment ions. For labels that target amine groups (e.g., mTRAQ), these shared fragments correspond to y-ions that lack lysine residues. For labels consistently located on the C-terminus (e.g., SILAC), they correspond to b-ions [28]. Depending on the label type, MaxQuant determines what ions are supposed to be shared and unique between channels. While all ions contribute to peptide identification, only unique fragments are taken for quantification.

The MaxDIA workflow has been updated to accommodate multiplexed signals. Each library query is extrapolated into separate MS2 spectra for any number of labels and further used in the 'Library search' step. Independently, MS1 multiplexes are assembled during the 'Feature detection' step. The 'Multiplexing' module integrates MS1 and MS2 information from these preceding steps. Finally, the MaxLFQ algorithm has been adapted to normalize and quantify different labeled channels as separate samples (Figure 4.1B).

The first step in the MultiplexDIA workflow is the extrapolation of the predicted library to account for multiplexed states. This process utilizes the spectrum prediction model to generate unlabeled fragmentation spectra. Depending on the type and amount of non-isobaric labels, unmodified library queries are further adjusted along the m/z axis to generate labeled spectra. The type of label and the peptide sequence determine the type and the number of shared/unique peptides.

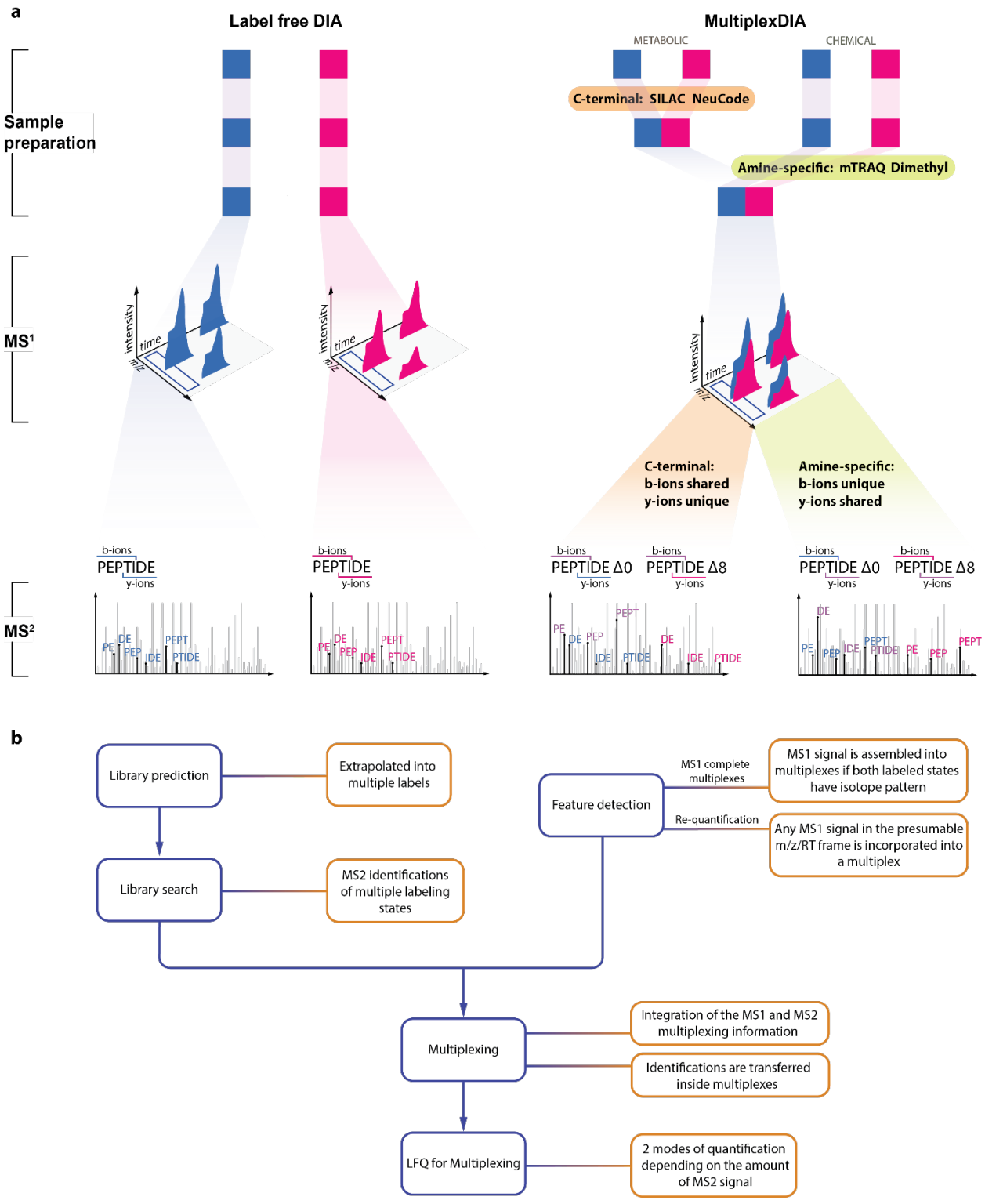


Figure 4.1: Overview. a. Schematic representation of data acquisition during label-free and labeled DIA approaches. **b.** MultiplexDIA module workflow of the MaxDIA.

Whenever extracted features are matched on the multiplexed queries in the library, they are automatically assembled into multiplexes. If some of the labeled states lack an underlying identification, multiplexes can still be detected among MS1 precursors. The MultiplexDIA workflow assesses the correct m/z distance between MS1 features that exhibit an isotope pattern, allowing a confident transfer of a single-channel identification to the remaining multiplexed MS1 features, termed ‘complete MS1 multiplex transfer’ (Figure 4.2).

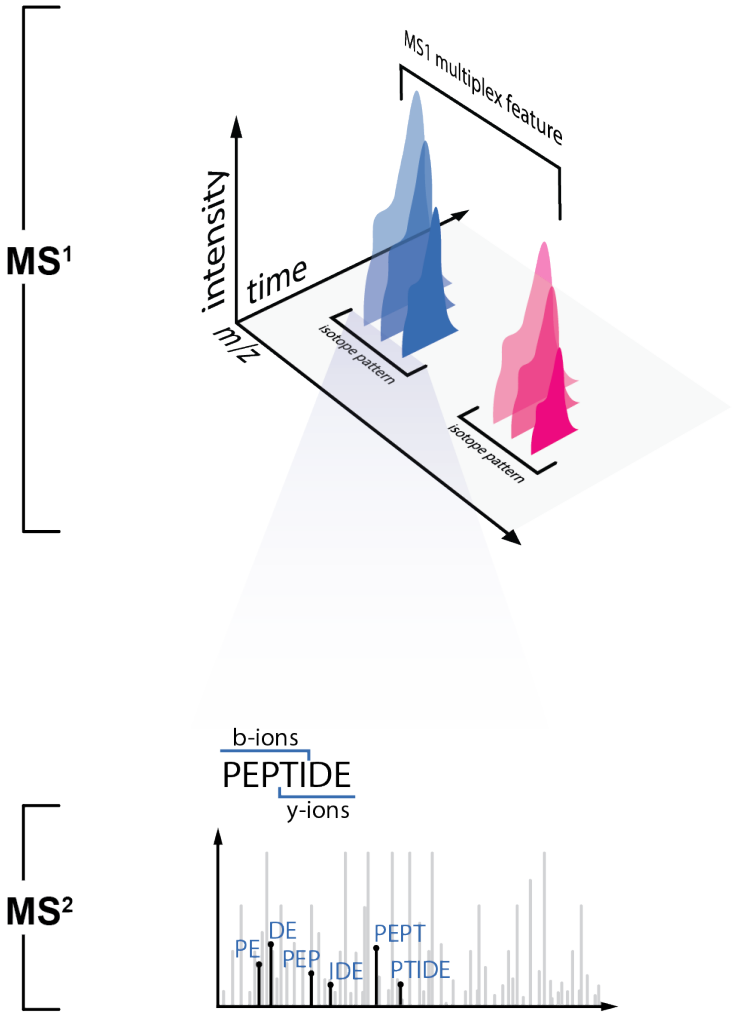


Figure 4.2: Identification transfer inside the complete MS1 multiplex. The complete mode of the MS1 multiplex assembly requires precursors to have isotope patterns and to maintain the m/z distance corresponding to the used labels. If one of the precursors has an underlying MS2 identification, it can be transferred onto the rest of the precursors in the multiplex without one.

As multiplexing and library search are processed independently, features in the same multiplex might have differing sequence identifications. In this case, MS/MS information has prevalence over the fact that MS1 features were assembled into a multiplex, thus restricting the MS1 transfer.

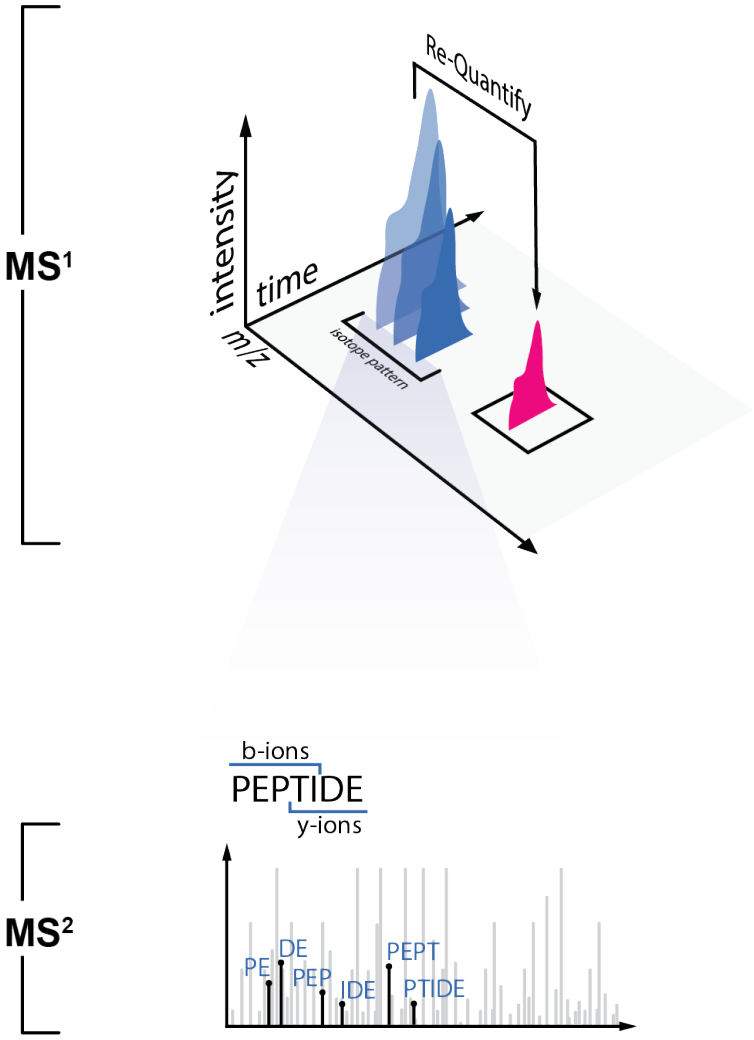


Figure 4.3: Requantify. The Re-Quantify mode of the MS1 multiplex assembly integrates any signal in the presumable m/z-RT frame of the missing channel and assigns it the identification of the present one.

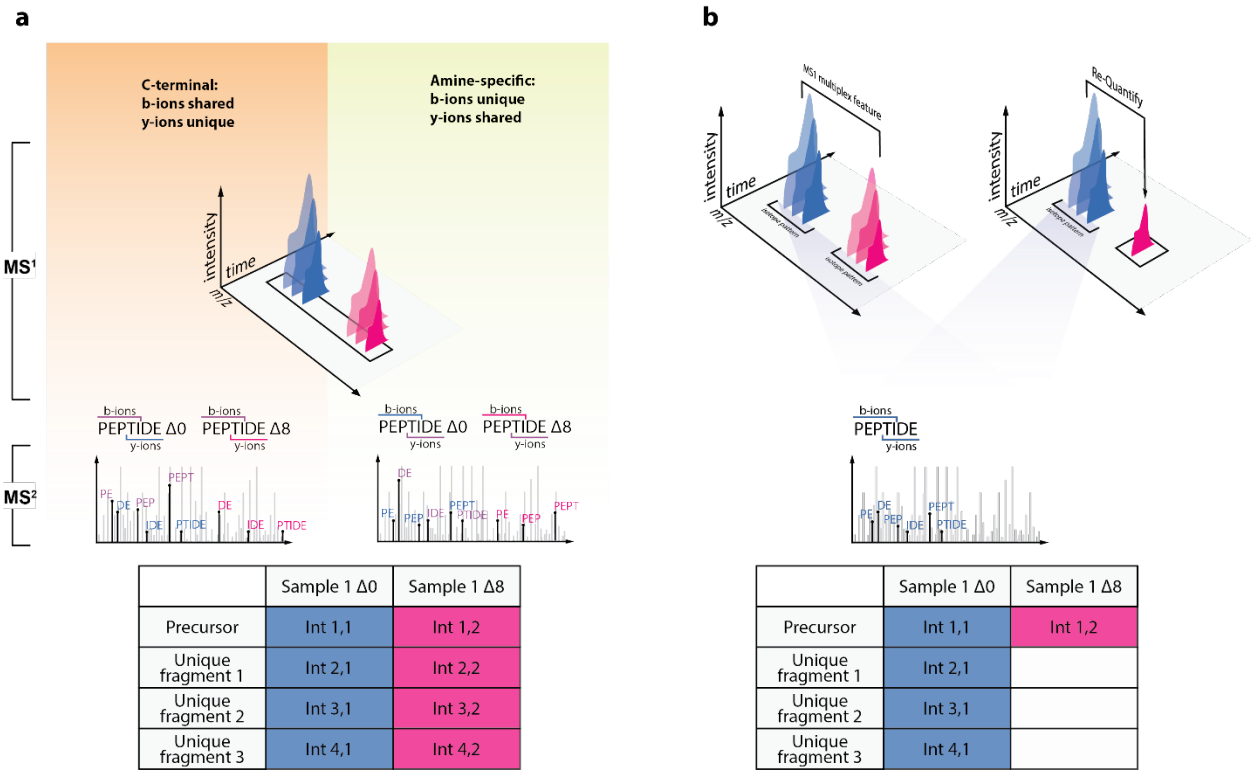
A more relaxed approach to the transfer of identifications between labels involves a 'Re-Quantify' algorithm, previously established in the DDA workflow (Figure 4.3) [22]. Even if a labeling state is entirely absent in the MS1 landscape, Re-Quantify integrates all the signal it

can find in the m/z-RT (m/z-RT-IM in case of the timsTOF data) frame of the supposed missing channel, provided that at least one channel has been identified.

During a regular MaxDIA workflow, ions shared between different identifications are reassigned to the higher-scoring match and excluded from lower-scoring matches in the 'Second matching' step. This approach is based on the assumption that the presence of those shared ions can be explained by the presence of a single sequence-charge-modification combination, thereby preventing data overinterpretation. Multiplexing labels are treated as fixed modifications and follow the same logic: whenever precursors from the same multiplex are co-isolated, their shared fragments are retained only in the best-scoring identification.

4.2 Utilizing MaxLFQ in MultiplexDIA

Whenever one of the channels is considered to be acquired under constant conditions (reference channel), a common approach is to compare channel ratios between samples [18, 31]. This, however, limits the possible channel usage and the overall throughput of the experiment. MultiplexDIA manages to normalize channels of all samples for their independent use by utilizing the established MaxLFQ algorithm (Figure 4.4A). A generalization of the MaxLFQ algorithm treats M LC-MS runs with N labeling states as M×N samples for normalization and quantification. By default, it incorporates both MS1 and MS2 signals but can use either of those exclusively. Quantifiable signal includes precursors and unique fragments, which correspond to y-ions for peptides with C-terminal labels, b-ions for peptides with amine labels only on N-terminus, and all ions for peptides with amine labels both on N-terminus and C-terminal lysine. In the case of identification transfer inside the multiplex, only the precursor ratio is used for quantification (Figure 4.4B).



	Sample 1 Δ0	Sample 1 Δ8
Precursor	Int 1,1	Int 1,2
Unique fragment 1	Int 2,1	Int 2,2
Unique fragment 2	Int 3,1	Int 3,2
Unique fragment 3	Int 4,1	Int 4,2

	Sample 1 Δ0	Sample 1 Δ8
Precursor	Int 1,1	Int 1,2
Unique fragment 1	Int 2,1	
Unique fragment 2	Int 3,1	
Unique fragment 3	Int 4,1	

Figure 4.4: MaxLFQ with multiplexing. a. MaxLFQ treats different labelling states as separate samples for normalization and quantification. **b.** Identification transfer between labels results in a possible ratio using the precursor MS1 signal.

By incorporating both precursor and fragment signals, the DIA version of the MaxLFQ algorithm exhibits robust and reliable standard deviation distribution across any number of samples [23]. However, all MS1-transferred features (through complete MS1 multiplexes or Re-Quantify) can only exhibit precursor ratios with other samples, bringing back the DDA MaxLFQ-specific issues, namely the large ratio stabilization [7]. To improve the quantification performance without losing advantages of fragment quantification, the large ratio stabilization algorithm is enabled when there is a single precursor ratio between samples, with at least one of the precursors lacking the MS/MS identification.

4.3 Application to the bulk multiplexed dataset

4.3.1 Identification and quantification performance

Distributions of *E. coli* ratios are plotted in the form of a boxplot. Actual log10 ratios are plotted against the ground truth log10 ratios.

The boxplots are split into three parts. For MaxQuant (Figure 4.5), these are constituted by:

- Label-free sample results;
- Multiplexed results without any identification transfer between labels;
- Multiplexed results with identification transfer between complete MS1 multiplexes.

For DIA-NN (Figure 4.6), these are constituted by:

- Label-free sample results;
- Multiplexed results without any identification transfer between labels;
- Multiplexed results with identification transfer between labels using the built-in 'intensity translation' algorithm.

The quantification performance can be estimated through three parameters:

- inter-quantile range (IQR) of the ratio distribution: the narrower the range, the more precise the quantification;
- distance between the median of the ratio distribution and the target value;
- the number of ratios as a proxy of the completeness of the quantifiable data.

Each boxplot has three underlying barplots, highlighting these statistics. In the boxplot part, diagonal dashed lines indicate the expected location of the boxplots. In the barplot part, reference dashed lines are added to assist comparisons between software.

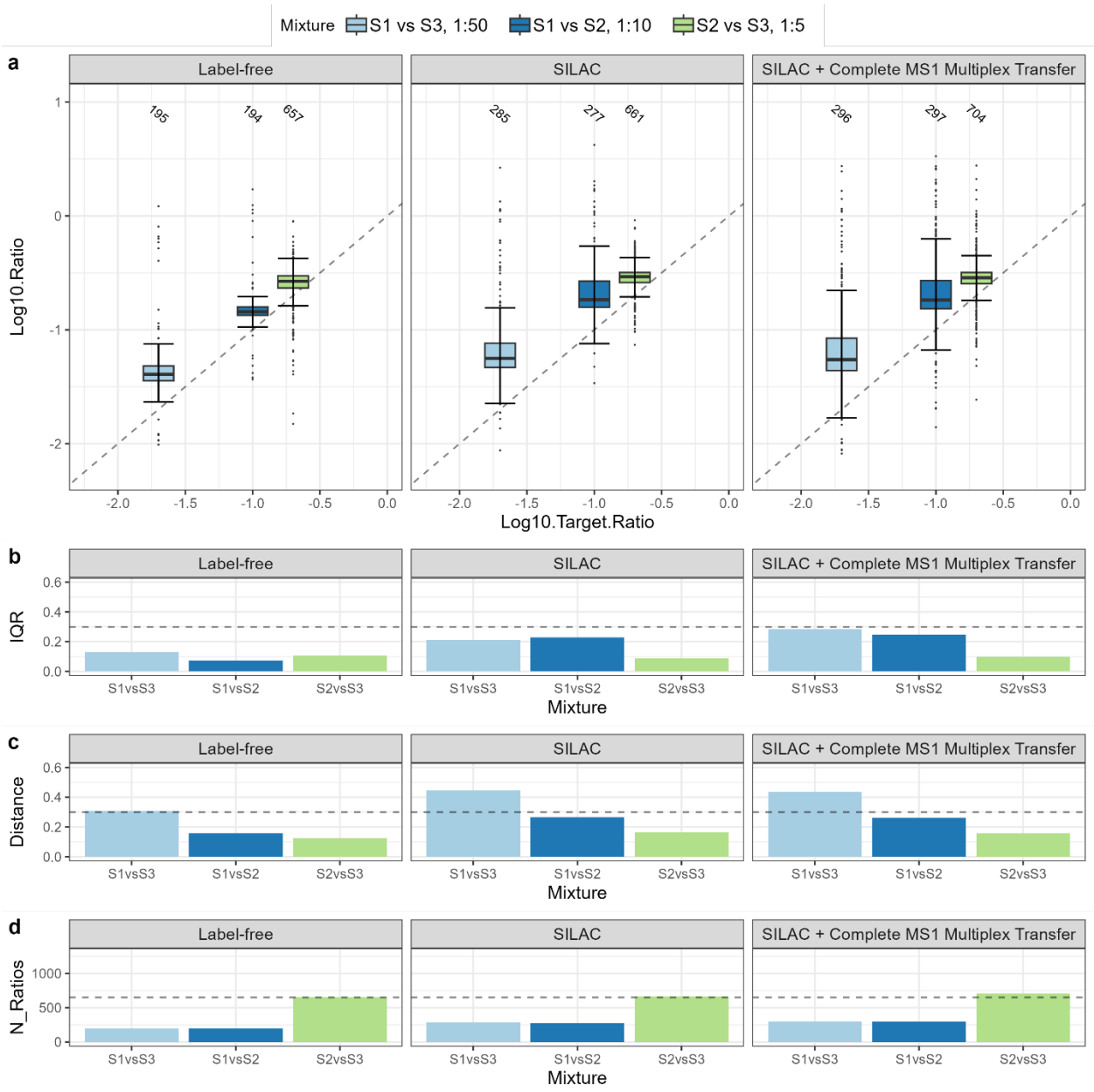


Figure 4.5: MaxQuant identification and quantification performance on the bulk dataset. **a.** Boxplots of different *E. coli* ratio distributions. Actual log10 ratios are plotted against the expected (target) log10 ratios. The number of ratios is highlighted on top of the boxplots. Plots are split into label-free samples, multiplexed samples without the label transfer, and multiplexed samples with complete MS1 multiplex label transfer. Dashed lines indicate the expected positions of the boxplots. **b-d.** Barplots, describing separate features of the corresponding boxplots: inter-quantile range (IQR), distance between the median and the target value, and number of ratios. Dashed lines serve as references across plots.

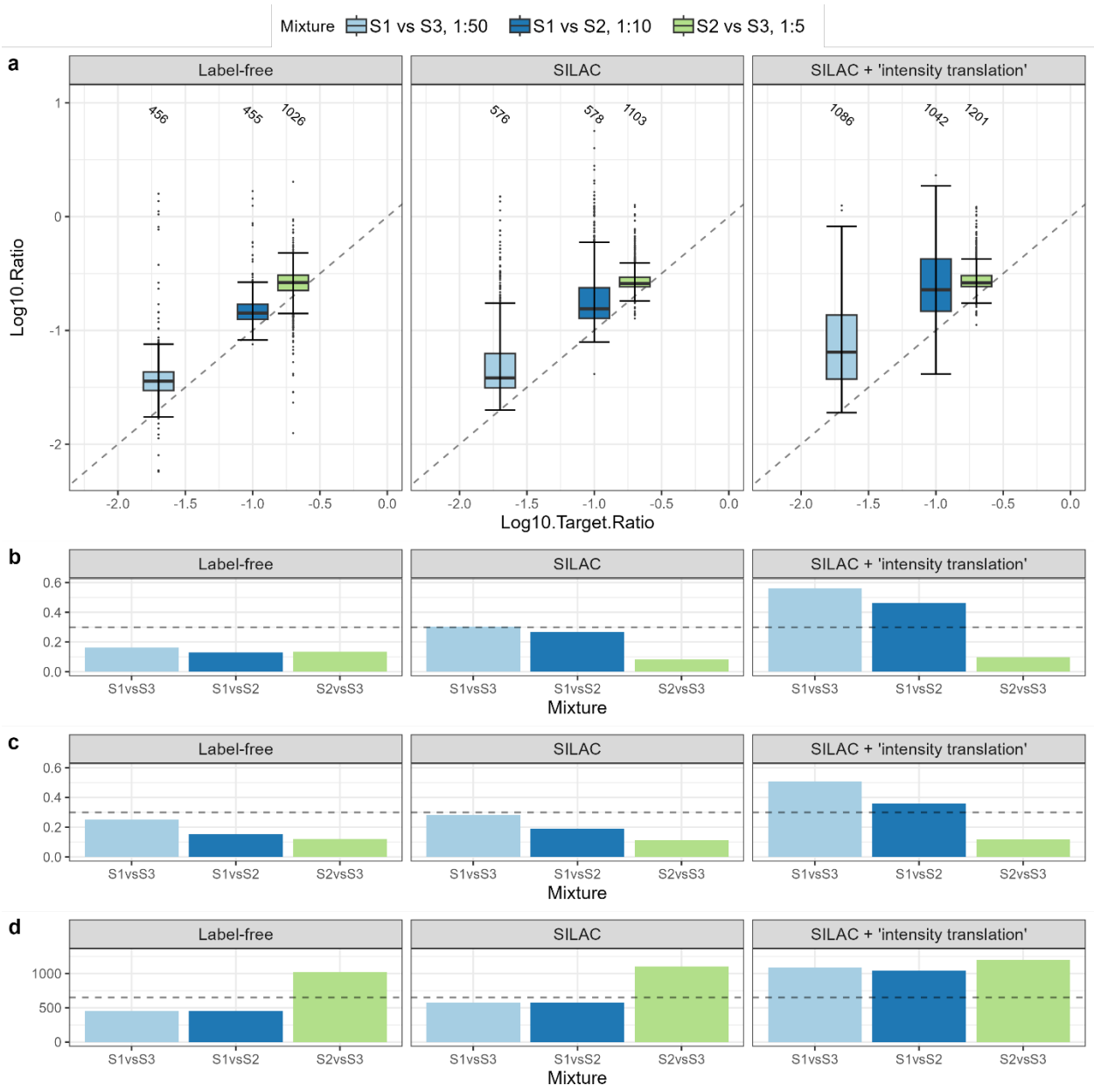


Figure 4.6: DIA-NN identification and quantification performance on the bulk dataset. a. Boxplots of different *E. coli* ratio distributions. Actual log10 ratios are plotted against the expected (target) log10 ratios. The number of ratios is highlighted on top of the boxplots. Plots are split into label-free samples, multiplexed samples without the label transfer, and multiplexed samples with 'intensity translation' label transfer. Dashed lines indicate the expected positions of the boxplots. **b-d.** Barplots, describing separate features of the corresponding boxplots: inter-quantile range (IQR), distance between the median and the target value, and number of ratios. Dashed lines serve as references across plots.

Inside MaxQuant results, one can observe the drastic increase in quantifiable S1/S3 and S1/S2 ratios by 46% and 43% when comparing label-free and multiplexed results without transfer, with IQR and distance to the median increasing by similar amounts. The complete MS1 multiplex transfer further increases the number of ratios, while slightly increasing the IQR in the S1/S3 ratios. S2/S3 steadily increases the number of ratios across all three conditions without changes in the IQR and the distance.

DIA-NN demonstrates a similar increase to MaxQuant in the IQR when comparing label-free and multiplexed results without transfer, with a slight increase in the distance. For the identification transfer between labels, DIA-NN employs an algorithm aimed at maximizing the transfer events. This results in almost doubling the number of ratios in S1/S3 and S1/S2, at a cost of doubling the IQR and the distance.

We've developed the complete MS1 multiplex transfer algorithm as an identification transfer method that doesn't interfere with quantification accuracy. In this regard, it would make sense to compare it directly to DIA-NN's multiplexing setup without the identification transfer. To make this comparison clearer, barplots of boxplot properties, grouped by the software, were made (Figure 4.7). On average, MaxQuant shows lower IQR and further distance from the median both in LFQ and SILAC comparisons. The main bottleneck for MaxQuant remains a subpar identification performance on the timsTOF data, which leads to lower numbers of ratios both in LFQ and SILAC parts. One can attribute the further median distance of MaxQuant to the incomplete ratio distribution.

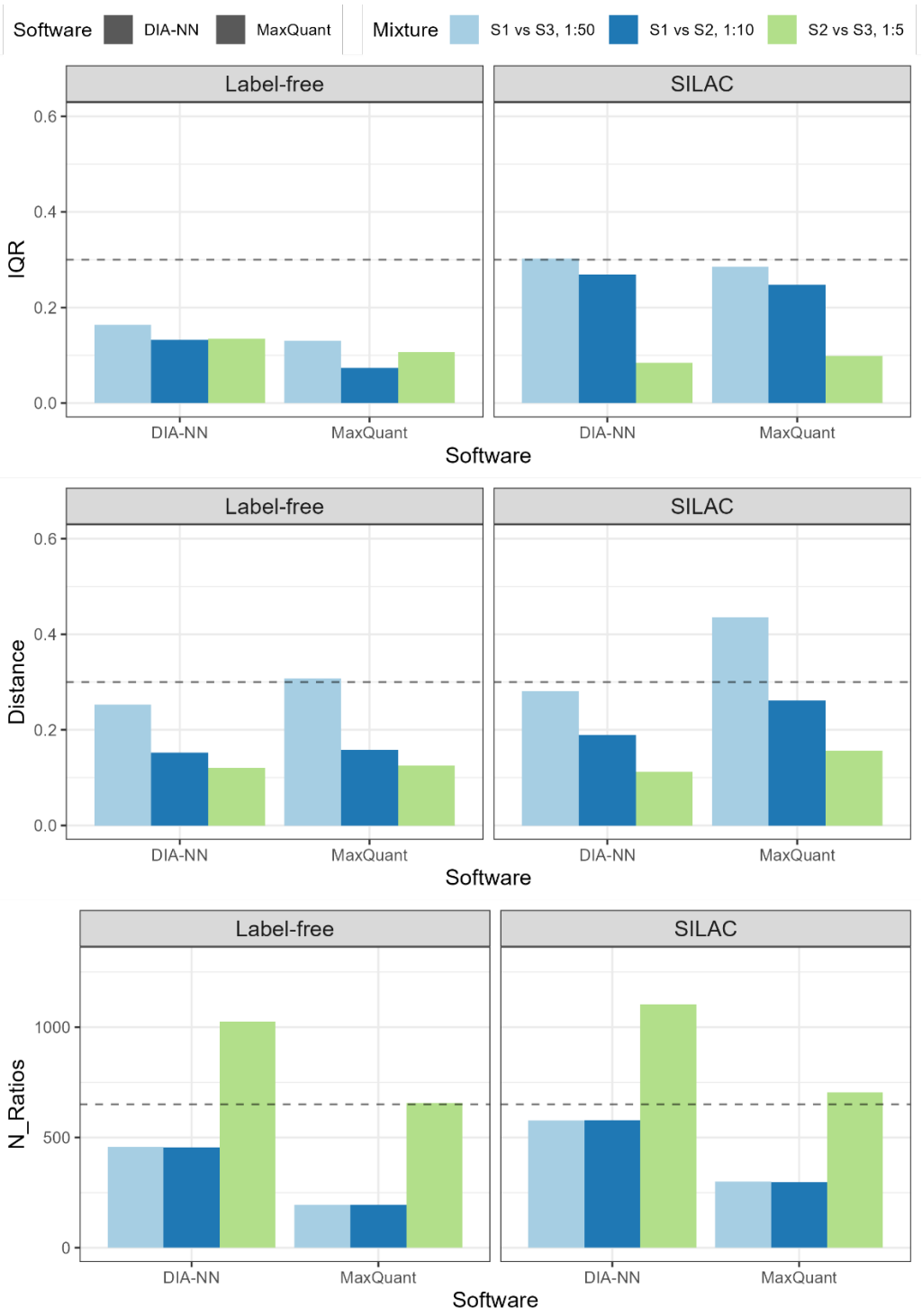


Figure 4.7: Barplots of the bulk dataset's boxplot properties, grouped by the software. LFQ part compares two label-free runs. SILAC part compares a multiplexed run without transfer on the DIA-NN's side and a multiplexed run with complete MS1 multiplex transfer on the MaxQuant's side. Dashed lines serve as references across plots.

4.3.2 False discovery rate control over differentially abundant protein groups

As described in section 3.1.7, this benchmark defines false positives as differentially abundant *H. sapiens* protein groups between samples in the significant area of p-value < 0.01 (-log10(p-value) > 2) and log2FC < -1. To compare the false discovery rate (FDR) performance of MaxQuant's multiplexing module with complete MS1 multiplex transfer and DIA-NN's multiplexing module without identification transfer, we split each distribution into protein groups identified both by multiplexed and label-free approaches, and protein groups identified exclusively through the use of multiplexing. Venn diagrams show the number of *E. coli* protein groups in each case for different sample ratios (Figure 4.8).

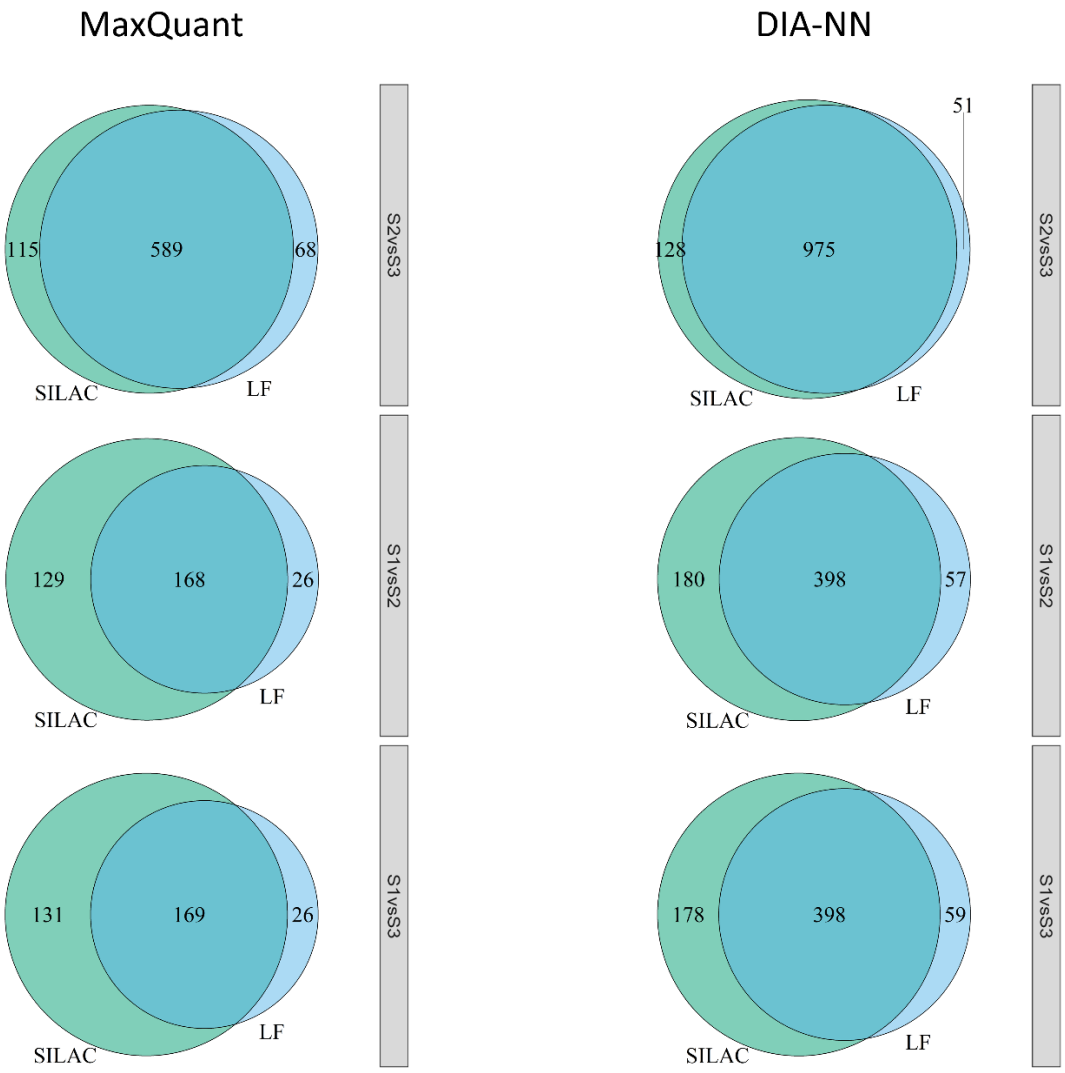


Figure 4.8: Venn diagrams of the *E. coli* protein group numbers. Label-free (LF, blue) and multiplexed (SILAC, green) approaches are intersected. Plots are separated by different sample ratios and software.

First, the volcano plot representation of the intersected ratios subset was investigated (Figure 4.9, 4.10), as the most direct way to compare the FDR performance between the label-free and multiplexed approaches. Interestingly, the MaxQuant label-free distribution of the S2/S3 ratios corresponds better to the expected log2FC and has less variation than that of DIA-NN. Both MaxQuant and DIA-NN demonstrate the drop in false positives in the SILAC samples compared to the label-free ones.

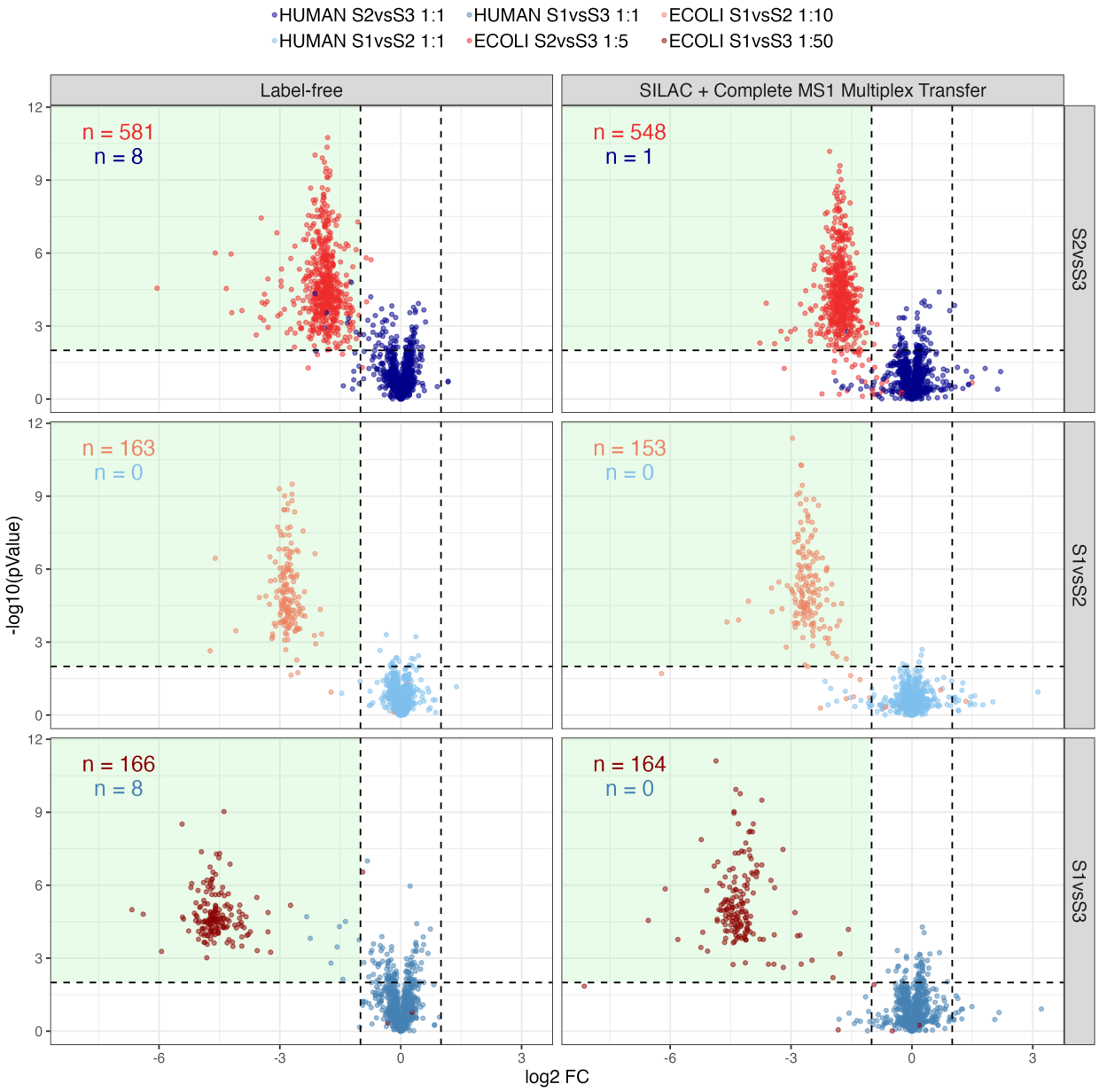


Figure 4.9: MaxQuant volcano plots of the intersected label-free/SILAC bulk results. -log10 of the p-value is plotted against the log2FC. Plots are separated by labeling approaches and increasing sample ratios. Different shades of red refer to the *E. coli* protein groups and blue to the *H. sapiens* protein groups. The numbers of those protein groups in the significant area (p-value < 0.01, log2FC < -1) are in the left upper corner. The significant area is highlighted in green.

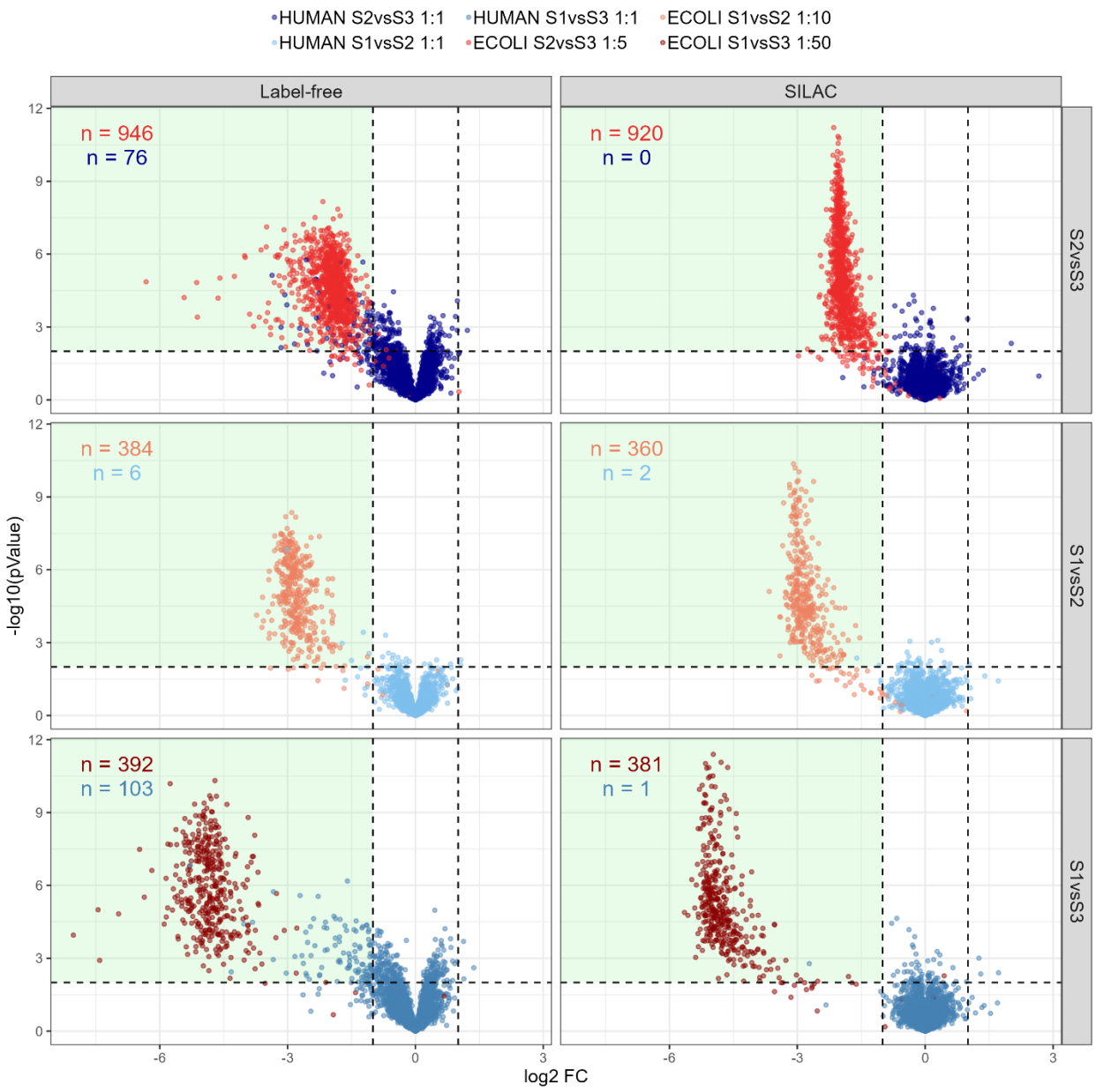


Figure 4.10: DIA-NN volcano plots of the intersected label-free/SILAC bulk results. -log10 of the p-value is plotted against the log2FC. Plots are separated by labeling approaches and increasing sample ratios. Different shades of red refer to the *E. coli* protein groups and blue to the *H. sapiens* protein groups. The numbers of those protein groups in the significant area (p-value < 0.01, log2FC < -1) are in the left upper corner. The significant area is highlighted in green.

ROC curves were made for further conclusions (Figure 4.11). Only sample ratios S2/S3 were taken due to the highest number of data points. In MaxQuant's case, both log2FC and p-value axes demonstrate the multiplexed approach outperforming the label-free one in the high-precision area, and the same or lower performance when moving towards the lower precision values. In DIA-NN's case, both log2FC and p-value axes show better FDR control for the SILAC dataset. One can also observe prominent differences between DIA-NN's and MaxQuant's label-free curves, with the latter showing better performance along the log2FC and p-value axes.

To investigate the subset of the data that was detected only by multiplexed approaches of MaxQuant and DIA-NN (Figure 4.8), volcano plots across different sample ratios were used (Figure 4.12, 4.13). One can see that the majority of the fold-change variation present in the total distribution of both software (Figure 4.5, 4.6) arises from SILAC-exclusive protein groups. Still, MaxQuant and DIA-NN are capable of distinguishing most of the *E. coli* significant and *H. sapiens* non-significant distributions.

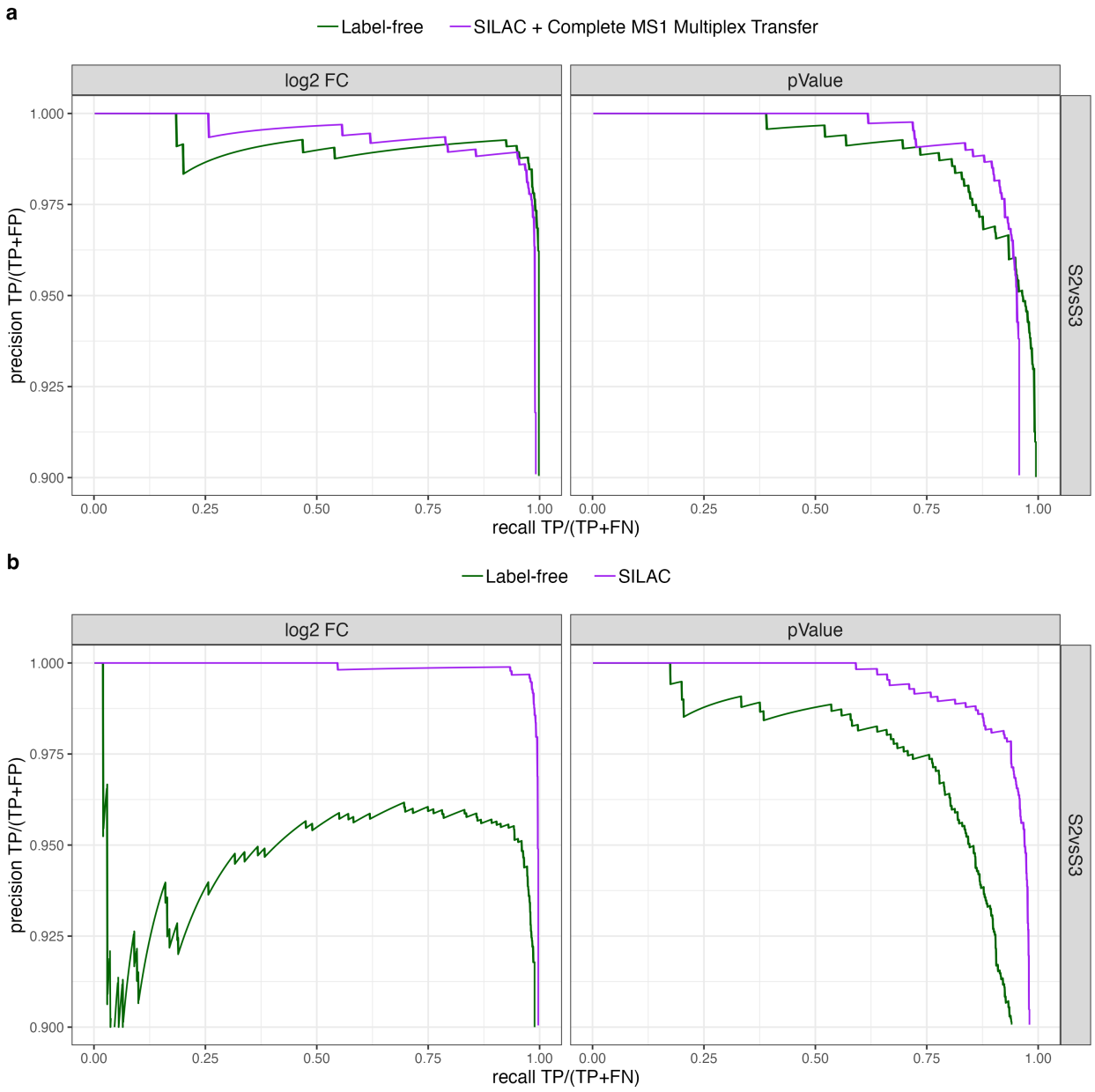


Figure 4.11: ROC-curves of the S2/S3 sample ratios. a. MaxQuant FDR performance along log2FC and Student's p-value axes. Precision is plotted against the recall. Green curves represent label-free samples, while purple curves correspond to SILAC-multiplexed samples. **b.** DIA-NN FDR performance along log2FC and Student's p-value axes.

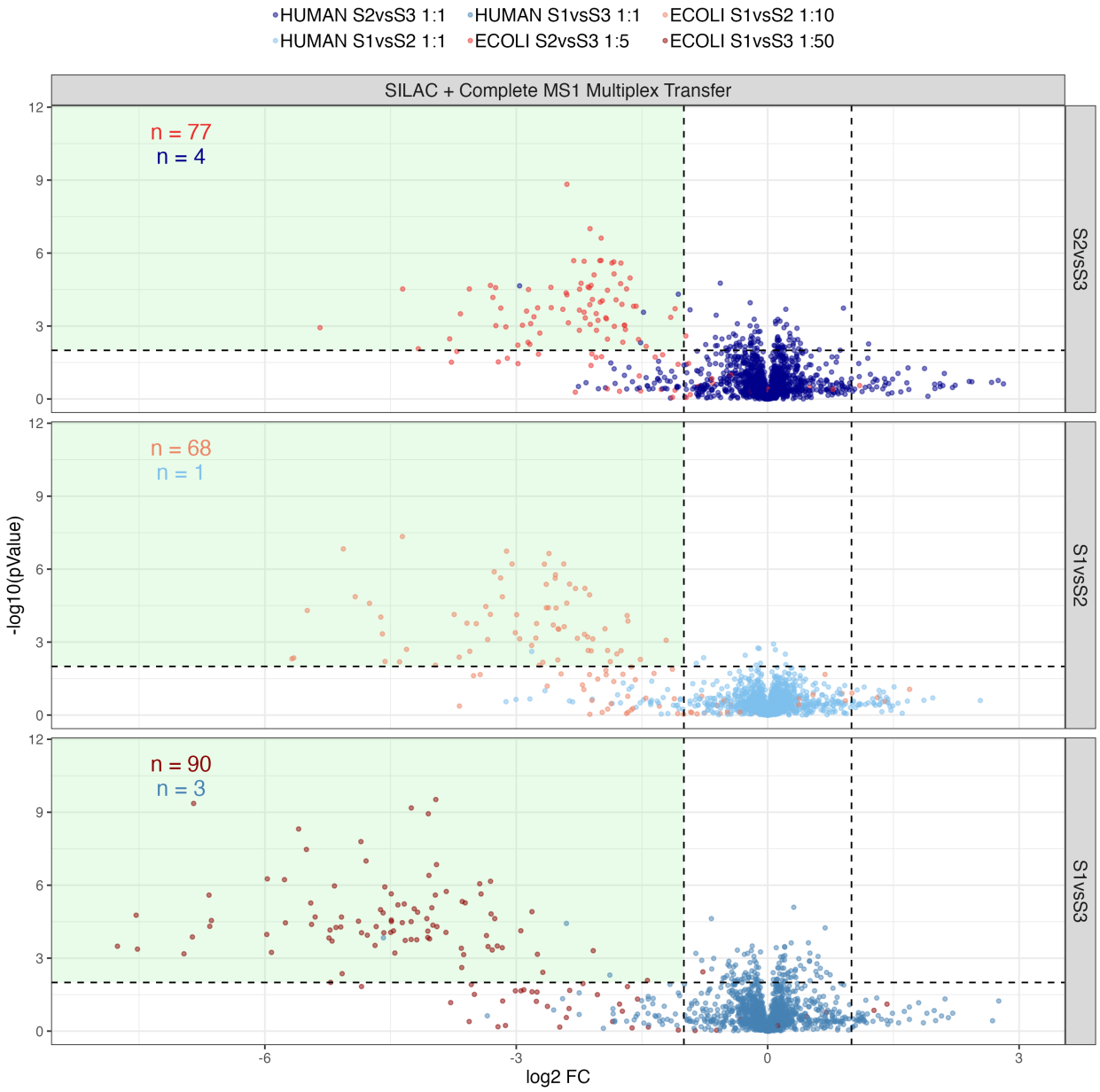


Figure 4.12: MaxQuant volcano plots of the exclusive SILAC bulk results. -log10 of the p-value is plotted against the log2FC. Plots are separated by labeling approaches and increasing sample ratios. Different shades of red refer to the *E. coli* protein groups, while blue ones correspond to the *H. sapiens* protein groups. The numbers of those protein groups in the significant area (p-value < 0.01, log2FC < -1) are in the left upper corner. The significant area is highlighted in green.

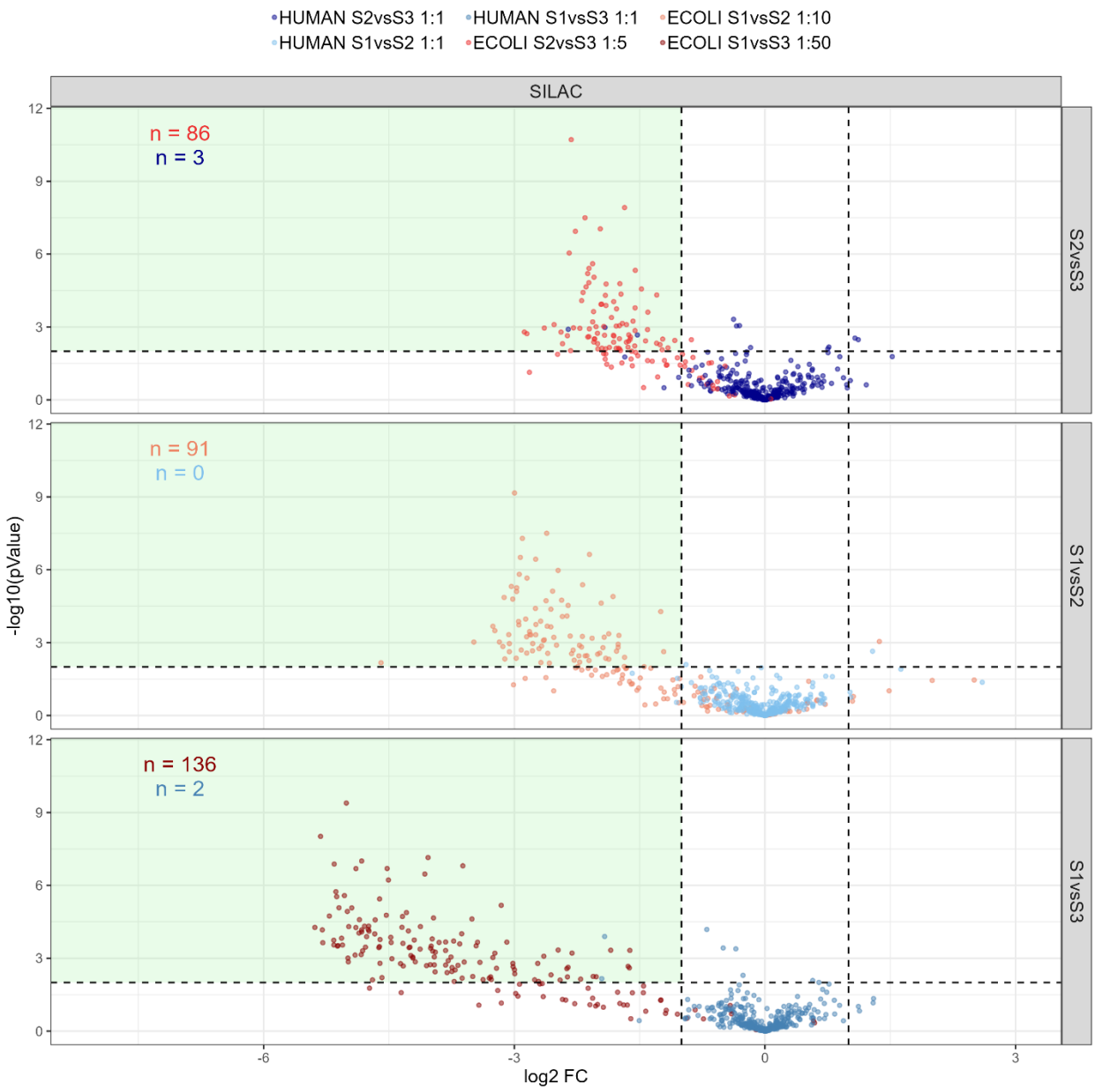


Figure 4.13: DIA-NN volcano plots of the exclusive SILAC bulk results. -log10 of the p-value is plotted against the log2FC. Plots are separated by labeling approaches and increasing sample ratios. Different shades of red refer to the *E. coli* protein groups, while blue ones correspond to the *H. sapiens* protein groups. The numbers of those protein groups in the significant area (p-value < 0.01, log2FC < -1) are in the left upper corner. The significant area is highlighted in green.

4.4 Application to the single-cell-like multiplexed dataset

4.4.1 Identification and quantification performance

The data analysis workflow repeats that of the bulk dataset. First, the log10 of light channel *E. coli* ratios between samples are plotted against the ground truth log10 ratios in the form of a boxplot (Figure 4.14). Compared to the bulk dataset, one cannot see a drastic difference, but a steady increase in the number of ratios between label-free, plain SILAC, and SILAC with transfer between complete MS1 multiplexes.

While the distance to the median remains close in all three conditions, the IQR remains the same or increases in ratios S2/S3, S3/S4, and S2/S4 (1:2 up to 1:10) and remains the same or decreases in ratios S3/S5 and S4/S5 (1:10 to 1:50). In general, ratios involving S5 sample show better quantification performance than those with lower amounts of proteomes. Sample S2/S5 is an exception, as it represents the most drastic 1:100 ratio and increases the IQR between the label-free and plain SILAC parts.

This dual effect may be attributed to the exclusion of shared ions during the quantification process. In low-abundance samples, multiplexing enhances identifications through transfer between labels and by improving the detection of otherwise weak shared fragmentation ions. However, only precursor and unique fragment ions contribute to the quantifiable signal, and the detection of the full unique fragment series in low-abundant samples is less consistent compared to the higher-abundant ones.

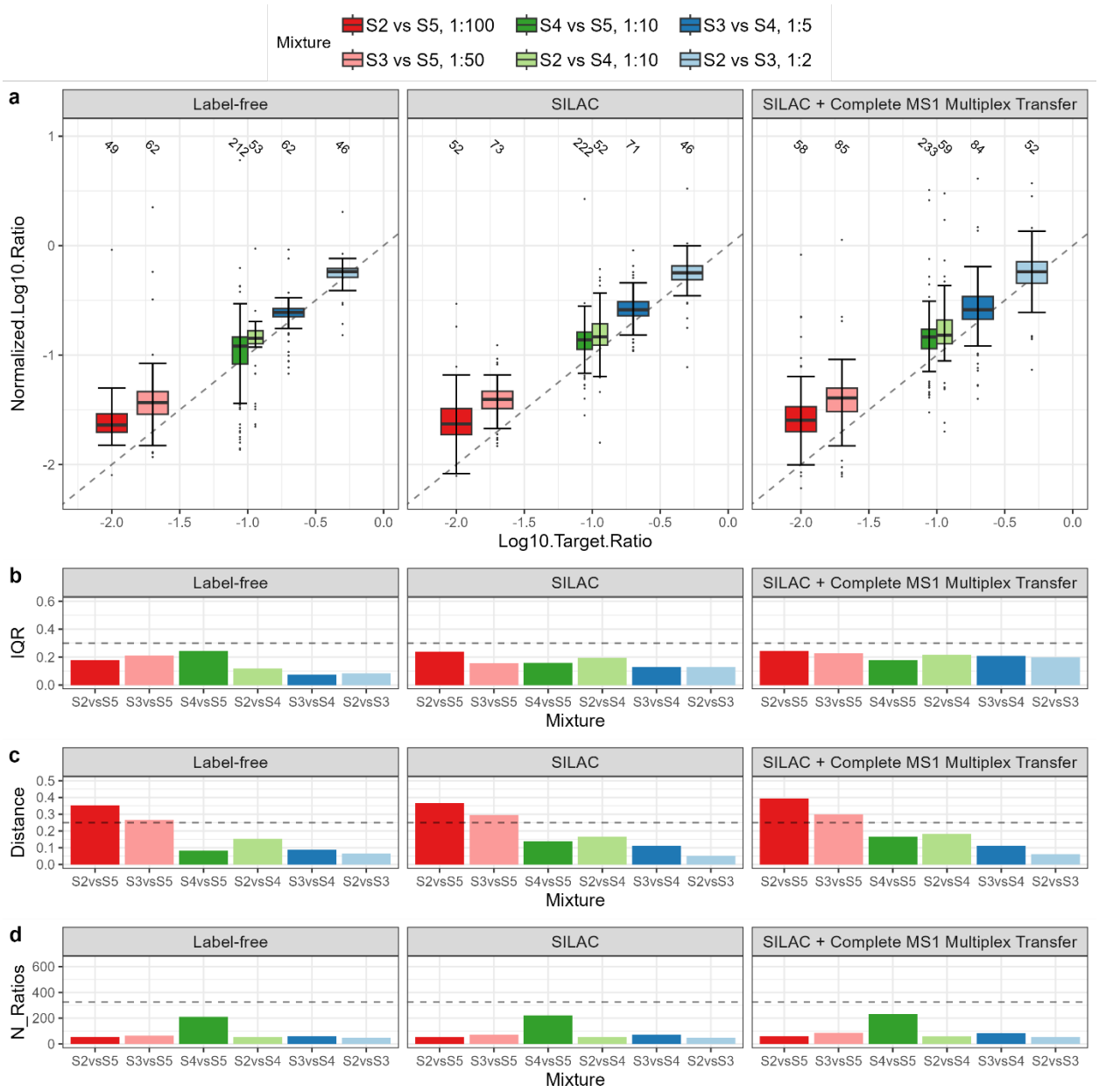


Figure 4.14: MaxQuant identification and quantification performance on the single-cell-like dataset. a. Boxplots of different *E. coli* ratio distributions. Actual log10 ratios are plotted against the expected (target) log10 ratios. The number of ratios is highlighted on top of the boxplots. Plots are split into label-free samples, multiplexed samples without the label transfer, and multiplexed samples with complete MS1 label transfer. Dashed lines indicate the expected positions of the boxplots. **b-d.** Barplots, describing separate features of the corresponding boxplots: inter-quantile range (IQR), distance between the median and the target value, and number of ratios. Dashed lines serve as references across plots.

Consistent with the bulk dataset, DIA-NN demonstrates a better IQR and distance control in the plain SILAC analysis compared to the SILAC + 'intensity translation' identification transfer, although the latter approach results in a higher number of identifications (Figure 4.15). An outstanding observation here is that ratios S2/S5, S2/S4, and S2/S3 are all decreasing in numbers, as well as IQR and distance, when comparing the label-free and SILAC approaches, suggesting that DIA-NN's multiplexing module may hinder identification performance in very low-abundant samples like S2 (Figure 4.15, 4.16).

MaxQuant's label-free analysis, on average, shows lower IQR than DIA-NN's counterpart. Distance is similar between the two except for the most extreme ratios S2/S5 and S3/S5 (Figure 4.17). Comparing MaxQuant's SILAC + complete MS1 multiplex transfer module with the SILAC module of DIA-NN, one can observe an increased IQR and distance for MaxQuant. This, however, can be explained by the consistent increase in identification rate across label-free and SILAC approaches in MaxQuant's case, in contradiction to DIA-NN. The total identification rate across label-free and SILAC approaches remains a major issue in MaxQuant analysis of the timsTOF data.

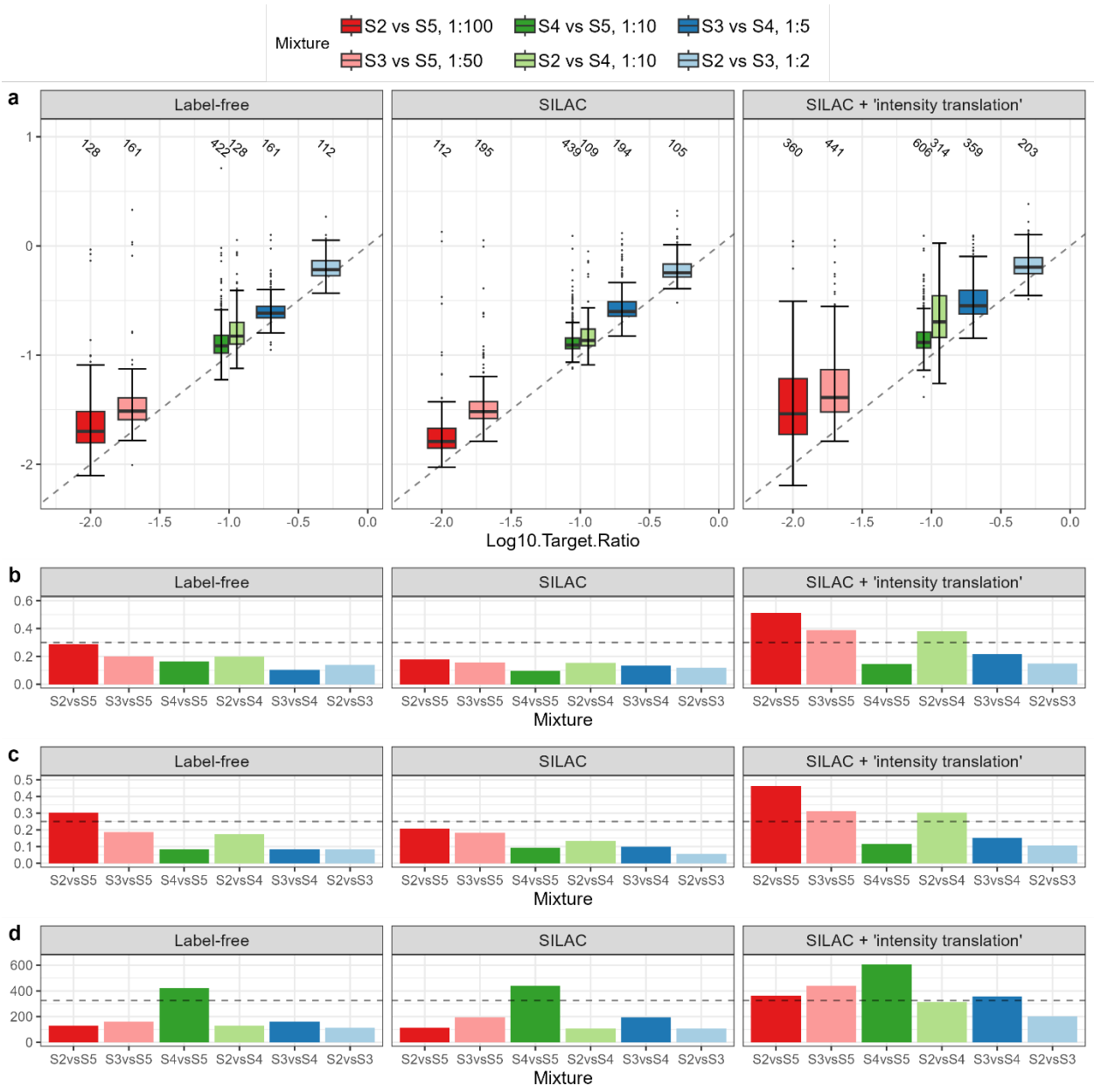


Figure 4.15: DIA-NN identification and quantification performance on the single-cell-like dataset. a. Boxplots of different *E. coli* ratio distributions. Actual log10 ratios are plotted against the expected (target) log10 ratios. The number of ratios is highlighted on top of the boxplots. Plots are split into label-free samples, multiplexed samples without the label transfer, and multiplexed samples with 'intensity translation' label transfer. Dashed lines indicate the expected positions of the boxplots. **b-d.** Barplots, describing separate features of the corresponding boxplots: inter-quantile range (IQR), distance between the median and the target value, and number of ratios. Dashed lines serve as references across plots.

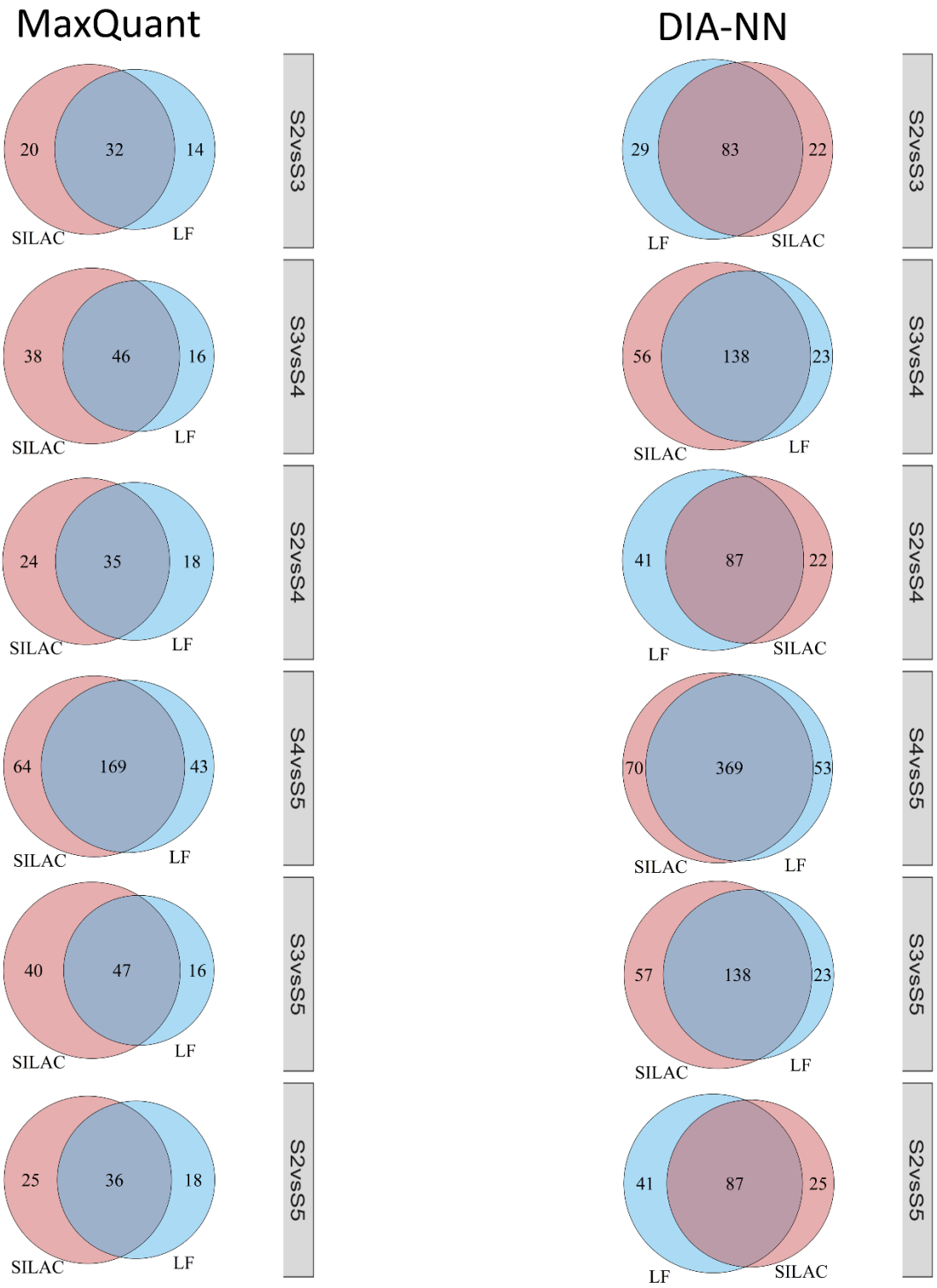


Figure 4.16: Venn diagrams of the *E. coli* protein group numbers. Label-free (LF, blue) and multiplexed (SILAC, red) approaches of both software are intersected. Plots are separated by different sample ratios and software.

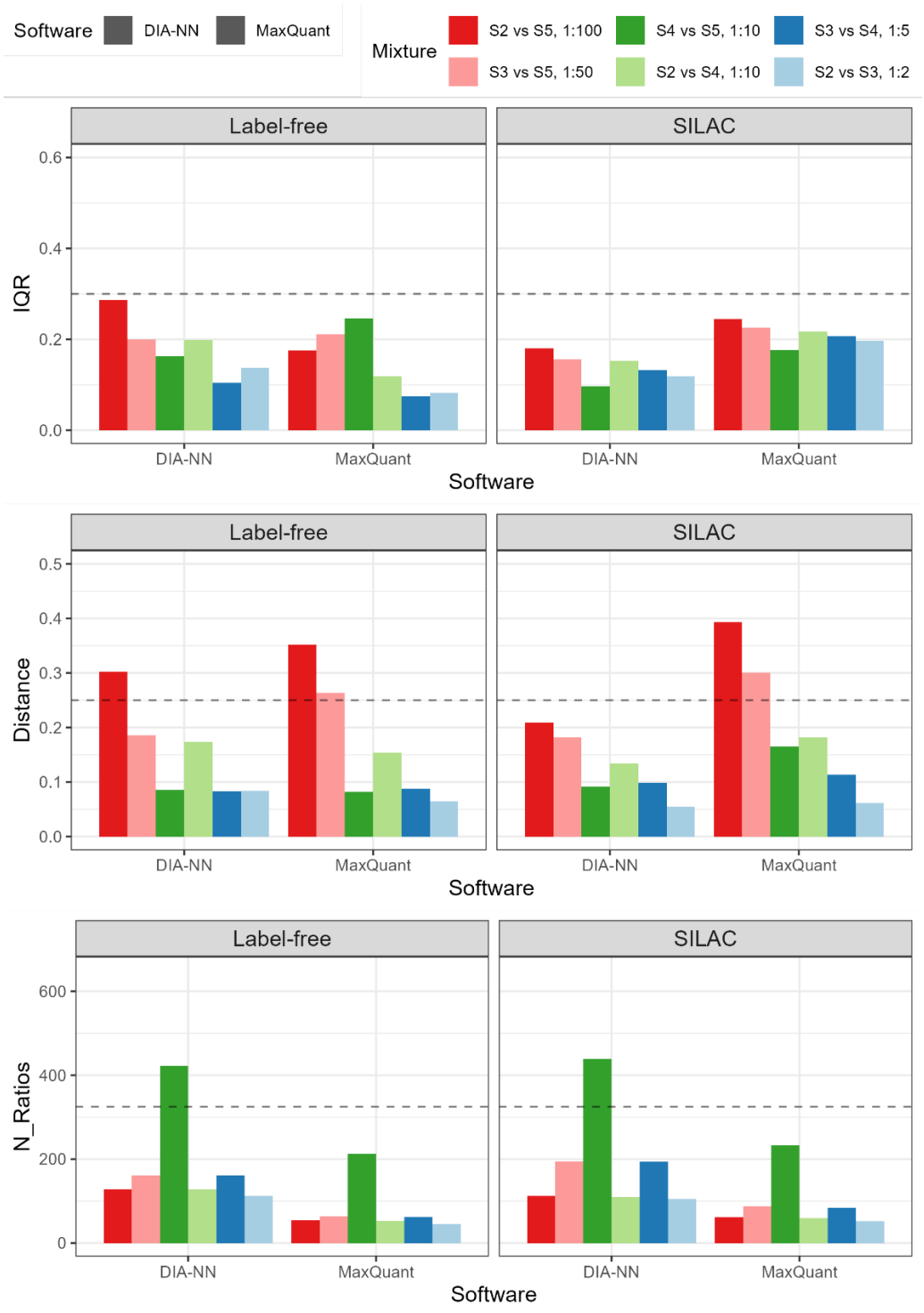


Figure 4.17: Barplots of the single-cell-like dataset's boxplot properties, grouped by the software. LFQ part compares two label-free runs. SILAC part compares a multiplexed run without transfer on the DIA-NN's side and a multiplexed run with complete MS1 multiplex transfer on the MaxQuant's side. Dashed lines serve as references across plots.

4.4.2 False discovery rate control over differentially abundant protein groups

In single-cell-like datasets, the primary objective is to maximize the number of differentially abundant proteins while maintaining control over false positives. To evaluate this, the total amount of protein groups for MaxQuant's multiplexing module with complete MS1 multiplex transfer and DIA-NN's multiplexing module without identification transfer, as well as for label-free approaches, was visualized in a volcano-plot manner (Figures 4.18, 4.19).

For sample ratios S2/S3, S3/S4, S3/S5, and S4/S5, both software tools increase the number of truly significant *E. coli* identifications while simultaneously reducing false-positive *H. sapiens* hits when comparing SILAC with label-free data. In contrast, ratios S2/S4 and S2/S5 show less consistency in MaxQuant and DIA-NN results. Both software show the decrease in the number of differentially abundant *E. coli* and *H. sapiens* protein groups for the ratio S2/S4. MaxQuant increases the number of true positives and false positives in S2/S5, while DIA-NN demonstrates the opposite effect by decreasing both of them.

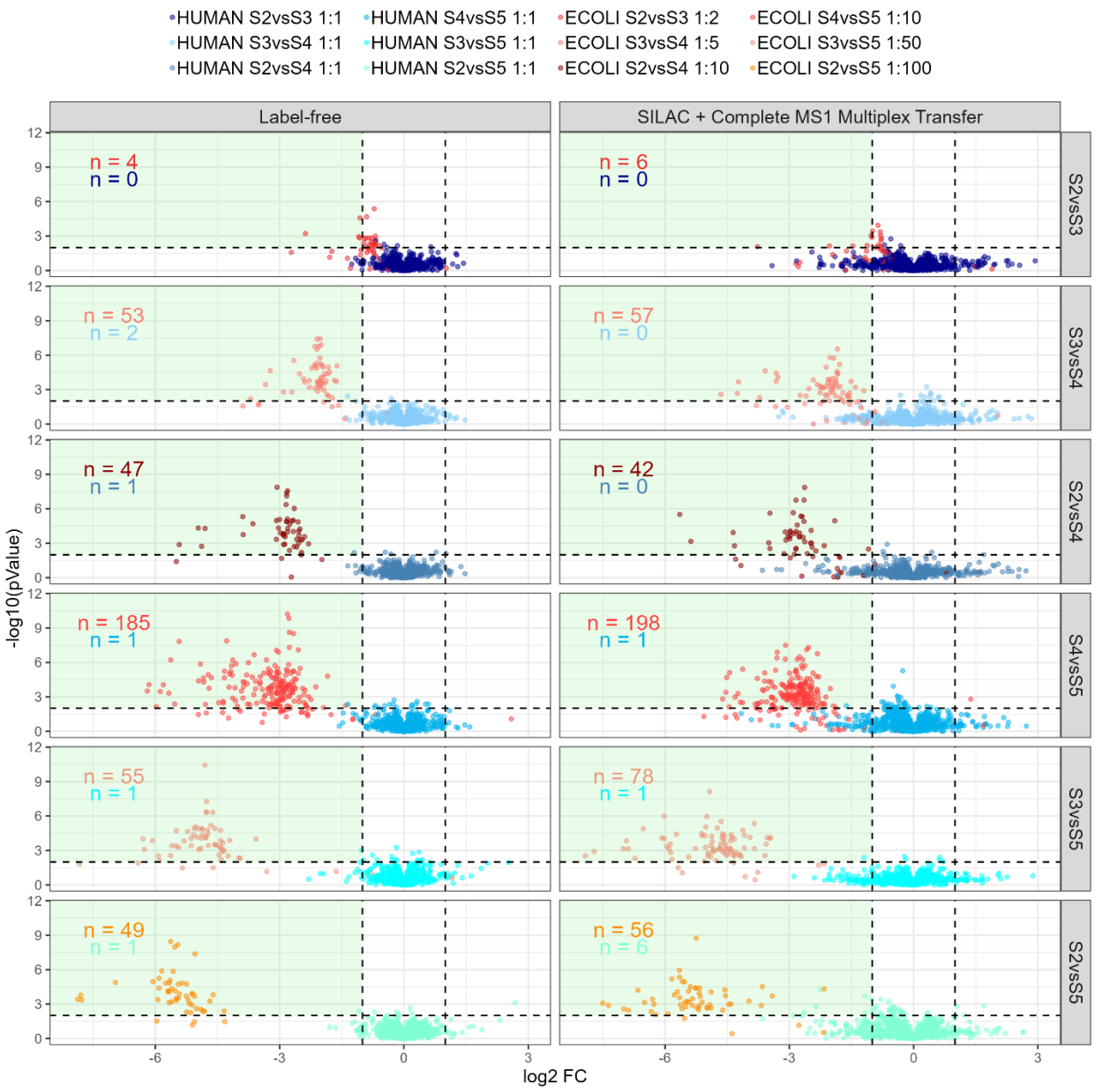


Figure 4.18: MaxQuant volcano plots of the total label-free/SILAC single-cell-like results. -log10 of the p-value is plotted against the log2FC. Plots are separated by labeling approaches and increasing sample ratios. Different shades of red refer to the *E. coli* protein groups and blue to the *H. sapiens* protein groups. The numbers of those protein groups in the significant area (p-value < 0.01, log2FC < -1) are in the left upper corner. The significant area is highlighted in green.

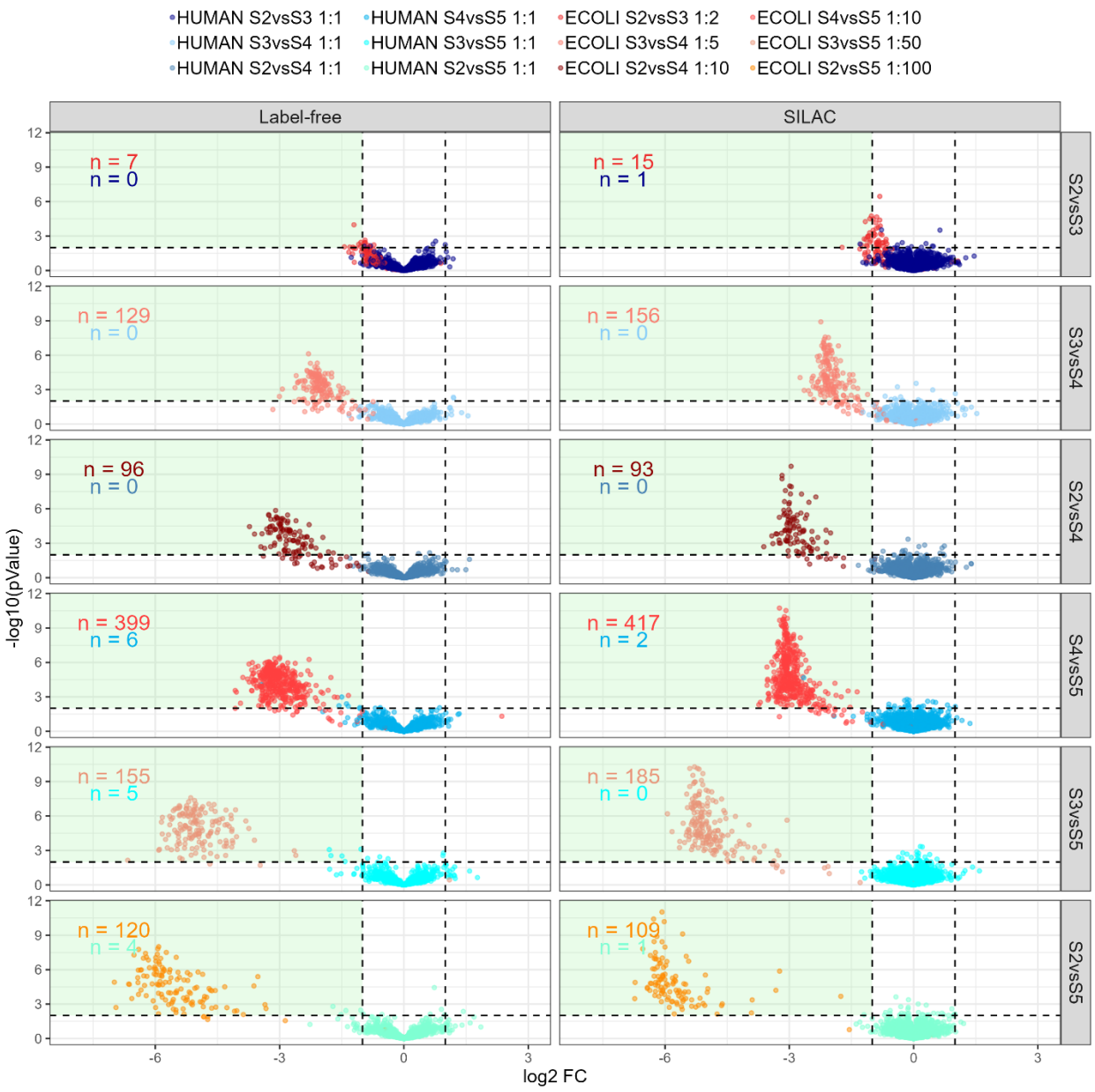


Figure 4.19: DIA-NN volcano plots of the total label-free/SILAC single-cell-like results. -log10 of the p-value is plotted against the log2FC. Plots are separated by labeling approaches and increasing sample ratios. Different shades of red refer to the *E. coli* protein groups and blue to the *H. sapiens* protein groups. The numbers of those protein groups in the significant area (p-value < 0.01, log2FC < -1) are in the left upper corner. The significant area is highlighted in green.

4.5 Exploring the Re-Quantification algorithm performance in DIA

As an alternative to the complete MS1 multiplex transfer, we offer a Re-Quantify algorithm, which was repurposed from the multiplexed DDA workflow. This approach adopts a more relaxed strategy regarding false-positive control, aiming to maximize identification transfer between labels by integrating any detectable signal at the position of a missing label.

The bulk dataset was re-analyzed with Re-Quantify as an option for identification transfer between labels (Figure 4.20). One can observe a large increase in the number of ratios for S1/S2 and S1/S3 when comparing the plain SILAC and SILAC + Re-Quantify. This, however, is coupled with at least a three-times increase in the IQR and distance. This less stringent approach to identification transfer is comparable to the strategy implemented in DIA-NN, where a significant rise in identifications is similarly accompanied by reduced quantification accuracy relative to standard SILAC. MaxQuant provides the option to enable the Re-Quantify mode of identification transfer in the settings, although low quantification accuracy of the transferred IDs should be accounted for.

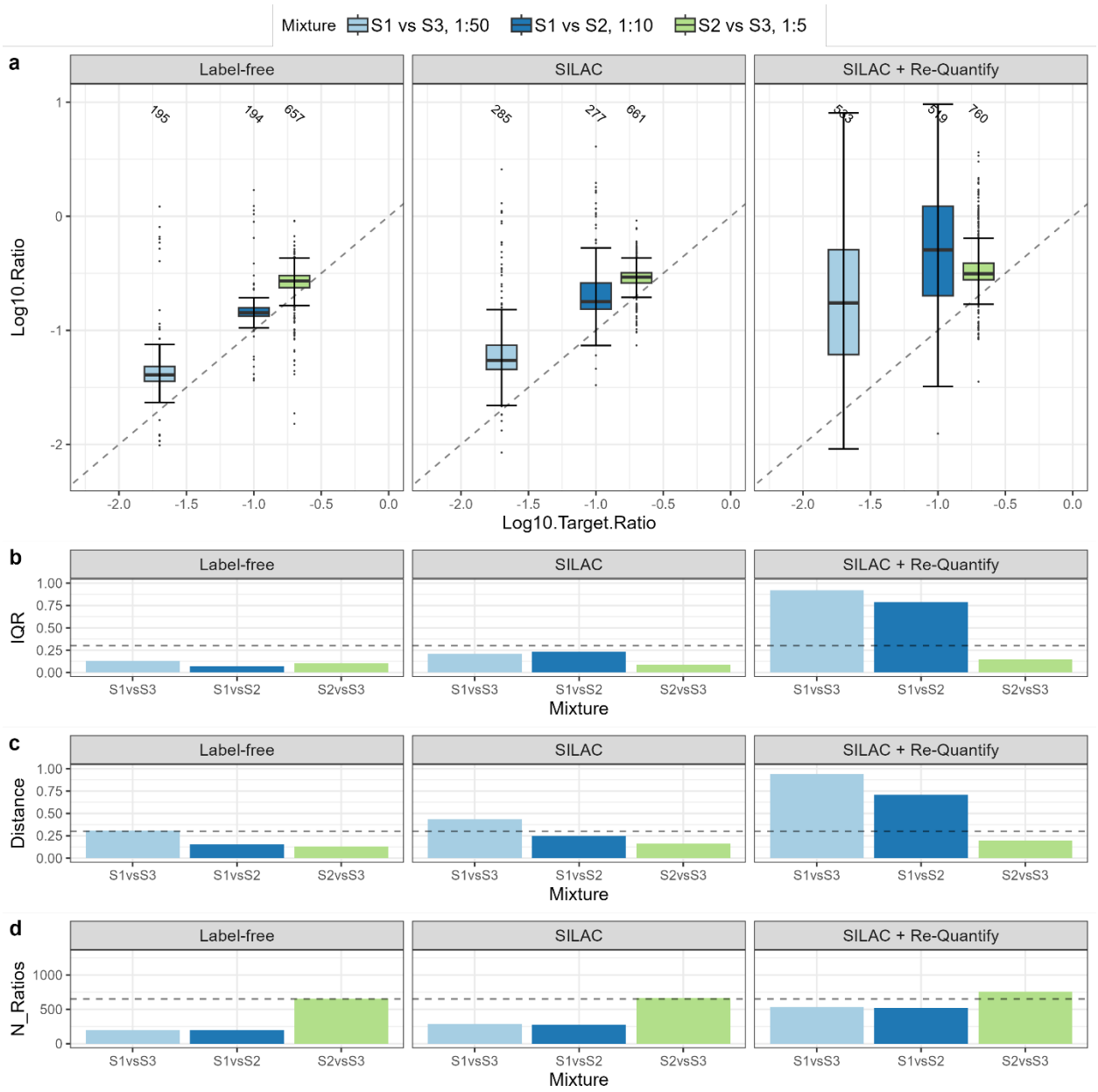


Figure 4.20: MaxQuant identification and quantification performance on the bulk dataset with Re-Quantify option enabled. a. Boxplots of different *E. coli* ratio distributions. Actual log10 ratios are plotted against the expected (target) log10 ratios. The number of ratios is highlighted on top of the boxplots. Plots are split into label-free samples, multiplexed samples without the label transfer, and multiplexed samples with Re-Quantify label transfer. **b-d.** Barplots, describing separate features of the corresponding boxplots: inter-quantile range (IQR), distance between the median and the target value, and number of ratios.

4.6 Application to the mTRAQ labeling dataset

MaxQuant results demonstrate a substantial difference in identification performance between label-free and mTRAQ analysis, with the latter underperforming (Figure 4.21). We hypothesized that this drastic difference with SILAC can be explained by the MS/MS library spectra generation and trained a separate spectrum prediction model using the results of three mTRAQ DDA files, taken from the same study [28]. Using this model, we re-analyzed the benchmarking mTRAQ DIA data and observed a 31% increase for *S. cerevisiae* and *H. sapiens* protein group ratios between light and medium channels (A/B, Figure 4.22).

Surprisingly, for DIA-NN, the total amount of protein group ratios is lower in the mTRAQ part of the data compared to the label-free one, while the IQR of the mTRAQ results increases (Figure 4.23).

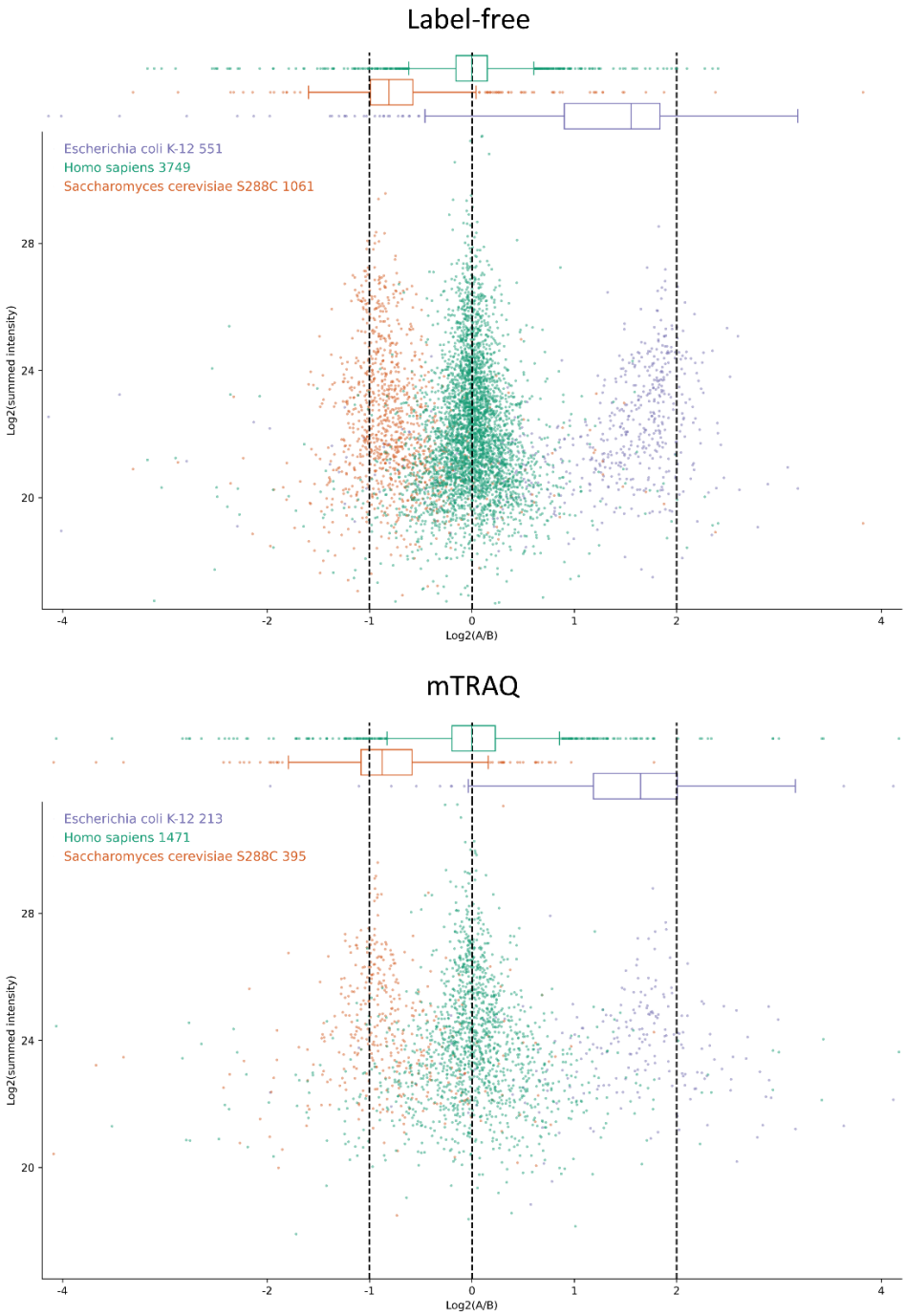


Figure 4.21: MaxQuant identification and quantification performance on the mTRAQ dataset. Log2 of summed intensities between light and medium channels is plotted against the log2 of the corresponding ratio. Plots are separated into label-free and mTRAQ multiplexed samples. Different species-specific ratios are highlighted with separate colors and expected ratio values: *E. coli* – purple, 2; *H. sapiens* – green, 0; *S. cerevisiae* – orange, -1.

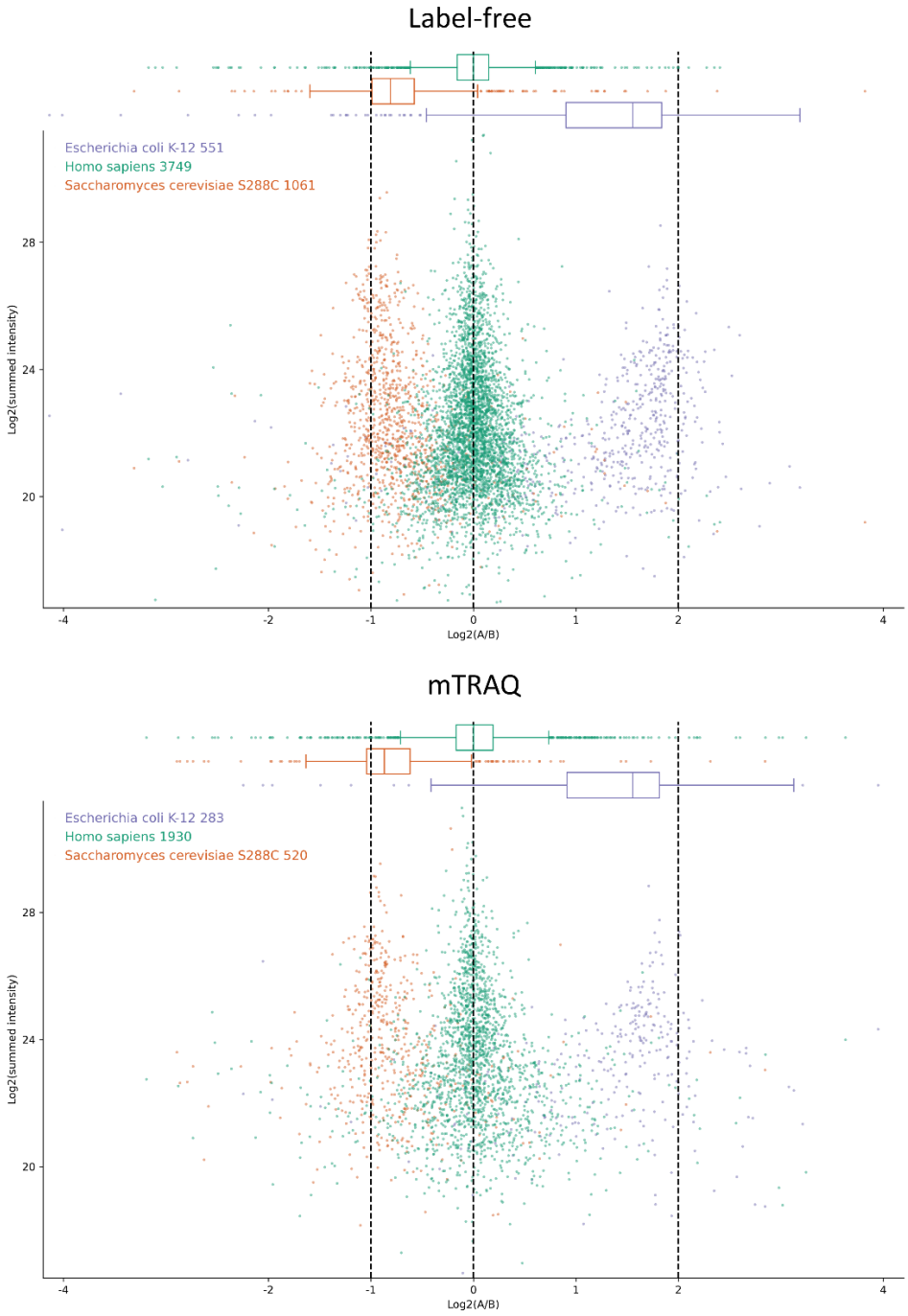


Figure 4.22: MaxQuant identification and quantification performance on the mTRAQ dataset using mTRAQ-specific spectrum prediction.

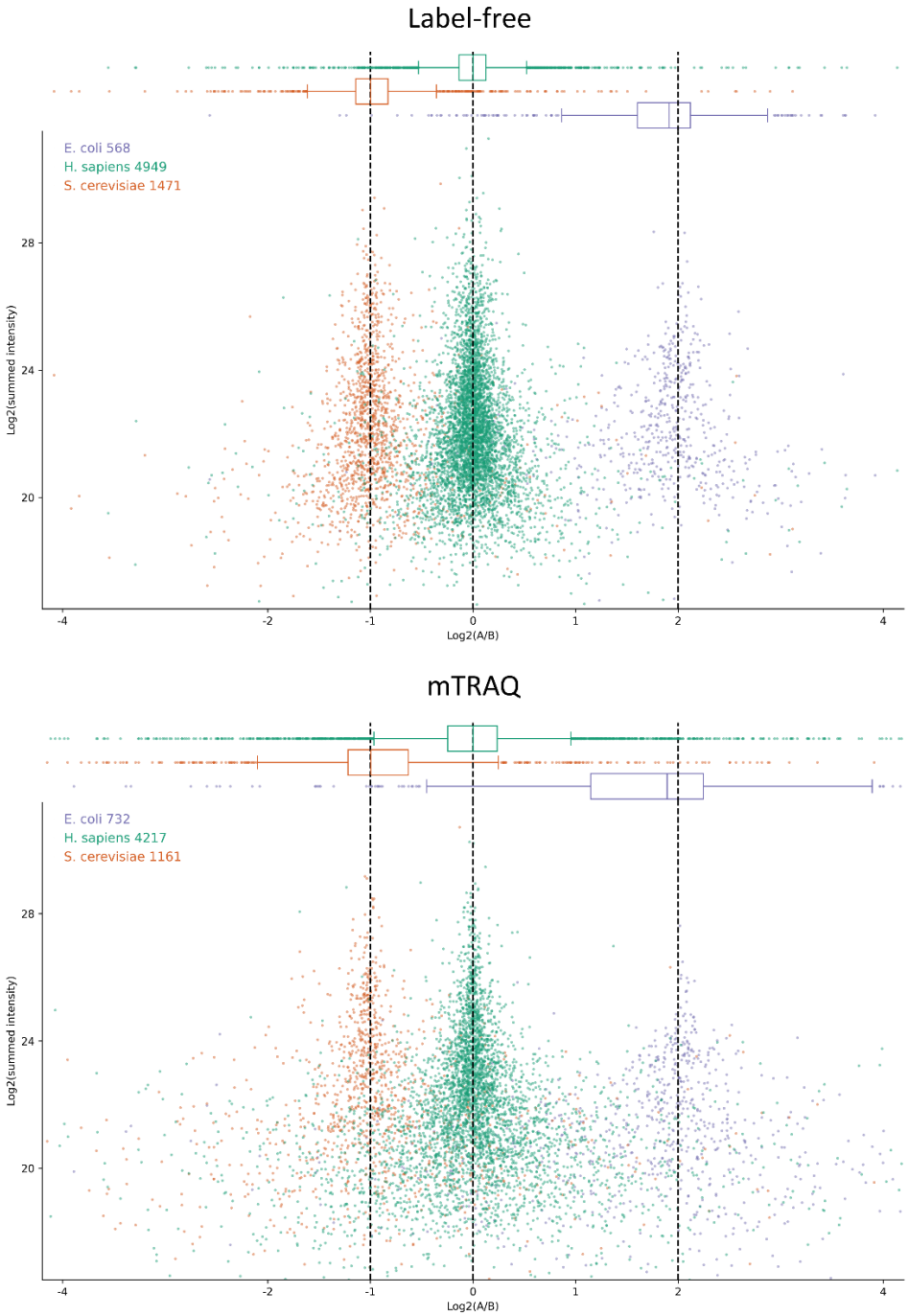


Figure 4.23: DIA-NN identification and quantification performance on the mTRAQ dataset. Log2 of summed intensities between light and medium channels is plotted against the log2 of the corresponding ratio. Plots are separated into label-free and mTRAQ multiplexed samples. Different species-specific ratios are highlighted with separate colors and expected ratio values: *E. coli* – purple, 2; *H. sapiens* – green, 0; *S. cerevisiae* – orange, -1.

5 Discussion

5.1 MaxDIA and DIA-NN analysis of the SILAC data

Compared to the label-free approach, the MultiplexDIA module of MaxDIA demonstrates identification, quantification, and false discovery improvements that are also observed in DIA-NN's plexDIA results.

MultiplexDIA consistently increases the number of quantifiable ratios both in bulk and in single-cell-like SILAC results, while maintaining quantification metrics (IQR and the distance to the expected median) at a comparable level with DIA-NN. Only the single-cell-like SILAC ratio distributions of DIA-NN demonstrate a lower IQR and a shorter distance. However, if no translation of identifications is enabled in DIA-NN, the number of ratios between samples drops in the single-cell-like SILAC part compared to the label-free one, which is not the case for MultiplexDIA of MaxQuant. Such a discrepancy in the identification rate explains the difference in the quantification accuracy.

When comparing shared protein groups between label-free and SILAC bulk results, one can observe improvement of the log2FC accuracy and a higher confidence in the FDR control with fewer *H. sapiens* data points in the significant area of p-value < 0.01 and log2FC < -1. Interestingly, plain label-free results of MaxDIA demonstrate better log2FC accuracy and FDR control than those of DIA-NN.

There are several algorithmic differences between the multiplexing modules of MaxDIA and DIA-NN worth highlighting. DIA-NN's plexDIA utilizes the library search of the multiplexed precursor states, but only selects the most highly confident identification as the retention time reference point for further re-extraction of the remaining channels. One can see from the DIA-NN's plain SILAC results that this strategy doesn't compromise MS1-level quantification accuracy when all channels are identified during the library search. It makes sense, since multiplexing labels like SILAC shouldn't change eluting properties of the peptides, resulting in all the multiplexed states having the same retention time. The only way this approach can account for the quality of the translated signal is through the channel q-value. When comparing

DIA-NN's SILAC and SILAC + 'intensity translation' results of bulk datasets, one can see that applying channel and global q-value thresholds to every labeled state (SILAC part) works well. However, once the translation of missing labels is added (SILAC + 'intensity translation' part) by removing thresholds on one of the channels, it disturbs the quantification performance. It also leaves an open question about the accuracy of re-extracting the MS2-level signal, since it is much more complex. In general, the translation algorithm is likely to pick up previously identified precursor signal if there was any identification or will pick up any signal in its presumable m/z-RT-IM window.

MultiplexDIA module of MaxDIA employs an alternative strategy: it leaves the priority to the MS1/MS2 library identifications and only performs a MS1-level transfer in the absence of the latter. The 'Re-Quantify' algorithm of MaxQuant resembles that of DIA-NN and results in a similar quantification performance. On the other hand, complete MS1 multiplex transfer adds a layer of confidence to the transferred MS1 signal by accounting for the presence of an isotope pattern. An important note is that MaxQuant never transfers or re-extracts MS2-level signal, which is only acquirable through library matching and never overwritten by the transferring events.

Another advancement of the MultiplexDIA module is the utilization of the MaxLFQ algorithm. This allows for the incorporation of MS1 alongside MS2-level signals into the quantification of the multiplexes, which is generally avoided by other software, nevertheless showing comparable quantification and FDR performance on the SILAC benchmarks. As seen from the DIA-NN results postprocessing, the normalization process of the multiplexed data usually involves reserving a single channel as the reference one (heavy in this instance) to normalize precursor signals across samples. MaxLFQ circumvents it by treating each channel as a separate sample, normalizing them through the same algorithm used in label-free runs. Not only does it simplify the postprocessing, but it also frees up the presumable reference channel for biologically relevant samples.

An important issue one would have to address in MaxDIA is the overall identification rate on the timsTOF SCP instrument. The number of quantifiable ratios of DIA-NN label-free and

SILAC results, both acquired through the timsTOF SCP, is doubled compared to the corresponding results of MaxDIA.

5.2 MaxDIA and DIA-NN analysis of the mTRAQ data

Another direction of development is required for the mTRAQ data analysis in MaxDIA. Demonstrated in section 4.6 of the results, the rate of MaxDIA quantifiable ratios in label-free QE Orbitrap analysis is only around 20% lower than that of DIA-NN, which is a much better state than that of timsTOF. However, the use of the mTRAQ label itself is hindering the identification rate significantly.

Metabolic labels like Arg10 and Lys8 shift the mass of the fragments/precursors by 8 Da and do not add any additional chemical group to them, as those heavy amino acids are incorporated into the sequence of the corresponding fragments/precursors. On the other hand, three isotope forms of the mTRAQ produce mass shifts of 140, 144, and 148 Da per derivatized site, respectively [41]. Such a substantial addition to the peptide sequence changes not only the m/z, but also the intensities of the fragmentation spectra. Thus, a standard model for library prediction is not suitable for the mTRAQ analysis.

To tackle this, we trained a spectrum prediction model on the results of three DDA mTRAQ files, which led to a 31% increase for *S. cerevisiae* and *H. sapiens* protein group ratios. The issue of the low identification performance compared to the label-free approach persists, but one can focus on increasing the size of the training data to circumvent the problem.

DIA-NN also demonstrates an overall drop in the number of ratios between samples A and B when switching to the mTRAQ workflow, given that transfer between labels was disabled. There is no citable information on whether DIA-NN incorporates a specifically trained mTRAQ spectrum prediction model, but such an observation may highlight similar issues that DIA-NN experiences during the mTRAQ data analysis.

6 References

1. Tyanova, S., et al., *The Perseus computational platform for comprehensive analysis of (prote)omics data.* Nature Methods, 2016. **13**(9): p. 731-740.

2. Cox, J. and M. Mann, *Quantitative, high-resolution proteomics for data-driven systems biology.* Annu Rev Biochem, 2011. **80**: p. 273-99.

3. Eliuk, S. and A. Makarov, *Evolution of Orbitrap Mass Spectrometry Instrumentation.* Annual Review of Analytical Chemistry, 2015. **8**(1): p. 61-80.

4. Meier, F., et al., *Parallel Accumulation-Serial Fragmentation (PASEF): Multiplying Sequencing Speed and Sensitivity by Synchronized Scans in a Trapped Ion Mobility Device.* J Proteome Res, 2015. **14**(12): p. 5378-87.

5. Wells, J.M. and S.A. McLuckey, *Collision-induced dissociation (CID) of peptides and proteins.* Methods Enzymol, 2005. **402**: p. 148-85.

6. Olsen, J.V., et al., *Higher-energy C-trap dissociation for peptide modification analysis.* Nature Methods, 2007. **4**(9): p. 709-712.

7. Cox, J., et al., *Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ.* Mol Cell Proteomics, 2014. **13**(9): p. 2513-26.

8. Hu, Q., et al., *The Orbitrap: a new mass spectrometer.* Journal of Mass Spectrometry, 2005. **40**(4): p. 430-443.

9. Scheltema, R.A., et al., *The Q Exactive HF, a Benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer.* Mol Cell Proteomics, 2014. **13**(12): p. 3698-708.

10. Makarov*, A., *Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis.* February 10, 2000.

11. Meier, F., et al., *diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition.* Nature Methods, 2020. **17**(12): p. 1229-1236.

12. Meier, F., et al., *Online Parallel Accumulation–Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer.* Molecular & Cellular Proteomics, 2018. **17**(12): p. 2534-2545.

13. Boesl, U., *Time-of-flight mass spectrometry: Introduction to the basics.* Mass Spectrometry Reviews, 2017. **36**(1): p. 86-109.

14. Dupree, E.J., et al., *A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of This Field.* Proteomes, 2020. **8**(3): p. 14.

15. Hu, A., W.S. Noble, and A. Wolf-Yadlin, *Technical advances in proteomics: new developments in data-independent acquisition.* F1000Research, 2016. **5**: p. 419.

16. Pino, L.K., et al., *Improved SILAC Quantification with Data-Independent Acquisition to Investigate Bortezomib-Induced Protein Degradation.* J Proteome Res, 2021. **20**(4): p. 1918-1927.

17. Chen, X., et al., *Quantitative proteomics using SILAC: Principles, applications, and developments.* Proteomics, 2015. **15**(18): p. 3175-92.

18. Geiger, T., et al., *Super-SILAC mix for quantitative proteomics of human tumor tissue.* Nat Methods, 2010. **7**(5): p. 383-5.

19. Bilbao, A., et al., *Processing strategies and software solutions for data-independent acquisition in mass spectrometry.* PROTEOMICS, 2015. **15**(5-6): p. 964-980.

20. Chapman, J.D., D.R. Goodlett, and C.D. Masselon, *Multiplexed and data-independent tandem mass spectrometry for global proteome profiling.* Mass Spectrometry Reviews, 2014. **33**(6): p. 452-470.

21. Geiger, T., J. Cox, and M. Mann, *Proteomics on an Orbitrap Benchtop Mass Spectrometer Using All-ion Fragmentation.* Molecular & Cellular Proteomics, 2010. **9**(10): p. 2252-2261.

22. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nat Biotechnol, 2008. **26**(12): p. 1367-72.

23. Sinitcyn, P., et al., *MaxDIA enables library-based and library-free data-independent acquisition proteomics.* Nat Biotechnol, 2021. **39**(12): p. 1563-1573.

24. Webb-Robertson, B.J., et al., *Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics.* J Proteome Res, 2015. **14**(5): p. 1993-2001.

25. Ong, S.-E. and M. Mann, *Mass spectrometry–based proteomics turns quantitative.* Nature Chemical Biology, 2005. **1**(5): p. 252-262.

26. Bennett, H.M., et al., *Single-cell proteomics enabled by next-generation sequencing or mass spectrometry.* Nature Methods, 2023. **20**(3): p. 363-374.

27. Petelski, A.A., et al., *Multiplexed single-cell proteomics using SCoPE2.* Nature Protocols, 2021. **16**(12): p. 5398-5425.

28. Derks, J., et al., *Increasing the throughput of sensitive proteomics by plexDIA.* Nat Biotechnol, 2023. **41**(1): p. 50-59.

29. Wang, Z., P.K. Liu, and L. Li, *A Tutorial Review of Labeling Methods in Mass Spectrometry-Based Quantitative Proteomics.* ACS Meas Sci Au, 2024. **4**(4): p. 315-337.

30. Pappireddi, N., L. Martin, and M. Wühr, *A Review on Quantitative Multiplexed Proteomics.* ChemBioChem, 2019. **20**(10): p. 1210-1224.

31. Welter, A.S., et al., *Combining Data Independent Acquisition With Spike-In SILAC (DIA-SiS) Improves Proteome Coverage and Quantification.* Mol Cell Proteomics, 2024. **23**(10): p. 100839.

32. Ong, S.-E. and M. Mann, *A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC).* Nature Protocols, 2006. **1**(6): p. 2650-2660.

33. Zhang, G., K. Deinhardt, and T.A. Neubert, *Stable Isotope Labeling by Amino Acids in Cultured Primary Neurons*, in *Methods in Molecular Biology*. 2014, Springer New York. p. 57-64.

34. Tolonen, A.C. and W. Haas, *Quantitative Proteomics Using Reductive Dimethylation for Stable Isotope Labeling.* Journal of Visualized Experiments, 2014(89).

35. Wu, Y., et al., *Five-plex isotope dimethyl labeling for quantitative proteomics.* Chemical Communications, 2014. **50**(14): p. 1708.

36. *<lee-et-al-2010-quantitative-analysis-of-mtraq-labeled-proteome-using-full-ms-scans.pdf>.*

37. Mertins, P., et al., *iTRAQ labeling is superior to mTRAQ for quantitative global proteomics and phosphoproteomics.* Mol Cell Proteomics, 2012. **11**(6): p. M111 014423.

38. Ito, T. and M. Hiramoto, *Use of mTRAQ derivatization reagents on tissues for imaging neurotransmitters by MALDI imaging mass spectrometry: the triple spray method.* Anal Bioanal Chem, 2019. **411**(26): p. 6847-6856.

39. Chen, X., et al., *Quantitative proteomics using SILAC: Principles, applications, and developments.* PROTEOMICS, 2015. **15**(18): p. 3175-3192.

40. Demichev, V., et al., *DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput.* Nature Methods, 2020. **17**(1): p. 41-44.

41. Tian, X., H.P. Permentier, and R. Bischoff, *Chemical isotope labeling for quantitative proteomics.* Mass Spectrometry Reviews, 2023. **42**(2): p. 546-576.

Acknowledgements

First of all, I would like to express my gratitude to Dr. Jürgen Cox for the chance to be a part of this amazing team, his guidance on proteomics and programming, and his overall support to keep this project going. You are an example of what it takes to be a truly professional and distinctive scientist.

I would also like to thank my Thesis Advisory Committee members, Prof. Dr. Johanna Klughammer and Prof. Dr. Maria Robles, for their valuable insights into my project's content and its structure.

I am also very grateful to my previous colleague, Assistant Prof. Dr. Pavel Sinitcyn, who introduced me to the field of proteomics and placed his trust in me when I applied to the Cox group. You are the glue that keeps the whole proteomics community (and a bit more) together.

Many thanks to my colleagues Shamil, Walter, and Juan for our joint efforts in delving into the code and setting up countless benchmarks. Our work together helped me grow professionally, and I hope you have the same experience.

Peli, Carlo, and Jinqui, you guys are simply the best. Not only do you work relentlessly on our lab projects, but somehow you also manage to help everyone around you. It is the rarest skill to always make things better, regardless of the situation.

Helen, I will bother you with Geoguessr even after I defend, be ready.

My deepest gratitude to the Russian LunchTM at the LMU for the therapeutic politics discussions and for making socializing in Munich possible.

I would like to express a very special thank you to my wife, Nadezhda. My first year of PhD was the hardest time of my life, but even in St. Petersburg, you reached out to support me every day, and you are still doing it now.

Last but not least, I would like to thank my mom for all the unconditional support and guidance she gave me to pursue the things in life that I care about. No one believes in me as much as she does, and I will be forever grateful for that. Мама, спасибо!