Computational solutions for quality control of mass spectrometry-based proteomics

Proefschrift voorgelegd tot het behalen van de graad van doctor in de wetenschappen aan de Universiteit Antwerpen te verdedigen door

Wout Bittremieux

Promotoren
prof. dr. Kris Laukens
prof. dr. Bart Goethals

Faculteit Wetenschappen
Antwerpen 2017

Universiteit Antwerpen

Faculteit Wetenschappen

Computational solutions for quality control of mass spectrometry-based proteomics

Proefschrift voorgelegd tot het behalen van de graad van doctor in de wetenschappen aan de Universiteit Antwerpen, te verdedigen door

Wout BITTREMIEUX

Promotor: prof. dr. Kris Laukens
prof. dr. Bart Goethals

Antwerpen, 2017

Computational solutions for quality control of mass spectrometry-based proteomics
Nederlandse titel: *Computationale oplossingen voor de kwaliteitscontrole van massaspectrometrie-gebaseerde proteoomanalyse*

> I can believe things that are true and things that aren't true and I can believe things where nobody knows if they're true or not.
>
> I believe that life is a game, that life is a cruel joke, and that life is what happens when you're alive and that you might as well lie back and enjoy it.
>
> —Neil Gaiman, American Gods

Acknowledgments

As I was trying to figure out whether doing a PhD was really something for me I read a series of sensational stories describing how a PhD can go horribly wrong for a multitude of reasons. I mostly still remember reading "The Ph.D. Grind" by Philip J. Guo, an epic 100-page memoir on his experiences as a PhD student. One part in particular stuck with me. Guo concludes with the F-word: upon asking whether doing a PhD is *fun*, he replied that most of all it was *fulfilling*. And I agree, my PhD was definitely fulfilling: I learned a lot of new interesting things and was able to overcome meaningful challenges. However, most importantly, my PhD was a lot of fun as well, with never a dull moment in these past four years! I am fortunate to have spent these years in the company of truly awesome colleagues, to have visited wonderful places all around the world, and to have met many brilliant and kind people.

First of all I have to thank my two fantastic promotors, Kris and Bart. Kris, I could not have wished for a better supervisor. You gave me all the opportunities I could have hoped for, and then some! You indulged my newbie biological questions, always showed confidence in my abilities to tackle even the hardest problems, and encouraged me to follow my interests in exploring new research questions. Bart, it was nice to have someone around who does not go glassy-eyed upon mentioning the words 'complexity analysis' and who speaks computer science instead of all that biology the whole time. It is all just data in any case.

Next, I owe a major thanks to the amazing members of the BioData Mining research unit. More than just colleagues, all of you have become close friends. I thoroughly enjoyed our many lunch discussions (and yes, I admit that I might have trolled a *tiny* little bit upon *some* occasions), our legendary ReBootCamp trips and late conference evenings, and all our fun social activities. Nghia, I admire your work ethic. Pieter, I admire your willingness to lend a helping hand on other people's projects. Stefan, I admire your extensive knowledge of trivia facts. Bart, I admire your commitment to the bioinformatics community. Aida, I admire how you hold your ground in a male-dominated environment. Charlie, I admire your joy for life. Nicolas, I admire your devotion to all things geeky. Pieter Moris, I admire your attention to online privacy. Chris, I admire how you are able to do without all those delicious braais. Danh, I admire your decision to live abroad in an alien culture. Furthermore, a heartfelt thanks to all the other members of the ADReM research group I enjoyed working with: Anthony, Boris, Cheng, Elyne, Emin, Emmanuel, Floris, Hernan, Jilles, Joeri, Koen S., Koen V., Len, Maarten, Martin, Reuben, Sandy, Stephen, Tayena, Thomas, Tim, and Toon.

My research has been enriched by all the smart people I have had the opportunity to work with in national and international collaborations. Being part of the smoothly running InSPECtor project has significantly contributed to the successful completion of my PhD. As principal investigators of the InSPECtor project I had two additional wise mentors in the form of Dirk and Lennart. Some of our nights at several ASMS conferences are especially memorable. Also thanks to other members of the InSPECtor team: Alex, Ana Silvia, Andrea, Hanny, Jan, Jérôme, Kurt, Nicolas, Pieter, and Sven. Furthermore, a big thanks to Dave. Thank you for hosting me in Cape Town, it was an absolute pleasure to work with you. I relished going on adventures in the beautiful South African countryside, but most of all I massively enjoyed meeting some great people. Thank you to all the hubbers and friends: Anzaan, Lizma, Margaretha, Marisa, Melanie, Ncite, Rob, Serej,

and Yessica. You and all other people at Stellenbosch University have made me feel welcome from the get-go. I also want to quickly thank Jan, Frank, Marc, and Tom from the DBTI research group at Hasselt University. You sparked my interest in doing research.

Ten slotte wil ik mijn hele familie nog bedanken. Jullie doen heel veel voor mij, waarvoor ik soms onvoldoende mijn appreciatie toon. Moeke, bedankt voor de onvoorwaardelijke steun doorheen de jaren. Mart & Bert, bedankt om die maanden met mij te kunnen overleven. Zoë, bedankt om telkens als ik in het weekend thuis kwam weer verheugd te zijn om mij te zien.

Thank you all!

— Wout Bittremieux
Antwerp, February 2017

Abstract

Mass spectrometry is an advanced analytical technique that can be used to identify and quantify the protein content of complex biological samples. Unfortunately mass spectrometry-based proteomics experiments can be subject to a large variability, which forms an obstacle to obtaining accurate and reproducible results. Therefore, to inspire confidence in the generated results a comprehensive and systematic approach to quality control is an essential requirement.

In this dissertation we present several computational solutions for quality control of mass spectrometry-based proteomics. In order to successfully employ comprehensive quality control procedures to assess the validity of the experimental results three basic requirements need to be fulfilled: (i) descriptive quality control metrics that characterize the experimental performance should be defined; (ii) the basic technical infrastructure to unambiguously store and communicate quality control data has to be available; and (iii) advanced analysis techniques are needed to derive actionable insights from the quality control data.

First, we show how secondary metrics that are not related to the spectral data, such as instrument metrics and environment variables, provide a complementary view on the experimental quality. We present the user-friendly Instrument MONitoring DataBase (iMonDB) toolset to manage and visualize these secondary metrics. Second, we introduce the Human Proteome Organization (HUPO) – Proteomics Standards Inititative (PSI) Quality Control working group, whose aim it is to provide a unifying framework for quality control data. We show how the standard qcML file format for mass spectrometry quality control data can be used as the focal point of a strong community-driven ecosystem of quality control tools and methodologies. Third, we present an unsupervised outlier detection workflow to automatically discriminate low-quality mass spectrometry experiments from high-quality mass spectrometry experiments. We show how this workflow can replicate expert knowledge in a data-driven fashion, enabling the substitution of time-consuming manual analyses by automated decision-making. Finally, we show how approximate nearest neighbors indexing can be used to speed up spectral library open modification searching by several orders of magnitude, leading to a record number of spectrum identifications in a minimal processing time.

We conclude with an overview of potential future steps that can be taken to further improve computational quality control methods for mass spectrometry-based proteomics, as well as discussing some of the opportunities to apply advanced machine learning techniques in this field with related challenges.

Samenvatting

Massaspectrometrie is een geavanceerde analytische techniek die gebruikt kan worden om eiwitten in complexe biologische stalen te identificeren en kwantificeren. De resultaten van een massaspectrometrie-gebaseerde proteoomanalyse kunnen echter fel variëren, hetgeen een belemmering vormt voor de nauwkeurigheid en de reproduceerbaarheid van de resultaten. Opdat onderzoekers vertrouwen kunnen hebben in de gegenereerde resultaten is het daarom noodzakelijk om op een grondige en systematische manier aan kwaliteitscontrole te doen.

In dit proefschrift rijken we enkele computationele oplossingen aan voor de kwaliteitscontrole van massaspectrometrie-gebaseerde proteoomanalyses. Hierbij focussen we op drie fundamentele eisen waaraan voldaan moet worden: (i) beschrijvende kwaliteitsmetingen die de experimentele performantie kenmerken moeten gedefinieerd worden; (ii) de technische infrastructuur om op een eenduidige wijze kwaliteitsgerelateerde informatie op te slaan en te communiceren moet beschikbaar zijn; en (iii) geavanceerde analysetechnieken zijn nodig om handelbare inzichten af te leiden uit de kwalitatieve data.

In eerste instantie tonen we aan hoe secundaire metingen die niet van de spectrale data afgeleid zijn, zoals instrumentparameters en omgevingsvariabelen, een aanvullende blik op de experimentele kwaliteit bieden. We presenteren de gebruiksvriendelijke Instrument MONitoring DataBase (iMonDB) software voor het beheren en visualiseren van deze secundaire metingen. Vervolgens introduceren we de Quality Control werkgroep van het Human Proteome Organization (HUPO) – Proteomics Standards Inititative (PSI), dewelke als doel heeft om een universeel kader voor kwaliteitscontrole te ontwikkelen. We bespreken hoe het qcML standaard bestandsformaat voor de opslag van kwaliteitsgerelateerde informatie voor massaspectrometrie experimenten gebruikt kan worden als het centrale element van het ecosysteem van tools en methodieken voor kwaliteitscontrole. Hierna presenteren we een outlier detectie workflow die automatisch laagkwalitatieve experimenten van hoogkwalitatieve experimenten kan onderscheiden en we laten zien hoe deze workflow manuele interventies kan vervangen. Tot slot bespreken we hoe indexeringstechnieken gebruikt kunnen worden om het identificeren van ongekende spectra met behulp van een spectrale bibliotheek te optimaliseren. Dit stelt ons in staat om een recordaantal spectra te identificeren in een minimale verwerkingstijd.

We sluiten af met een overzicht van de potentiële stappen die in de toekomst gezet kunnen worden ter verdere verbetering van computationele methodes voor kwaliteitscontrole van massaspectrometrie-gebaseerde proteoomanalyses. Verder bespreken we enkele uitdagingen om geavanceerde machine learning technieken toe te passen op dit gebied.

Contents

List of publications

Pieter Kelchtermans, **Wout Bittremieux**, Kurt De Grave, Sven Degroeve, et al. "Machine Learning Applications in Proteomics Research: How the Past Can Boost the Future". In: *PROTEOMICS* 14 (4-5 Mar. 2014), pp. 353–366. DOI: 10.1002/pmic.201300289.

Wout Bittremieux, Pieter Kelchtermans, Dirk Valkenborg, Lennart Martens, et al. "jqcML: An Open-Source Java API for Mass Spectrometry Quality Control Data in the qcML Format". In: *Journal of Proteome Research* 13.7 (July 3, 2014), pp. 3484–3487. DOI: 10.1021/pr401274z.

Mathias Walzer, Lucia Espona Pernas, Sara Nasso, **Wout Bittremieux**, et al. "qcML: An Exchange Format for Quality Control Metrics from Mass Spectrometry Experiments". In: *Molecular & Cellular Proteomics* 13.8 (Aug. 1, 2014), pp. 1905–1913. DOI: 10.1074/mcp.M113.035907.

Trung Nghia Vu, **Wout Bittremieux**, Dirk Valkenborg, Bart Goethals, et al. "Efficient Reduction of Candidate Matches in Peptide Spectrum Library Searching Using the Top k Most Intense Peaks". In: *Journal of Proteome Research* 13.9 (Sept. 5, 2014), pp. 4175–4183. DOI: 10.1021/pr401269z.

Stefan Naulaerts, Pieter Meysman, **Wout Bittremieux**, Trung Nghia Vu, et al. "A Primer to Frequent Itemset Mining for Bioinformatics". In: *Briefings in Bioinformatics* 16.2 (Mar. 2015), pp. 216–231. DOI: 10.1093/bib/bbt074.

Wout Bittremieux, Hanny Willems, Pieter Kelchtermans, Lennart Martens, et al. "iMonDB: Mass Spectrometry Quality Control through Instrument Monitoring". In: *Journal of Proteome Research* 14.5 (May 1, 2015), pp. 2360–2366. DOI: 10.1021/acs.jproteome.5b00127.

Pieter Meysman, Yvan Saeys, Ehsan Sabaghian, **Wout Bittremieux**, et al. "Discovery of Significantly Enriched Subgraphs Associated with Selected Vertices in a Single Graph". In: *Proceedings of the 14th International Workshop on Data Mining in Bioinformatics - BIOKDD '15*. Sydney, Australia, Aug. 10, 2015, p. 8.

Wout Bittremieux, Pieter Meysman, Lennart Martens, Dirk Valkenborg, et al. "Unsupervised Quality Assessment of Mass Spectrometry Proteomics Experiments by Multivariate Quality Control Metrics". In: *Journal of Proteome Research* 15.4 (Apr. 1, 2016), pp. 1300–1307. DOI: 10.1021/acs.jproteome.6b00028.

Evelyne Maes, Pieter Kelchtermans, **Wout Bittremieux**, Kurt De Grave, et al. "Designing Biomedical Proteomics Experiments: State-of-the-Art and Future Perspectives". In: *Expert Review of Proteomics* 13.5 (Apr. 25, 2016), pp. 495–511. DOI: 10.1586/14789450.2016.1172967.

Pieter Meysman, Yvan Saeys, Ehsan Sabaghian, **Wout Bittremieux**, et al. "Mining the Enriched Subgraphs for Specific Vertices in a Biological Graph". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (June 7, 2016), p. 1. DOI: 10.1109/TCBB.2016.2576440.

Wout Bittremieux, Dirk Valkenborg, Lennart Martens, and Kris Laukens. "Computational Quality Control Tools for Mass Spectrometry Proteomics". In: *PROTEOMICS* (Early view Oct. 17, 2016). DOI: 10.1002/pmic.201600159.

List of figures

List of tables

List of abbreviations

ABRF	Association of Biomolecular Resource Facilities
AGC	automatic gain control
ANN	approximate nearest neighbors
Annoy	Approximate Nearest Neighbors Oh Yeah
ANOVA	analysis of variance
API	application program interface
ASMS	American Society for Mass Spectrometry
AUC	area under the curve
BSA	bovine serum albumin
ConvNet	convolutional neural network
CPTAC	Clinical Proteomic Tumor Analysis Consortium
cRAP	common Repository of Adventitious Proteins
CV	controlled vocabulary
DDA	data-dependent acquisition
DIA	data-independent acquisition
DL	deep learning
ELN	electronic lab notebook
ESI	electrospray ionization
FDR	false discovery rate
FWHM	full width at half maximum
GAN	generative adversarial network
GIS	geographic information system
GPM	Global Proteome Machine
GPU	graphics processing unit
GUI	graphical user interface
HPP	Human Proteome Project
HTML	Hyper Text Markup Language
HUPO	Human Proteome Organization

IAA	iodoacetic acid
IAM	iodoacetamide
ID	identification
iMonDB	Instrument MONitoring DataBase
iPRG	Proteome Informatics Research Group
IQR	interquartile range
iRT	indexed retention time
iTRAQ	isobaric tags for relative and absolute quantitation
JAXB	Java Architecture for XML Binding
JPA	Java Persistance API
JPQL	Java Persistence Query Language
LC	liquid chromatography
LIMS	laboratory information management system
LoOP	Local Outlier Probability
MALDI	matrix-assisted laser desorption/ionization
MBR	match-between-runs
MIAPE	Minimum Information About a Proteomics Experiment
ML	machine learning
MS	mass spectrometry
MS/MS	tandem mass spectrometry
MSI	Metabolomics Standards Initiative
NCI	National Cancer Institute
NIST	National Institute of Standards and Technology
NN	neural network
NPC	Netherlands Protemic Center
OCR	optical character recognition
OMS	open modification searching
PCA	principal component analysis
PDF	Portable Document Format
PNNL	Pacific Northwest National Laboratory
ppm	parts per million
PRIDE	PRoteomics IDEntifications
PRM	parallel reaction monitoring
PSI	Proteomics Standards Inititative

PSM	peptide-spectrum match
PTM	post-translational modification
PTXQC	Proteomics Quality Control
QA	quality assurance
QC	quality control
RNN	recurrent neural network
ROC	receiver operator characteristic
RT	retention time
SILAC	stable isotope labeling with amino acids in cell culture
SimpatiQCo	SIMPle AuTomatIc Quality COntrol
SProCoP	Statistical Process Control in Proteomics
SQL	Structured Query Language
SRM	selected reaction monitoring
SSM	spectrum-spectrum match
t-SNE	t-Distributed Stochastic Neighbor Embedding
TCGA	The Cancer Genome Atlas
TIC	total ion current
TMT	tandem mass tags
TOF	time-of-flight
wwPDB	Worldwide Protein Data Bank
XML	eXtensible Markup Language

Chapter 1

Introduction

Mass spectrometry (MS) is an advanced analytical technique that can be used to identify and quantify the protein content of complex biological samples [3]. The output of an MS experiment typically consists of a large collection of unknown mass spectra to which corresponding peptide sequences need to be assigned [196]. When mass spectrometry was initially developed as an analytical technique, more than two decades ago, the only way to identify the generated spectra was through a time-consuming manual investigation that crucially depended upon expert knowledge. As technological improvements vastly increased the size of the output of an MS experiment, computational techniques to identify unknown mass spectra were soon developed [75]. Nowadays there are multiple tools that allow to automatically and efficiently identify mass spectra. Despite this computational progress, the results of an MS experiment are often still subject to large variability [251], and only a third of all generated spectra can typically be reliably identified. Therefore, suitable quality control (QC) techniques are of vital importance to inspire confidence in the generated results. This dissertation presents various computational solutions for QC of mass spectrometry-based proteomics experiments, from defining metrics that characterize the quality of the experiments, to automatically discriminating low-quality experiments from high-quality experiments using data mining techniques, to using optimized algorithmic approaches to identify the generated spectral data.

1.1 Outline of the dissertation

This dissertation is structured as follows:

Chapter 2: Quality control in mass spectrometry-based proteomics

We start by introducing the various stages that typically comprise an MS experiment and why it is of crucial importance that the experimental results can be trusted unambiguously. Relevant issues that impact the quality of an experiment are raised, along with suggestions on how to prevent these issues from occurring. We discuss how to account for potential sources of variability that can have an influence on the output results. Specialized QC samples can be used to systematically assess the performance of an MS system. We present the different types of QC samples, their specific properties, and how they can be incorporated into the analytical workflow.

Chapter 3: Computational quality control tools

After relevant QC considerations applicable to the wet-laboratory part of an MS experiment were introduced in chapter 2, we move to the subsequent computational interpretation of the generated data. We review the different software tools that exist to examine the quality of the experimental data. We catalog these tools according to the data source they use to compute their QC metrics and the granularity of experiments they can be applied to. The strengths and weaknesses of each of the tools is reviewed along with recommendations on how they can be used. Finally we discuss how QC metrics can be computed from varying data sources and provide a general comparison of the efficacy of the different types of QC metrics.

Chapter 4: Monitoring secondary quality control metrics

In this chapter we present a specific tool, the Instrument MONitoring DataBase (iMonDB), in full detail. The iMonDB is a unique QC tool in that it does not use the spectral data to provide a quality assessment but instead monitors secondary measurements. First we show that instrument parameters and settings are ideally suited to be used as QC metrics as they provide a low-level view on the instrument performance and can be computed for any type of MS experiment irrespective of its specific analytical procedure. Furthermore, we show how commodity hardware can be used to automatically measure environmental variables, which can have a considerable yet often neglected influence on the experimental results. The relevance of monitoring these secondary metrics is illustrated through realistic use cases.

Chapter 5: Making quality control more accessible

We introduce efforts by the Human Proteome Organization (HUPO) – Proteomics Standards Inititative (PSI) Quality Control working group. The aim of this working group is to provide the basic technical necessities to support a robust community-driven QC ecosystem. To this end its primary goal is to establish the qcML format as a standard file format for QC data, along with accompanying software libraries supporting this format. Notably we provide a detailed technical discussion of the jqcML open source Java application program interface (API) for data in the qcML format and how it aids developers on including support for the qcML format in their tools.

Chapter 6: Unsupervised quality assessment of experiments

In this chapter identification-free QC metrics are used to discriminate between low-quality and high-quality MS experiments. We present an unsupervised method based on multivariate outlier detection techniques to automatically detect low-quality experiments. Furthermore, a specialized outlier interpretation scheme is used to extract explanatory QC metrics for each anomalous observation. Using frequent itemset mining, the most relevant QC metrics characterizing a diminished performance are identified, and based on these metrics the high-quality and low-quality experiments are compared. We show that these techniques produce results that conform to previously published independent expert knowledge despite requiring minimal manual input.

Chapter 7: Optimized open modification spectral library searching

In this chapter we present a novel paradigm for spectral library searching. Although an open modification searching (OMS) strategy can be used to correctly identify unknown mass spectra containing unexpected or unconsidered post-translational modifications (PTMs), leading to a significant increase in identification performance, because of the search space explosion inherent to OMS, this is a computationally very expensive task which consequently takes a long time. We illustrate how approximate nearest neighbors (ANN) indexing techniques can be applied to spectral library searching to efficiently prune the search space during OMS, achieving a speed-up of several orders of magnitude.

Chapter 8: Conclusion

We end with a summarization of the contributions that have been presented in this dissertation and a reflection on their impact. Finally, we mention some open bioinformatics problems for mass spectrometry-based proteomics and highlight some interesting avenues for future work.

Chapter 2

Quality control in mass spectrometry-based proteomics

Abstract

Mass spectrometry is a highly complex analytical technique, and mass spectrometry-based proteomics experiments can be subject to a large variability, which forms an obstacle to obtaining accurate and reproducible results. Therefore a comprehensive and systematic approach to quality control is an essential requirement to inspire confidence in the generated results. A typical mass spectrometry experiment consists of multiple different phases including the sample preparation, liquid chromatography, mass spectrometry, and bioinformatics stages. We review potential sources of variability that can impact the results of a mass spectrometry experiment occurring in all of these steps, and we discuss how to monitor and remedy the negative influences on the experimental results. Furthermore, we detail how specialized quality control samples of varying sample complexity can be incorporated into the experimental workflow and how they can be used to rigorously assess detailed aspects of the instrument performance.

Preface

This chapter is in preparation to be published as:

Wout Bittremieux et al. "Quality Control in Mass-Spectrometry-Based Proteomics". In: *Mass Spectrometry Reviews* (Manuscript submitted)

2.1 Introduction

Proteomics is a crucial domain in modern biological and biomedical research [3]. The current method of choice to identify and quantify complex protein samples is often liquid chromatography (LC) followed by mass spectrometry (MS). The importance of these techniques is exemplified by their use in large-scale research initiatives, such as the two recent attempts at providing a draft of the human proteome [138, 276] or the ongoing Human Proteome Project (HPP) by the Human Proteome Organization (HUPO) [154, 160, 175, 199, 200], where LC-MS techniques are used to identify, quantify, and characterize the human proteome.

As illustrated in figure 2.1, a bottom-up LC-MS experiment consists of multiple different stages. First, various sample preparation measures ensure that the biological samples are optimally suited

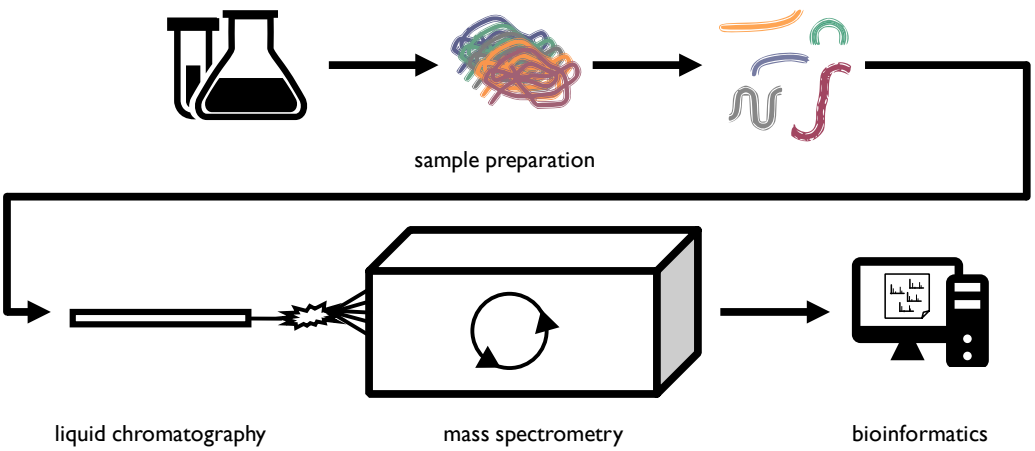


Figure 2.1: Generally considered an LC-MS experiment consists of a sample preparation, a liquid chromatography, a mass spectrometry, and a bioinformatics stage.

for MS analysis. Typical steps include denaturation, reduction, and alkylation of the proteins. Because MS instruments (generally) cannot process intact proteins directly, the denatured proteins are subsequently digested into a peptide mixture through proteolytic cleavage. Next, this peptide mixture is processed through liquid chromatography, which separates the peptides based on their hydrophobicity. After liquid chromatography the peptides get ionized to obtain a charge, and the derived spectra are generated in the mass spectrometer. Whether this is done in a data-dependent acquisition (DDA) or data-independent acquisition (DIA) manner, for a typical discovery experiment in both approaches as many spectra as possible are identified through tandem mass spectrometry (MS/MS), whereas for a targeted experiment specific peptides of interest are exclusively monitored [216]. Finally, the generated spectral data is interpreted through various bioinformatics means [196, 197]. Peptides can be identified from the mass spectra through sequence database searching [76], spectral library searching [105, 233], or *de novo* sequencing [185]; and the peptides can be mapped back to their originating proteins through protein inference [120]. Additionally, protein quantification [13, 102] and other advanced analyses may be performed.

As succinctly described above, performing a mass spectrometry experiment is an intricate process, and each of these different steps has to be optimized to acquire accurate and reproducible results. Unfortunately, despite the many recent technological and computational advances the results of an experiment can still be subject to a large variability [2, 251]. As represented in figure 2.2, this variability can originate from multiple sources [217]: the different stages of an LC-MS experiment can each exhibit stochastic behavior and influence one another, contaminants can inadvertently be present [136], and the optimal computational interpretation is often not obvious [19]. Furthermore, instrument drift and sample degradation can introduce a longitudinal variability [20, 46]. Most notably, instrument interventions, such as a preventive maintenance, have a considerable influence upon the results [20]. Especially in regards to current large-scale studies this is of major importance, as measurements obtained at different times can only be correctly compared with each other if they were obtained under consistent and comparable conditions.

Therefore it is of vital importance that appropriate quality assurance (QA) and quality control (QC) measures are taken to monitor and control the existing variability [8, 172, 177, 249], something

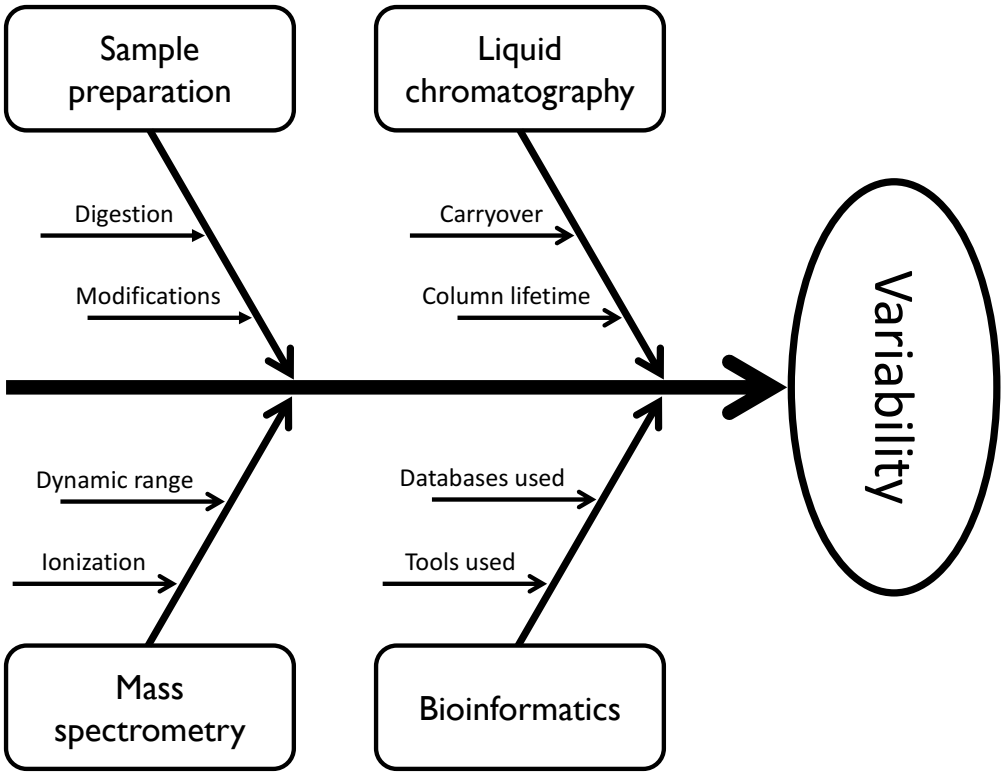


Figure 2.2: An Ishikawa diagram (non-exhaustively) highlighting some of the major sources of variability in each of the stages of an LC-MS experiment. These and other sources of variability will have specific impacts on the obtained results, as will be discussed further.

that is mandatory to inspire confidence in the obtained results. A systematic approach to quality control makes it possible to objectively assess the quality of an MS experiment, and empirical quantitative measures enable the intra-study, intra-laboratory, and inter-laboratory comparison of the performance of mass spectrometry runs [123]. As mentioned previously, these quality assessments are crucial to validate the results produced by long-term multi-site projects [46], such as the HUPO's Human Proteome Project [154, 160, 175, 199, 200] or the studies conducted by the National Cancer Institute (NCI) Clinical Proteomic Tumor Analysis Consortium (CPTAC) [235, 252, 288]. Furthermore, with so many different factors that can impact the experimental results it is important to carefully consider the various influences independently of each other. To this end, for example, a Pareto chart is a helpful visualization technique, as it can be used to represent the contribution of each individual factor to the total variability [23].

Here we will detail the origin of some common causes of variability that can influence the results of a mass spectrometry experiment and which steps should be taken to avoid them. Notably we will highlight how QC samples can be incorporated into the experimental workflow to systematically assess the instrument performance. Mass spectrometry is an advanced and versatile technique, and it can be used for a wide variety of applications. As a result, there is no definitive consensus on which QC methodology to employ [176], nor is it possible to establish a single uniform approach to quality control. Instead we will broadly review some of the representative QC approaches, discuss general considerations, and show how these steps can be used to monitor the various elements of an MS workflow.

2.2 Managing LC-MS variability through quality control

First we will briefly introduce the different types of QC samples that can be employed to monitor the performance of a mass spectrometry experiment and how these samples can be incorporated within the experimental workflow. Next we will highlight some of the problems that can arise during the different stages of an LC-MS experiment, how they negatively influence the experimental results, and how QC methodologies can be used to detect these problems.

2.2.1 Types of QC samples

QC samples can range from a simple peptide mixture to a single protein digest to a complex whole-cell lysate, and each of these types of samples can be employed in a specific fashion to analyze the system performance.

Relatively simple samples consist of a single protein [143], such as bovine serum albumin (BSA), enolase, or cytochrome c; or a protein mixture containing a few proteins [23, 142]; for clarity we will further denote this type of QC samples as 'QC1'. Notably BSA is often used as such a sample because of its historical application in a variety of experimental procedures and its low cost. Furthermore, BSA is usually quite dissimilar from the protein content of the biological samples under consideration, which helps to minimize negative influence on the experimental results due to potential cross-contamination. QC1 samples are typically run on a very frequent basis, i.e. daily or multiple times per day, to quickly evaluate the instrument performance, and they are especially of use to efficiently and systematically assess the LC performance based on observed peak widths and retention times (RTs). As running QC samples takes up valuable instrument time, there is a trade-off between time spent running them and time (and precious sample content) lost due to performing biological runs while the instrument was in a suboptimal state, leading to inferior results. To minimize this trade-off, the QC1 samples are typically run using a short gradient so they can be performed on a frequent basis without unduly occupying an excessive amount of instrument time.

QC samples with a higher sample complexity, denoted 'QC2', consist of a whole-cell lysate, such as a yeast lysate [18, 206], a HeLa cell lysate [143], or a *Pyrococcus furiosus* lysate [278]. QC2 samples are executed using settings equivalent to those for the biological runs to integrally simulate their performance. As this requires more instrument time than the simple QC1 runs, QC2 samples are carried out on a less frequent basis, typically once every week [215]. In contrast to the QC1 single protein digests, complex QC2 samples are used to primarily evaluate the MS performance. Given the sensitivity of current (Orbitrap) mass spectrometers, it is important to inject small amounts of QC samples (e.g. nanogram amounts of peptides; approaching the limit of detection) in order to sufficiently stress the machine and detect potential flaws [143, 215]. Using such low quantities has the additional advantage that it also helps in preventing or reducing cross-contamination of the biological samples by the QC samples. An important consideration to take into account when running QC2 samples is that ideally their characteristics should reflect those of the biological samples. For example, if phosphoproteomics experiments are conducted on a regular basis, it is important to not only perform a general quality assessment, but also to specifically evaluate the ability to detect phosphorylated peptides and proteins [143]. An example of a recent large-scale project where specialized complex QC samples were used is the CPTAC System Suitability (CompRef) Study [71], whose objective it was to validate mass spectrometry protocols by the participants. The CompRef samples were compiled for use within the CPTAC cancer studies and consisted of human-in-mouse xenograft tumor tissue to closely resemble the biological samples. These CompRef samples were first used as a preliminary validation of the workflow during the System Suitability Study, and subsequently acted as QC samples

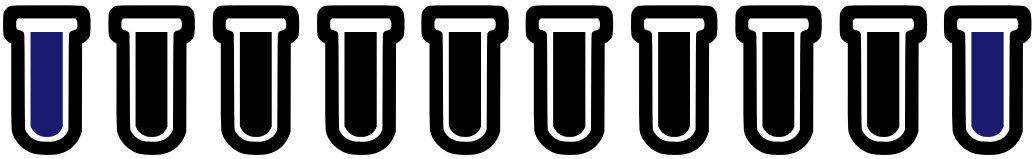
during successive CPTAC studies to characterize human colon and rectal cancer [235, 288] and to evaluate the longitudinal stability of quantitative proteomics techniques [252]. Another example of a complex sample used in a recent high-profile, multi-site study is the hybrid QC sample used by Navarro et al. [194] to benchmark software tools for label-free proteome quantification. This sample consisted of tryptic digests of human, yeast, and *Escherichia coli* proteins mixed in defined proportions to enable the evaluation of both precision and accuracy of label-free quantification, and it was used to assess and ultimately improve the performance of several software tools [194].

Mixtures of synthetic peptides are a slightly different type of QC samples. Depending on the complexity of their composition, these mixtures can be run and evaluated individually, similar to the QC1 samples but with an even further simplified sample content, or they can be spiked into other samples. By spiking a well-defined mixture into the biological samples quality control can be performed in parallel with the biological analyses and a direct link between the qualitative information and the experimental data can be established. Similarly, the synthetic peptides can be spiked into one of the above QC samples, typically a complex QC2 whole cell lysate, to combine the advantages of both types of samples into a single MS run. An important consideration when spiking synthetic peptides into other samples is that these peptides should not overlap with the original sample content. This can be avoided by using artificial, synthetically modified, peptides that are dissimilar from any naturally occurring peptides [78], or by isotopically labeling the synthetic peptides so that their mass is dissimilar from the mass of their naturally occurring peptide variants [24, 208, 209]. These synthetic peptide mixtures are especially important to evaluate the performance of targeted approaches, such as selected reaction monitoring (SRM). To be able to consistently monitor the transitions of specific peptides and to optimally schedule SRM experiments chromatographic stability is an essential prerequisite, which can be evaluated using these well-characterized peptides as their transitions should exhibit minimal run-to-run variation. Synthetic mixtures can be produced in-house or purchased from commercial vendors, and they are often composed in such a fashion that they can be used to examine specific performance characteristics, as we will discuss below.

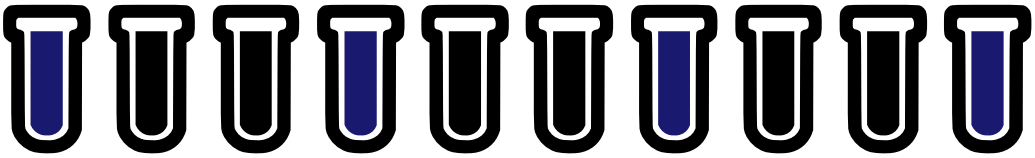
2.2.2 Incorporating QC samples

As illustrated in figure 2.3, QC samples can be combined with the biological samples in several ways [21]. This is tightly linked to the experimental design [169]: how many controls, replicates, etc., are used cannot be considered independently from the use of QC samples.

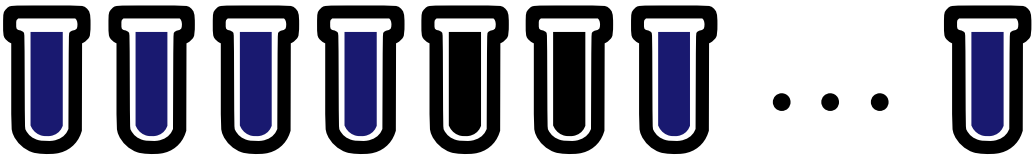
Typically, simple QC1 samples are run at the start and end of each batch of experiments, or at least once a day in case of larger batches, as shown in figure 2.3a. Another approach is to systematically interleave the QC samples after each fixed number of biological samples, as shown in figure 2.3b. This limits the amount of sample loss that can occur due to an intermediate reduction in instrument performance. For example, Zhang et al. [288] report that during their CPTAC study benchmark tumor xenograft samples were run after every five biological samples, and BSA samples were run after every ten biological samples. As frequently running QC samples decreases the throughput of the biological samples, QC1 samples are often run using a short LC gradient to minimize their run time, as mentioned previously. On the other hand, QC2 samples require more instrument time as they closely simulate the biological runs to allow a comprehensive performance evaluation. Therefore, they are typically run less frequently, on the order of once a week. However, as instruments have been getting more powerful, the importance of these complex QC2 samples has increased and they are run on a more frequent basis. Further, as shown in figure 2.3c, a reference set detailing the expected performance might be required to statistically interpret the subsequent QC runs [23, 215]. This reference set can often be derived



(a) QC samples are run at the start and at the end of a batch to assess the batch quality.



(b) QC samples can be interleaved with the biological samples within a single batch to detect an intermediate decrease in performance and avoid undue sample loss.



(c) A reference set of high-quality QC measurements is used as the basis to characterize the performance.

Figure 2.3: The experimental workflow can incorporate the QC samples (blue) through various combinations with the biological samples (black).

from historical high-quality data. In the absence of such measurements it might be necessary to run multiple QC samples successively prior to the start of an experiment. For example, when switching to a new QC standard sample or when employing a novel protocol the new data cannot be compared to the historical measurements and a reference set might need to be compiled explicitly. Likewise, if multiple LC columns are combined interchangeably with the same MS instrument a separate reference set for each of the two columns has to be used, as performance characteristics are column- and instrument-dependent.

2.2.3 Quality control throughout the experimental workflow

As mentioned previously, a typical LC-MS experiment consists of several different stages. Broadly this process can be divided in the four following phases [143]: (i) sample preparation, including proteolytic digestion of the proteins; (ii) separation through liquid chromatography; (iii) mass spectrometry analysis; and (iv) computational data interpretation. All these steps can introduce significant variability that needs to be controlled in order to obtain reproducible results, so the ideal QC methodology should be able to assess the performance of each of these stages.

For each of these phases we will highlight potential sources of variability, and we will detail how structured quality control methodologies can be implemented to detect and control this variability.

Sample preparation

Sample preparation enables the analysis of complex biological samples by mass spectrometry techniques, and entails steps from the initial sample collection up to the proteolytic digestion

and sample storage prior to the actual LC-MS analysis. As the results of an experiment depend on the initial sample quality this step is of vital importance to acquire trustworthy results [35].

Due to the wide variety in sample origin and experimental applications, each with their specific peculiarities and points of attention, it is impractical to cover all existing sample preparation techniques. However, appropriate sample preparation steps for a bottom-up LC-MS experiment typically include denaturation with a chaotrope, reductant, and/or alkylating agent followed by tryptic digestion of the proteins before the resulting peptides can be further processed [66, 210]. All these different steps will introduce a certain degree of variability in the output results, which needs to be monitored and controlled.

Unexpected modifications During denaturation the secondary and tertiary structure of the proteins are removed by interrupting their non-covalent bonds. Additionally, covalent disulfide bridges are cleaved via reduction, after which the proteins are alkylated to prevent the reformation of these disulfide bonds. A complete unfolding of the proteins is required to be able to achieve a full enzymatic cleavage into peptides, but these steps can also introduce unexpected post-translational modifications (PTMs) [133].

The chaotropic agent urea is often used for protein denaturation. An important consideration is that urea can cause artificial carbamylation [241]. In aqueous solutions urea dissociates upon heating and over time. One of its degradation products is isocyanate, which covalently reacts with protein N-termini and ε-amine groups of lysines (and arginines to an extremely limited extent) to form carbamyl derivates [144]. Prolonged incubation of protein samples in urea buffers can induce undesired carbamylation, which will occur at a higher rate if old urea or elevated temperatures are employed. Artificially introduced carbamylation is obviously detrimental for studies that investigate the effect of in vivo carbamylation, which has been related to protein ageing. However, general issues are that carbamylation hampers proteolytic digestion with trypsin, blocks amino groups from isotopic/isobaric labeling, and changes peptide charge states, retention times, and masses [247]. Therefore it is important to avoid the formation of urea-induced carbamylation during sample preparation. This can be done by minimizing the generation of cyanates or by removing active cyanates from the solution. Since urea only degrades in aqueous solution it should be prepared freshly [144]. Other strategies involve maintaining the sample at a low temperature [108, 174], lowering the pH [240], or using a variety of buffers [144, 163, 247]. To verify that unexpected carbamylation is not present in an excessive amount appropriate search settings during peptide identification should be used, i.e. a variable carbamylation PTM should be considered.

Another source of unexpected modifications comes from the alkylation step. Alkylation ensures that after disulfide bridges have been cleaved using a reductant the proteins remain unfolded by preventing reformation of the disulfide linkages. For this step a commonly used alkylation agent is iodoacetamide (IAM). Through reaction with IAM a carbamidomethyl group is attached to cysteine residues to prevent these from reforming disulfide bridges, which results in a monoisotopic mass difference of 57.021464 Da. A potential issue is that overalkylation with IAM will cause N-terminal carbamidomethyl modifications as well [37]. Alternatively, alkylation can be done through carboxymethylation with iodoacetic acid (IAA), which adds a monoisotopic mass of 58.005479 Da. Similar to IAM, overalkylation with IAA will result in N-terminal modifications [279]. Therefore care has to be taken that the sample solution should not be overexposed to either IAM or IAA during alkylation, and appropriate search settings specifying the corresponding N-terminal modifications should be employed to verify this.

A prevalent modification that can easily be misinterpreted is the nonenzymatic deamidation of asparagines and glutamines to aspartates and glutamates respectively, whose rate can increase

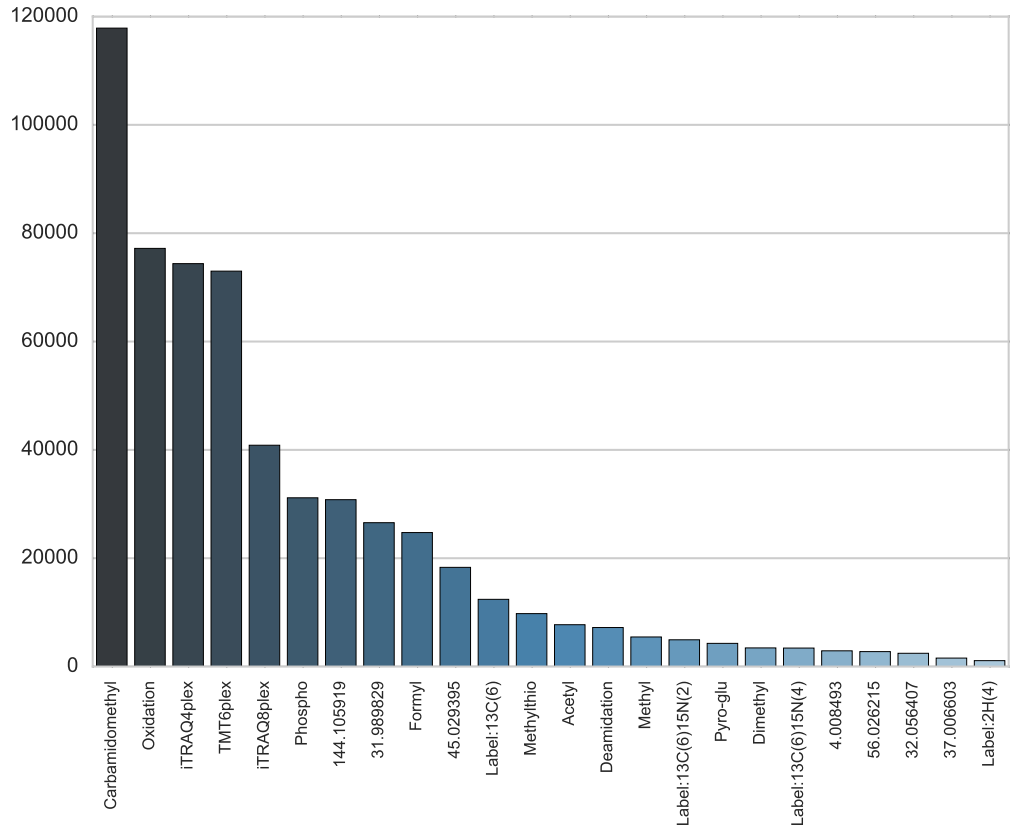


Figure 2.4: The most frequently observed modifications in the PRIDE database [267] based on the 789745 spectral clusters in the human spectral library generated by PRIDE Cluster (version 2015-04) [107]. Note that a single spectrum can potentially contain multiple modifications, and these modifications can have both a biological or an artificial origin.

dramatically during prolonged incubations in digestion buffers at a mildly alkaline pH [110, 148]. As deamidation adds a monoisotopic mass of 0.984016 Da, when not correctly considering this modification the ^{13}C peaks of amidated peptides can be misassigned as monoisotopic peaks of the corresponding deamidated ones, although current high-resolution instruments are able to unambiguously distinguish these peaks. As before it is important to carefully perform the sample preparation and use the correct identification search settings to verify that unexpected modifications have not been introduced.

Suitable search settings are essential to correctly interpret the generated data. Importantly, any expected modifications as well as modifications that can be involuntarily introduced, as discussed previously, should be specified correctly. A recent analysis indicates that unexpected or unconsidered modifications account for missing identifications of a large proportion of unassigned spectra [49]. Similarly, an analysis of 19 million spectral clusters based on previously unidentified spectra deposited in the PRIDE database [267] illustrates the extent to which unidentified spectra can be traced back to unexpected or unconsidered PTMs [107]. Figure 2.4 shows the most prevalent modifications present in the human spectral library generated by PRIDE Cluster [107].

Proteolytic digestion stability In a bottom-up LC-MS proteomics experiment proteins are not analyzed directly, instead they are cleaved into peptides through proteolytic digestion. For this task trypsin is currently the most frequently used protease [262]. Advantages of trypsin are its low cost and its high cleavage specificity and activity. Furthermore, tryptic peptides have various desirable characteristics: their mass is within the preferred mass range for mass spectrometry analysis (based on an *in silico* digestion of all proteins in UniProtKB/Swiss-Prot [255] unique fully tryptic peptides have a median length of 12 amino acids and an interquartile range between 8 and 20 amino acids) and they are ideally suited to carry at least two defined positive charges [243].

At its most basic level, trypsin cleaves exclusively and systematically C-terminal of arginine and lysine, unless followed by a proline [134]. Nevertheless, the formation of semitryptic and nonspecific peptides during protein digestion can still happen due to multiple reasons, although these peptides show a decreased repeatability [251] and they are often not considered during the subsequent bioinformatics analysis, resulting in missing or incorrect identifications [135, 166]. Moreover, most notably for targeted and other quantification experiments consistency of the detectable peptides is of crucial importance.

There are many factors that can influence the digestion stability. One of these is the manner in which the preceding sample preparation steps were performed, and Proc et al. [218] have shown that the choice of chaotropic agents, surfactants, and solvents significantly influences the digestion reproducibility. Other factors that have an influence are the temperature and the pH at which the digestion is carried out, the enzyme-to-substrate ratio, and the duration of the digestion. At a higher temperature the thermal denaturation of trypsin results in a loss of tryptic activity and autolysis [84, 164], while a lower pH improves trypsin stability over an extended digestion period [164]. Meanwhile, although enzyme-to-substrate ratios reported in the literature range from $1:100$ to as high as $1:2.5$, Loziuk et al. [164] have shown that at excessive enzyme-to-substrate ratios an "overdigestion" of peptides caused by increased tryptic autolysis occurs, which may lead to the generation of nonspecific and very small peptides. Similarly, Hildonen, Halvorsen, and Reubsaet [114] recommend a limited digestion time, to avoid a complete digestion as this leads to an increased number of small peptides that are not LC-MS detectable. Furthermore, not all trypsin is created equally, with the origin of the trypsin an important source of variability. Comparisons have shown that the number of missed cleavages, semitryptic peptides, and nontryptic peptides can vary significantly based on whether the trypsin is of bovine or porcine origin [270] and between different commercial trypsins [42, 44, 214].

To assess the digestion performance it is important to monitor the extent of missed cleavages, semitryptic peptides, and nontryptic peptides. Ideally fully tryptic peptides should be preferred as their formation is more reproducible when the trypsin digestion is able to proceed to a state of equilibrium [251]. In some cases semitryptic peptides might be desired as well to generate more detectable peptides and increase the protein sequence coverage [114]. Furthermore it is important to take into account that digestion efficiency is protein- and sample-dependent [81]. Therefore, there is no one-size-fits-all optimal digestion procedure; specialized protocols might be required to for example optimally monitor specific transitions in a targeted experiment.

DIGESTIF is a commercially available compound QC sample that can be used to evaluate the tryptic digestion efficiency [155]. The *DIGESTIF* standard is assembled from a protein scaffold and artificial peptides, with the amino acids flanking the cleavage sites of these peptides selected to either favor or hinder proteolytic cleavage. This allows to progressively monitor the digestion performance by checking which peptides are effectively generated compared to their theoretical cleavage specificity. Alternatively, to monitor the digestion performance Domon et al. [38, 91] inserted QC samples at various different moments during the experimental process. Prior to any sample preparation steps they start with a well-defined QC mixture of a few proteins, insert a first set of isotopically labeled peptides representing a subset of tryptic peptides of these proteins

prior to digestion, and a second set of isotopically labeled peptides (with the same amino acid sequences but a different isotope pattern) prior to the LC-MS analysis. By comparing the relative intensities of the unlabeled peptides, originating from the initial proteins, and the labeled peptides from the first set of labeled peptides the digestion efficiency can be assessed. Further, through comparison with the intensities of the labeled peptides from the second set the overall recovery of the full sample preparation workflow can be evaluated.

Although trypsin is by far the most popular protease, employing another protease or performing a multi-protease protein digestion can have specific advantages [96, 257]. A common alternative to trypsin is the combination of Lys-C and trypsin, which generates similar peptides and significantly reduces the number of missed cleavages [99]. Less frequently used proteases can be beneficial as well, for example to generate longer peptides for "middle-down" proteomics, although these proteases are usually not as thoroughly characterized as trypsin is, so care has to be taken [257]. In these situations a consistent and systematic QC methodology assumes even greater importance.

Sample loss Differential recovery of peptides due to nonspecific adsorption is a potential source of unexpected sample loss during sample preparation, leading to a reduced reproducibility. This sample loss can occur in all steps of a proteomics workflow, and care should be taken that suitable sample handling material is employed at all times. It should be taken into account that adsorption is peptide-specific [101], so optimized protocols might be required for specific situations. Furthermore, a systematic analysis of well-characterized QC samples can highlight signal loss.

The type of sample tubes that are used for peptide storage can result in a large variation in the results, with low-adsorption plastic tubes more suitable than regular plastic tubes or glass tubes [14, 145]. In contrast, hydrophobic peptides exhibit an increase in recoverability for glass tubes [261]. Furthermore, the addition of other compounds to the sample solution can be used to reduce sample loss due to competition of adsorption with the peptides [101, 245].

Adsorption does not happen exclusively to sample tubes; for example some peptides, including all sulfur-containing peptides, adsorb on the stainless steel injection needle as well [261]. As a rule, the more sample handling steps are undertaken, the more loss due to surface adsorption occurs [170]. Therefore, online and automated methods can help to reduce potential sample loss.

Contaminants Another important source of variability is the presence of contaminants in the sample [116, 136, 274]. Contaminants will compete with the spectra of interest during MS measurements and can cause ion suppression of low abundant peptides. Contaminants can often have seemingly innocuous origins, such as a lab member using a new perfume [213] or wearing a wool sweater [136]. It is important to be aware of potential sources of contaminants during all sample preparation steps to avoid undue contamination.

Some contaminants can be traced back to a prior sample preparation step. For example, trypsin autolysis artifacts can be generated during protein digestion, or polymeric interferences can leak from plastics employed in the laboratory. Other contaminants can be involuntarily introduced into the sample. One of the most prevalent contaminants is keratin, which is omnipresent and can originate from skin, hair, dust, etc.

Total elimination of all contaminants is virtually impossible, but suitable procedures can help to minimize contamination. To prevent contaminants as much as possible it is important to always work in a clean lab environment, wear suitable lab clothes, and use specialized equipment for a single task exclusively. To be able to detect contaminants it is necessary to specify them

in the identification search settings. A recent analysis of public data deposited in the PRIDE repository [267] indicates that a majority of commonly incorrectly identified spectra corresponds to contaminants such as albumin, trypsin, and keratin [107]. The MaxQuant software [56] has functionality to automatically include a built-in database of contaminant sequences during sequence database searching [57]. Otherwise, lists of commonly observed contaminants are publicly available. The common Repository of Adventitious Proteins (cRAP) [53] provides a resource of contaminant proteins, sourced from the Global Proteome Machine (GPM) [60]. Both a fasta file for use in sequence database searching and a spectral library in the X! Hunter format [58] are available. In addition, the PRIDE database [267] provides a spectral library of contaminants, generated through PRIDE Cluster [107].

Liquid chromatography

Prior to MS analysis peptides are typically processed using liquid chromatography to separate them based on their hydrophobicity. This adds a time dimension to the subsequently recorded MS data, which enables the mass measurement of individual peptides by spreading out the dense information content of a complex sample over the range of the LC gradient, and which provides orthogonal information for the peptide identification [205].

The LC phase is typically subject to more variable influences than any other component of the LC-MS system [15], and consequently it is the most common culprit of variability in the results of an experiment [227]. A rigorous monitoring of the chromatographic performance is therefore essential. Useful QC metrics include the peak shape (width and height), as sharper peaks generate higher signal intensities and can reduce oversampling [227]. A disproportionate level of signal intensity early or late in the gradient can indicate that the column should be serviced or replaced. An early signal can be caused by sample bleed, and a late signal can arise from peak tailing of either overloaded peptides or peptides with poor chromatographic behavior [227]. The RT of known peptides and their elution order can be used to measure differences between early (hydrophilic) and late (hydrophobic) peptides in the chromatographic gradient [1, 227].

Specialized QC samples can help to thoroughly monitor the performance of the LC system. By composing QC samples so that they contain peptides with varying hydrophobicities the elution profile of the LC gradient can be characterized and evaluated [43], as illustrated in figure 2.5. Notable are so-called indexed retention time (iRT) peptide standards. These peptides have standardized RTs spanning a wide gradient and can be used to normalize the RT of individual experiments [78]. Although RTs can be predicted through computational modeling [189], these predictions have a somewhat limited accuracy [221]. Instead, the reference RTs of the iRT peptides can be used to correct for variations in the RT of the other peptides detected in a single experiment or to align RTs across multiple experiments. Several QC standards containing iRT peptides have been proposed [24, 43, 79, 117, 155]. These standards mostly vary slightly in the range of the LC gradient they can cover, but some standards have further advanced properties. For example, the previously mentioned *DIGESTIF* standard can additionally be used to evaluate the tryptic digestion performance [155], while the *RePLiCal* standard consists of a synthetic protein that exclusively contains lysine-terminating peptides, which ensures that proteolytic digestion by both trypsin and Lys-C can be evaluated analogously [117].

It is of vital importance to avoid cross-contamination due to sample carryover. Carryover happens when an analyte originating from a previously analyzed sample reappears during a subsequent injection, which will result in interference with the active measurements. Carryover can occur because of interactions between the sample and various materials it comes into contact with, as mentioned previously, or when sample residues are trapped in dead volumes within system flow paths [121]. To minimize or avoid carryover suitable column washing steps should be

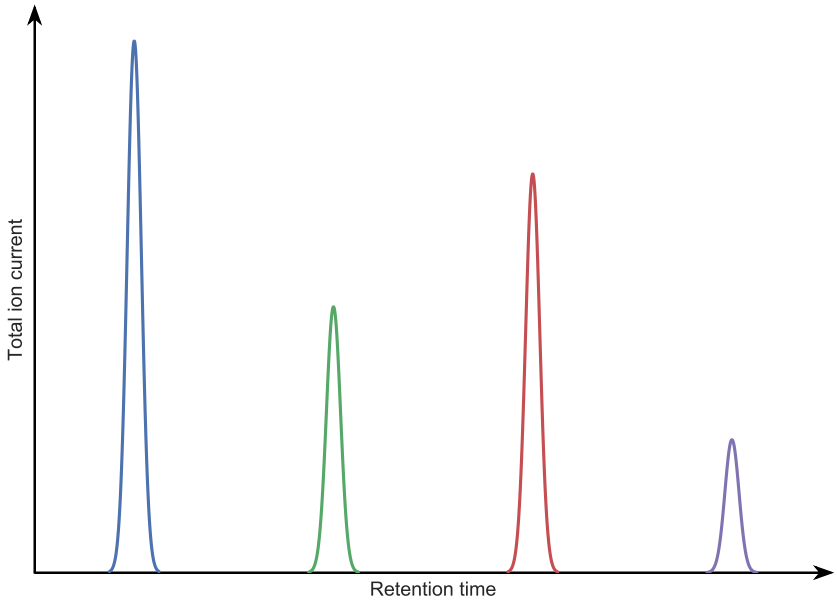


Figure 2.5: Depending on the composition of the QC samples the LC performance can be monitored using peptides that elute over the entire gradient, and the dynamic range can be monitored if peptides are present in varying concentrations.

employed [67, 187, 277]. The presence of carryover and the cleaning effectiveness can be tested by using blank injections between runs of different samples.

Mass spectrometry

As peptides elute from the LC column their mass over charge is measured in the mass spectrometer.

Prior to the mass measurement peptides are ionized through electrospray ionization (ESI). The spray stability can be checked by monitoring for drops in the ion current, which can indicate spray sputter [227]. Tryptic digests are expected to generate mainly peptides containing a 2+ charge, and a high rate of differently charged peptides can indicate ionization issues and will likely impact identification rates [227]. Besides due to an unstable proteolytic digestion, as mentioned previously, partially tryptic peptides can also originate from in-source peptide fragmentation [137]. It is possible to differentiate partially tryptic peptides originating from in-source fragmentation from other proteolytic-derived partially tryptic peptides as the former will have the same LC elution time as their parent peptides [137]. To measure high-quality spectra sufficient signal should be present. Various parameters can influence the internal instrument behavior and these should be carefully optimized [10, 130, 131, 281]. Interrelated instrument parameters influencing the signal-to-noise ratio are the maximum ion injection time and the automatic gain control (AGC), and the effective ion injection time can be monitored to detect problems with sample load [130]. By comparing the measured masses of known compounds,

which can either be explicitly added reference standards or systematically observed contaminants, the mass accuracy can be evaluated [227]. These known masses can further be used as lock mass during mass calibration if excessive mass deviations are observed [201]. The dynamic range can be monitored if peptides are present in varying concentrations, as illustrated in figure 2.5. QC samples can contain distinct peptides in different concentrations [43] or isotopically labeled variants of the same peptide at different ratios [24]. While the concentrations typically span two to four orders of magnitude, the ability to detect even the smallest concentrations indicates the capacity to detect low-abundant peptides over the observed extensive proteome dynamic range [293]. Furthermore, the sensitivity of the MS instrument can be evaluated by employing only small amounts of QC samples, as mentioned previously.

Bioinformatics data interpretation

Although the wet laboratory workflow is often considered to contribute the most variability to the results of an MS experiment and multiple studies have aimed to improve and standardize existing protocols, the bioinformatics data interpretation can likewise introduce major errors that are a cause of irreproducibility [19]. Already for the most fundamental task, mapping peptide sequences to spectra, there exist dozens of different search engines, each using a unique methodology, (possibly silent) assumptions, and peculiarities. Furthermore, even when using the same tool often different versions or parameter combinations can yield significantly dissimilar results. Although a careful evaluation can indicate the optimal search settings for a single tool [82, 263], the high volume of the data generated by MS techniques and the complexity of the bioinformatics tools is a barrier for a mutual objective assessment [93, 284]. The "ground truth" for evaluation is typically not known and the introductions of novel tools regularly lack a sufficient comparison to the state-of-the-art methodology. Nevertheless, to inspire confidence in the acquired results a robust computational and statistical interpretation according to community best practices should always be performed before reporting novel biological findings [232].

In the previous sections we have already mentioned several evaluation criteria that should be investigated to detect specific problems. A benchmark of overall performance that is often monitored is the identification rate in terms of peptide-spectrum matches (PSMs), identified peptides, and identified proteins. This gives a quick insight into the performance of the whole experimental set-up and can indicate whether more detailed quality assessments are required. Whereas for complex QC2 samples, such as a whole-cell lysate, the number of proteins is an often reported metric, for simple QC1 samples, consisting of only a single to a few proteins, the sequence coverage is usually more relevant. The appeal of these high-level QC metrics is that they give a quality assessment of the whole system in a single, easily interpretable metric. However, an MS experiment consists of multiple complex steps that are interrelated, and it might not be possible to identify the source of a decrease in performance based on only a single metric. Instead, sets of detailed QC metrics can be computed [227], highlighting individual performance aspects of the chromatography, the charge state distribution, the spectrum acquisition, etc. A disadvantage of these advanced QC metrics is that, unlike for the number of identifications or the sequence coverage where a higher value is usually better, their interpretation is often not straightforward and requires expert knowledge. Therefore, to establish value intervals of acceptable performance a high-quality reference set might be used, as described in section 2.2.3 [23, 215]. Furthermore, analyzing multiple metrics simultaneously requires a multivariate approach. Although this increases the complexity of the data analysis, recent research has shown some promising approaches for informed and automatic decision-making based on multivariate sets of advanced QC metrics [9, 30, 273]. Finally, it does not suffice to investigate QC metrics for a single experiment in isolation. Instead, the longitudinal performance should be examined. Through extensive monitoring of operation over time the technological passport of a mass spectrometer

can be established, and based on these highly detailed and instrument-specific insights the reliability of the experimental results can be diagnosed. Although not necessarily related to their biological relevancy, this constitutes the bare essentials required to inspire solid confidence in novel scientific findings.

2.3 Conclusion

Performing an LC-MS experiment is a highly complex activity and there exist a multitude of potential sources of variability that can influence the results and impact repeatability and reproducibility. We have tried to give an overview of some prevalent issues that can arise and how to detect them, but nevertheless we have only managed to cover the tip of the iceberg. Instituting a thorough QC methodology might initially seem like it requires a lot of effort and it occupies valuable instrument time without any immediate gains, but a systematic quality assessment pays off in the long run and is an indispensable prerequisite to inspire confidence in the acquired results. Especially in order to advance mass spectrometry techniques and use them as routine applications in a clinical setting a consistent analytical performance is a fundamental requirement [177, 237, 280].

Developments on both the experimental and computational front are needed to improve current QC methodologies, for which core facilities can act as an important driver [162]. As proteomics technologies have matured core facilities have concentrated the cutting-edge technical expertise necessary to obtain high-quality results, and they form an essential means of providing this in an affordable manner [184]. Core facilities have an incentive to support and develop robust quality assurance practices to demonstrate the quality of the generated data to their clients and stakeholders, and through their expert knowledge on a broad aspect of MS-based applications they are at the forefront of developing standardized QC workflows. Significant bioinformatics work is needed as well. All too often laboratories still only monitor detailed QC metrics in an empirical fashion when a malfunctioning is suspected, instead of on a systematic basis. This can be partly attributed to the relative absence of user-friendly tools and software suites that facilitate and encourage a methodical QC workflow. Although a few tools to compute advanced QC metrics exist [21, 32], they remain underused in part due to their limited ease-of-use. Nevertheless, to make further progress objective metrics rooted in a solid bioinformatics foundation are mandatory. The end goal should not be to merely understand QC issues retrospectively, but also to prevent them from happening by timely suggesting solutions. Eventually the QC tools should ideally be tightly coupled to the MS instrumentation to make automated decisions on the fly, avoiding subjective and time-consuming manual quality assessments to increase the throughput.

Finally, because of these obvious advantages we expect that the importance of quality control will only increase in the (near) future. Currently QC information is often not included in scientific publications, which might lead to uncertainty on the conducted methodology. Instead, in the future reporting this information might become formalized, similar to existing guidelines mandated by journals [253], and the QC metrics might become an integral part of a data submission to public data repositories [73, 178], with current work ongoing to provide the necessary technical basis for this goal [33]. Coupling comprehensive QC information to the experimental data will enable assessing the reliability of an experiment at a glance based on the instrument's technological passport. Especially in light of some historical occasions where claims turned out to be exaggerated [11, 80] and recent reports of the general reproducibility crisis in various scientific fields [12], an innate approach to quality control is mandatory to inspire confidence in and to advance the field of mass spectrometry-based proteomics.

Chapter 3

Computational quality control tools

Abstract

As a growing emphasis has been placed on quality control for mass spectrometry-based proteomics multiple computational quality control tools have been introduced. These tools generate a set of metrics that can be used to assess the quality of a mass spectrometry experiment. Here we review which different types of quality control metrics can be generated, and how they can be used to monitor both intra- and inter-experiment performance. We discuss the principal computational tools for quality control and list their main characteristics and applicability. As most of these tools have specific use cases it is not straightforward to compare their performance. For this survey we used different sets of quality control metrics derived from information at various stages in a mass spectrometry process and evaluated their effectiveness at capturing qualitative information about an experiment using a supervised learning approach. Furthermore, we discuss currently available algorithmic solutions that enable the usage of these quality control metrics for decision-making.

Preface

This chapter was previously published as:

Wout Bittremieux et al. "Computational Quality Control Tools for Mass Spectrometry Proteomics". In: *PROTEOMICS* (Early view Oct. 17, 2016). DOI: 10.1002/pmic.201600159

This chapter gives an overview of the available computational tools that can be used for quality control in mass spectrometry-based proteomics. Although this chapter already includes a brief introduction to the Instrument MONitoring DataBase (iMonDB) tool, it should be noted that chronologically the iMonDB was published prior to this chapter, and it will be presented in further detail in chapter 4.

3.1 Introduction

In the past decade mass spectrometry-based proteomics has evolved into an extremely powerful analytical technique to identify and quantify proteins in complex biological samples. This high-throughput approach can yield a considerable volume of complex data for each experiment. As it

has matured, over the last few years a growing emphasis has been placed on quality assurance (QA). This attention on quality assurance is of the utmost importance to safeguard confidence in the acquired results: in cases where this has been lacking mass spectrometry (MS) proteomics has sometimes suffered from exaggerated claims [11, 80]. To anticipate this evolution, a shift to "quality by design" is now taking place [249]. This means that the "designing and developing formulations and manufacturing processes ensure a predefined product quality." As such, quality assurance consists of multiple aspects of which QA is an essential component, but other elements such as a careful experimental design [45, 119, 169] are equally vital.

Whereas the experimental design has to be established prior to the initiation of an experiment, quality control takes place while or after the experimental results are obtained. Nonetheless, quality control and experimental design should not be discussed in isolation, as they are interwoven. For example, a quality control (QC) sample can consist of a single peptide, a single protein digest, or a complex lysate, and this decision influences the type of QC metric(s) that can be investigated [21, 143, 206]. Furthermore, one has to decide how many QC runs to include in the experiment and to what extent and in which order these QC runs are interleaved with the biological samples under consideration. The goal of quality control is then to leverage the experimental set-up to comprehend how well an instrument performs and how confident the results from the experiments are.

Related to the experimental design and based on the type of performance we want to monitor there are multiple approaches to quality control. A typical example consists of the use of QC samples with a simple sample content interleaved between the biological samples. The interesting aspect of such QC samples is that they have a controlled, limited, and known sample content. They are typically measured on a frequent basis, which allows to extract periodic information on the performance of the mass spectrometer. Of course, to understand this performance expressive QC metrics that provide information indicative of the quality of the experimental results need to be derived. Some straightforward and commonly used QC metrics include the number of identifications or the sequence coverage. Although these metrics give a global view of the performance, they do not allow us to pinpoint specific elements of the workflow where a failure might have arisen. Instead, more granular QC metrics providing information on the chromatography, the ion signal, the spectrum acquisition, etc., might be used.

Over the years dozens of QC metrics have been proposed, generated by a range of bioinformatics tools. In this chapter we will list the main QC tools and explain their use cases and capabilities. Furthermore, we will provide an empirical assessment of which type of QC metrics is most adequate in detecting low-quality experiments.

3.1.1 Quality control metrics

We can primarily distinguish QC metrics based on whether they represent information about a single experiment, or about multiple experiments, as illustrated in figure 3.1.

Intra-experiment metrics give information about a single experiment and are computed at the level of individual scans or identifications. These metrics show the evolution of a specific measure over the experiment run time, such as, for example a chromatogram of the total ion current (TIC) over the retention time (RT), or the mass accuracy of the identified spectra.

Inter-experiment metrics, on the other hand, assess a specific part of the quality of an experiment using a single measurement for the whole experiment. These values can subsequently be compared for multiple experiments, for example through a longitudinal analysis to evaluate the performance over time. Often an intra-experiment metric can be converted to an inter-experiment metric through summarization. This is illustrated in figure 3.1, where a TIC

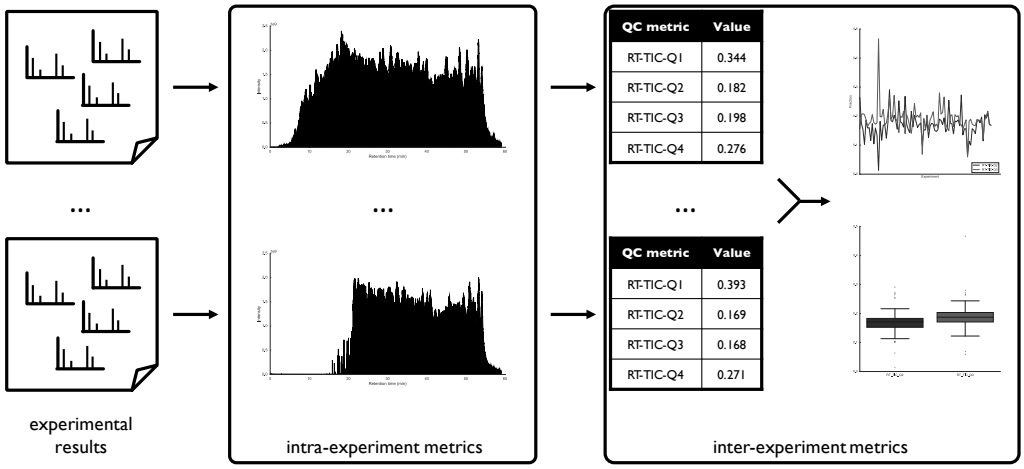


Figure 3.1: Intra-experiment metrics evaluate the quality of a single experiment, whereas inter-experiment metrics can be used to compare the quality of multiple experiments.

chromatogram enables the assessment of the chromatographic performance by visualizing the intensity distribution over the RT. Using summary statistics this continuous information can be converted to inter-experiment metrics detailing the fraction of the total RT that was required to accumulate a certain amount of the TIC, which gives a high-level assessment at the experiment level of the chromatographic stability.

To compare inter-experiment metrics multiple observations for different experiments are required. Therefore, QC tools that analyze these metrics usually include a database back-end for the persistent storage of historical data. On the other hand, intra-experiment metrics can be computed from only a single experiment and there is no comparison with external data. As a result, QC tools that exclusively generate intra-experiment metrics are generally easier to set up, as no external data storage needs to be provided. Because the use cases and requirements differ between these two types of tools, we will further make a distinction between tools that generate metrics for individual experiments, tools that compare a limited group of experiments and do not necessarily require a complex back-end for data storage, and tools for longitudinal tracking that store QC data for a large number of experiments.

A second distinction between various metrics can be made based on from which stage in a mass spectrometry workflow they represent the quality of the system. As shown in figure 3.2, we can distinguish between instrument metrics, identification (ID)-free metrics, and ID-based metrics.

ID-free metrics and ID-based metrics are similar in the sense that they are both computed from the spectral results. *ID-free* metrics are derived solely from the spectral results, i.e. from the raw spectral data directly generated by the mass spectrometer. These metrics aim to capture information over the whole mass spectrometry workflow and include for example the shape of the peaks or the course of TIC detailing the chromatography, the number of MS and tandem mass spectrometry (MS/MS) scans or the scan rate detailing the spectrum acquisition, or the charge state distribution detailing the ionization. The advantage of ID-free metrics is that they are generated directly from the raw spectral data, which makes it possible to instantly generate these metrics as soon as a mass spectrometry run has been completed.

ID-based metrics are derived from the spectral results as well, but they combine these data with

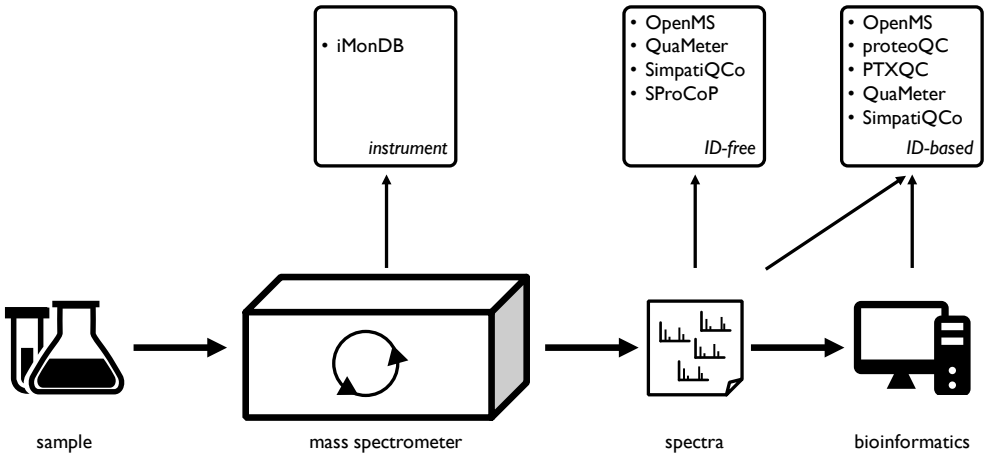


Figure 3.2: QC tools can capture qualitative information at different stages of a mass spectrometry experiment. For each type of QC metrics the representative tools are listed.

subsequently obtained identification results. Examples include aforementioned metrics such as the number of identifications in terms of peptide-spectrum matches (PSMs), peptides, or proteins; or the sequence coverage for a known sample. Other detailed metrics can be computed as well, for example by comparing the difference in RT for similar identifications to assess the chromatographic stability, the number of spectra identified as the same peptide to measure the dynamic sampling, or by linking information similar to the ID-free metrics with the identification results. Compared to ID-free metrics, the computation of ID-based metrics is somewhat more involved because it additionally requires the identifications results. Furthermore, the computation of ID-based metrics can be negatively influenced by suboptimal identification settings. However, in general the inclusion of identifications can provide a more detailed qualitative assessment of the experimental results.

Finally, *instrument* metrics do not look at the spectral data but derive information directly from instrument readouts. These are typically very sensitive, low-level metrics, such as the status of the ion source, the vacuum, or a turbo pump, depending on the type of instrument. An advantage of instrument metrics is that they directly indicate which part of the instrument is outside its normal range of operation. This facilitates troubleshooting and can be a driver for maintenance scheduling. On the other hand, these metrics cannot be directly related to the experimental results, instead they provide a secondary source of QC information. Furthermore, instrument metrics are instrument- and vendor-specific, and are typically not included in open file formats such as the mzML format [179].

Each distinct type of metric can give a different view on the quality of the data. However, not all metrics are always applicable; often metrics are especially relevant for a particular type of sample. For example, monitoring the sequence coverage is mostly applicable when using samples that contain a single protein digest, whereas the number of protein identifications is applicable to samples that consist of a complex lysate. Additionally, the type of experiment also plays an important role. For example, the number of identifications is very relevant for a discovery experiment, but less so for a targeted experiment. In contrast, instrument metrics are largely agnostic to the type of experiment and the sample content, but they can significantly vary between different instrument models and vendors.

3.2 Quality control tools

In recent years, quality control has become a key focus of attention in academic, industrial, and governmental proteomics laboratories. This trend is exemplified (and possibly driven) by the numerous QC tools that have been developed over the past few years. Initial work by Rudnick et al. [227] described for the first time how computational QC metrics can be used to objectively assess the quality of a mass spectrometry proteomics experiment. Whereas previously quality control was mostly performed manually by monitoring a few key measurements, this work showed how a comprehensive set of QC metrics can be used to thoroughly investigate the system performance. A set of 46 mainly ID-based metrics was defined and implemented in a pipeline of Perl programs by researchers at the National Institute of Standards and Technology (NIST), called NIST MSQC. This set of metrics has since then been reimplemented in several lab-specific data processing pipelines. Support for NIST MSQC itself has been discontinued in early 2016, but several of the reimplementations remain under active development.

It has been demonstrated that computational QC metrics provide objective criteria that can accurately capture the quality of a mass spectrometry experiment, and there has been a proliferation of tools that can compute such metrics. Here, we will detail the primary tools, their characteristics, and their usage. Table 3.1 provides an overview of the discussed tools.

3.2.1 Tools evaluating individual experiments

QuaMeter

QuaMeter was initially developed as a user-friendly and open-source alternative to NIST MSQC. NIST MSQC consisted of a graphical user interface (GUI) wrapper around multiple individual tools and scripts with various interdependencies, which resulted in a complex pipeline. Additionally, some elements of this pipeline could only be modified to a limited extent. NIST MSQC could exclusively compute metrics from Thermo Scientific raw files, and only supported three search engines to provide identifications: the NIST MSPepSearch or the SpectraST [152] spectral library search engines, or the OMSSA [94] sequence database search engine. These limitations restricted the applicability of NIST MSQC. Instead, QuaMeter consists of a single multi-platform command-line application that is able to compute QC metrics from raw files originating from instruments produced by multiple vendors. Using the ProteoWizard [48] library it is able to read spectral data stored in a wide variety of vendor-specific raw files (restricted to the Windows platform) and open standard file formats, such as mzML [179]. Further, it can utilize identification results produced by any search engine in the standard mzIdentML [126] or pepXML format through external processing using IDPicker [166].

The initial QuaMeter version [167] computed a set of 42 ID-based QC metrics equivalent to those defined by Rudnick et al. [227]. In a subsequent version QuaMeter improved upon this by also including functionality to compute a set of 45 ID-free QC metrics [273]. Both sets of metrics are inter-experiment summary metrics, although the output is exported to simple tab-delimited text files, so the visualization and analysis thereof has to be done using external software or code scripts. Without advanced visualization or analysis functionality QuaMeter focuses solely on computing QC metrics. Especially the set of ID-free metrics, which requires only the spectral data, can very easily be computed. For the set of ID-based metrics some prior processing of the identification results by IDPicker is required, which can make this process slightly more cumbersome. Only a limited configuration is required, and through the command-line functionality the computation can easily be automated. This makes QuaMeter a powerful tool that computes an extensive set of inter-experiment QC metrics.

Tool	Interface	Operating system	Experiment type	Instrument	ID-free	ID-based	Website
QuaMeter [167, 273]	command-line	Windows, Linux	discovery DDA	×	✓	✓	http://proteowizard.sourceforge.net/
OpenMS [272]	KNIME	cross-platform	discovery DDA	×	✓	✓	http://www.openms.de/
proteoQC [275]	R	cross-platform	discovery DDA	×	×	✓	http://bioconductor.org/packages/proteoQC
PTXQC [28]	R	Windows, cross-platform	discovery & quantification DDA	×	×	✓	https://github.com/cbielow/PTXQC
SProCoP [23]	Skyline	Windows	discovery, targeted SRM & PRM	×	✓	×	http://proteome.gs.washington.edu/software/skyline/tools/sprocop.html
SimpatiQCo [215]	web	Windows	discovery DDA	×	✓	✓	http://ms.imp.ac.at/?goto=simpatiqco
iMonDB [34]	GUI	Windows	any	✓	×	×	https://bitbucket.org/proteinspector/imondb/

Table 3.1: An overview of the discussed QC tools and their main characteristics.

OpenMS

OpenMS is a comprehensive open-source software library that offers a wide range of algorithms and tools for mass spectrometry-based proteomics and metabolomics [246]. It consists of various small processing tools that can be used to construct complex analysis workflows [7, 127]. These workflows can be designed visually using the KNIME workflow engine [26], where each tool functions as an individual node in the workflow.

The various OpenMS nodes can be used to build complex QC pipelines [272]. The provided QC nodes can compute a set of intra-experiment metrics, consisting of both ID-free and ID-based metrics. OpenMS supports a range of search engines to generate identifications for the ID-based metrics, for which there exist specific nodes, including Mascot [212], MS-GF+ [140], Myrimatch [250], OMSSA [94], and X!Tandem [59]. Example QC metrics include the number of spectra (identified or otherwise), peptides, and proteins; mass accuracy statistics; and the mass over charge and RT acquisition ranges. These metrics are complemented by various plots that provide further details, such as a TIC chromatogram, a histogram of the mass accuracy of the identified peptides, or a histogram of the charge distribution of the detected ion features. OpenMS exports this information to an eXtensible Markup Language (XML)-based qcML file [272], which can be visualized in a web browser through an embedded stylesheet, or to a Portable Document Format (PDF) report.

Due to the wealth of algorithms and tools that are available in the OpenMS software library, the provided QC workflows can potentially be easily extended to compute additional metrics. Furthermore, there is no need to be restricted to algorithms natively provided by OpenMS, as the available functionality can easily be extended through custom nodes, for example by using the built-in support for the R statistical programming language [220]. This makes it possible to build granular workflows and achieve a very fine-grained control, although expert knowledge of the OpenMS ecosystem and the KNIME environment is recommended to do so. The constructed workflows can subsequently be exported and shared. Both OpenMS and KNIME are cross-platform tools, ensuring the universal applicability of these workflows.

3.2.2 Tools comparing groups of experiments

proteoQC

The proteoQC package [275] for the R programming language [92, 220] can be used to generate a Hyper Text Markup Language (HTML) report detailing the experimental quality. Prior to executing proteoQC the experimental design has to be specified by configuring each spectral data file representing a sample as belonging to a specific fraction, technical replicate, and biological replicate. The generated QC report contains intra-experiment metrics for each individual sample, as well as aggregated information to compare samples at the level of their fractions, technical replicates, and biological replicates.

To generate a set of intra-experiment ID-based metrics for each sample, proteoQC uses the rTANDEM package [86] to interface the X!Tandem [59] sequence database search engine in R to provide identification results. For each sample some individual metrics and QC plots are generated, such as a breakdown of the precursor ion charge states, the mass accuracy, information on the number of spectra and peptides that were used to identify distinct proteins during protein inference, etc. Furthermore, when identifying the data proteoQC automatically adds the common Repository of Adventitious Proteins (cRAP) [53] database to the user-provided protein database. The cRAP database contains contaminants such as common laboratory proteins, like trypsin,

or contaminants transfered through dust or contact, like keratin, and proteoQC reports which of these contaminants were detected in the samples. Additionally, proteoQC reports on the reproducibility of the results by comparing the number of identified spectra, peptides, and proteins per fraction, technical replicate, and biological replicate, and their overlap between the replicates.

By incorporating the experimental design proteoQC can make informed comparisons between individual samples, which provides QC information on an additional level. Furthermore, proteoQC is fully cross-platform within the popular R programming language. However, as the QC pipeline has to be configured programmatically, some R experience is recommended to utilize proteoQC.

PTXQC

Proteomics Quality Control (PTXQC) [28] is an R-based QC pipeline for MaxQuant [56], a highly popular software suite for quantitative proteomics. Like MaxQuant, PTXQC supports a wide range of quantitative proteomics workflows, including stable isotope labeling with amino acids in cell culture (SILAC), tandem mass tags (TMT), and label-free quantification. After initial processing of the spectral data by MaxQuant, PTXQC uses the MaxQuant output results to compute various QC metrics. PTXQC requires as input the custom text files generated by MaxQuant and the MaxQuant configuration settings, and hence cannot be used to process any other type of data. As PTXQC is written in the R programming language, it is fully cross-platform. Additionally, easy drag-and-drop functionality to execute the QC analyses is provided for the Windows operating system.

PTXQC produces an extensive report that contains a set of 24 intra- and inter-experiment metrics. These metrics are divided into four categories corresponding to the specific MaxQuant output source the metrics are derived from: “ProteinGroups”, “Evidence”, “Msms”, and “MsmsScans”. The metrics cover a wide range of information, including the intensity of the detected features and peptides, the potential presence of contaminants, the mass accuracy of the identified peptides and fragments, the number of missed cleavages detailing the enzyme specificity, and the number of identified peptides and proteins. Other metrics are specifically related to the MaxQuant match-between-runs (MBR) [55] functionality. MBR aligns the RTs of multiple runs and transfers their identifications across features that have the same accurate mass and a similar RT, providing more data for the downstream quantification of proteins. PTXQC assesses the MBR performance by evaluating the RT alignment and by checking whether the identification transfer seems correct. All of these metrics are then visualized and compared between the different raw files that constitute the considered MaxQuant project using detailed figures. Furthermore, each of the metrics is converted to an individual score for each experiment using automated scoring functions. Most of these scores are absolute scores generated by comparing the observation to a threshold, for example such as whether the number of detected contaminants is too excessive, or generated by evaluating a specific characteristic of the observation, for example such as the extent to which the mass deviations are centered around zero. Other scores are computed for a single raw file using the other raw files as a reference, for example by comparing the number of missed cleavages in each individual raw file to the average number of missed cleavages. Finally, some other scores are evaluated relative to settings extracted from MaxQuant, such as the mass accuracy compared to the width of the precursor mass window. All these scoring functions generate inter-experiment metrics that are used to compare the quality of the different experiments. Usefully, PTXQC provides a heatmap overview of the inter-experiment metrics, which yields an assessment of the quality at a glance and facilitates pinpointing the low-performing experiments.

Although PTXQC can exclusively be used to analyze MaxQuant results, through this tight integration it is able to compute some highly relevant and specialized QC metrics. These metrics

do not only assess the quality of the spectral data, but also provide information on the subsequent bioinformatics processing by MaxQuant. Furthermore, the addition of a high-level heatmap at the start of the report is very useful to get a quick overview of the quality, after which the more detailed visualizations can be employed to further investigate potential problems.

SProCoP

Statistical Process Control in Proteomics (SProCoP) [23] is a QC script written in R [220] that can be used as a plugin [41] for the popular Skyline [168] tool for targeted proteomics. SProCoP applies well-established statistical process control techniques such as the Shewhart control chart and the Pareto chart. The purpose of a Shewhart control chart is to track performance over time and identify outliers that deviate excessively from the expected behavior. Further, the Pareto chart is a combination of a bar and line graph, which displays the number of deviating measurements for each metric along with its cumulative percentage, and provides feedback on which metrics are more variable and may require attention.

Using these statistical process control techniques SProCoP monitors the performance of five inter-experiment QC metrics based on targeted peptides present in QC samples with a known sample content or spiked into real samples: signal intensity, mass measurement accuracy, RT reproducibility, peak full width at half maximum (FWHM), and peak symmetry. Measurement thresholds are defined empirically based on a reference set of samples with a known good quality, after which the performance of other samples in the Skyline project can be investigated.

Through its integration with Skyline SProCoP is vendor-independent and can be used for a wide range of targeted and discovery workflows. Additionally these statistical process control techniques are available online [242] and have been implemented in the Panorama [234] repository for targeted proteomics from Skyline. Panorama AutoQC is a utility application that monitors for new data files and automatically invokes Skyline to process the data [22]. The QC metrics are stored in Panorama and the statistical process control charts similar to SProCoP can be visualized through the Panorama web application.

3.2.3 Tools for longitudinal tracking

SimpatiQCo

SIMPle AuTomatIc Quality COntrol (SimpatiQCo) [215] not only computes various QC metrics, it also stores and visualizes these metrics for longitudinal monitoring of the system performance. It uses a PostgreSQL database as back-end, and an Apache webserver to provide a web-based front-end for configuration and visualization.

SimpatiQCo can compute QC metrics from a limited selection of Thermo Scientific and SCIEX instruments. Raw files from these instruments can be uploaded to the web server manually, or can be added automatically through a “hot folder” that is monitored continuously for new raw files. These raw files are then submitted to a linked Mascot server for peptide identifications. Next, SimpatiQCo calculates a range of ID-free and ID-based QC metrics such as the number of MS and MS/MS scans, the number of identified PSMs and proteins, the TIC, and information on lock masses (if applicable). Further, specific peptides and proteins can be investigated in detail using metrics such as the peak area and width and the elution time of peptides of interest, and the protein sequence coverage. For each QC metric the range of acceptable values is learned based on the historical observations using robust statistical measures to take outlying values into account. This information is then displayed in the metric plots using a color-coded background band to

highlight deviating system performance. Further, external messages can be entered manually, for example pertaining to instrument maintenance. These messages will be superimposed on the metric plots to relate the external events to the evolution of the metrics.

SimpatiQCo consists of a number of different components, such as the database, the web server, and various processing tools. These components need to be installed individually, and although a step-by-step installation guide is available online, this complicated process is not recommended for novice users. Furthermore, not all of the configuration can be done through the graphical web-based client. For example, to process raw files these must be able to be linked to a specific instrument. Unfortunately, an instrument definition can only be created by manually adding a record in the corresponding table of the PostgreSQL database.

SimpatiQCo is a powerful tool to track system performance over time, albeit with some technical limitations. Namely, SimpatiQCo is only able to process raw files generated on a limited number of instrument models and only supports the commercial Mascot search engine for peptide identifications.

iMonDB

Unlike the previous tools the Instrument MONitoring DataBase (iMonDB) [34] does not compute metrics from the spectral results, but extracts instrument metrics from the raw files. The iMonDB uses a MySQL database to store its information. This database acts as a server, with two separate standalone GUI applications that can connect to the database as clients, each with a specific task: the iMonDB Collector processes raw files and stores the instrument metrics in the database, whereas the iMonDB Viewer retrieves the information from the database and visualizes it.

The iMonDB supports a wide range of instruments manufactured by Thermo Scientific, although it does not support other instrument vendors. Prior to extracting instrument metrics from a raw file, a corresponding instrument definition has to be created. This can be done through the iMonDB Collector, which allows the full configuration through its graphical user interface. Further, extraction of the instrument metrics can be done manually through the GUI, or can be done through command-line functionality provided by the iMonDB Collector. This command-line functionality can be used to automatically run the iMonDB Collector using an external scheduling tool, such as the native operating system scheduler.

The behavior over time of the metrics for each instrument can be viewed using the iMonDB Viewer. Similar to functionality provided by SimpatiQCo it is possible to add additional information pertaining to external events and show this on the metric plots to link this to the evolution of the metrics. It is also possible to export a PDF file of the external events for reporting purposes.

A unique aspect of the iMonDB is that this is the only tool that is able to systematically analyze instrument metrics. The advantage of these instrument metrics, which provide information at the lowest level, is their high sensitivity, which makes it possible to detect emerging defects in a timely fashion. However, because these metrics are instrument-dependent they are usually not retained during conversion to open formats, such as mzML [179]. Due to this limitation the iMonDB needs to work with vendor-specific raw files directly, which is currently limited to Thermo Scientific raw files. Furthermore, there is a multitude of instrument metrics that are extracted, which makes it hard to comprehend which metrics are most useful to monitor systematically, even for expert users. Nevertheless, these instrument metrics can be very useful to detect malfunctioning instrument elements before these have a deleterious effect on the experimental results, preventing potential loss of valuable sample content.

3.2.4 Other tools

As mentioned previously, NIST MSQC [227] was the first tool that generated computational QC metrics, although it was recently retired in early 2016.

Metriculator [254] is a web-based tool for storing and visualizing QC metrics longitudinally. However, Metriculator does not compute QC metrics directly but critically depends upon an embedded version of NIST MSQC. Unfortunately, the installation process for Metriculator is not very straightforward; it has many Ruby dependencies whose installation might fail, and which are presently outdated or even no longer supported.

LogViewer [248] is a simple visualization tool that presents a set of 11 instrument metrics, such as MS and MS/MS ions injection times, and ID-free metrics, such as the charge state and mass distributions. As input it uses log files from Thermo instruments exported by RawXtract [182], which has been deprecated presently.

A different approach is used by SprayQc [231]. Whereas the other discussed tools compute QC metrics post-acquisition, SprayQc directly interfaces with peripheral equipment to continuously monitor its performance. SprayQc is able to automatically track the stability of the electrospray through computer vision, the status of the liquid chromatography (LC) pumps, the temperature of the column oven, and the continuity of the data acquisition. In case a malfunctioning is detected SprayQc can automatically take corrective actions and warn the instrument operator. This is a valuable approach to minimize the loss of precious sample content and provide early notifications, and it can complement the other QC tools that provide a post-acquisition quality assessment.

3.3 Metrics evaluation

We compared various sets of metrics to assess their effectiveness in expressing the quality of a mass spectrometry proteomics experiment. Typically this is not a straightforward task because, as we have reviewed in the previous sections, each QC tool has its own characteristics and requirements, and use cases can vary as some tools are specific to certain experimental workflows and sample types. Meanwhile most tools also represent some of their QC information through visualizations. Although these quickly provide useful insights for human users, this data is not suitable for an objective, automatic comparison.

To compare different types of metrics we used the set of instrument metrics computed by the iMonDB [34], the set of ID-free metrics computed by QuaMeter [273], and the set of ID-based metrics as identified by Rudnick et al. [227]. These sets of metrics are very comprehensive and all of these inter-experiment metrics can readily be used to compare experiments to each other. To be able to determine whether or not these metrics can capture qualitative information about an experiment, we used a public dataset for which the quality of the experiments is known. The dataset consists of a number of complex quality control LC-MS runs performed on several different instruments at the Pacific Northwest National Laboratory (PNNL) [9]. Each sample had an identical content (whole cell lysate of *Shewanella oneidensis*), and the quality of the various runs has been manually annotated by expert instrument operators as being either “good”, “ok”, or “poor”. We split up the various runs depending on the instrument type, being either “Exactive”, “LTQ IonTrap”, “LTQ Orbitrap”, or “Velos Orbitrap”, with each of these instrument groups consisting of multiple individual instruments. We refer to the original publication by Amidan et al. [9] for further information on the experimental procedures and the dataset details.

This public dataset already contains the precomputed set of ID-free metrics by QuaMeter and the set of ID-based metrics by SMAQC [236] (the PNNL in-house reimplementation of the

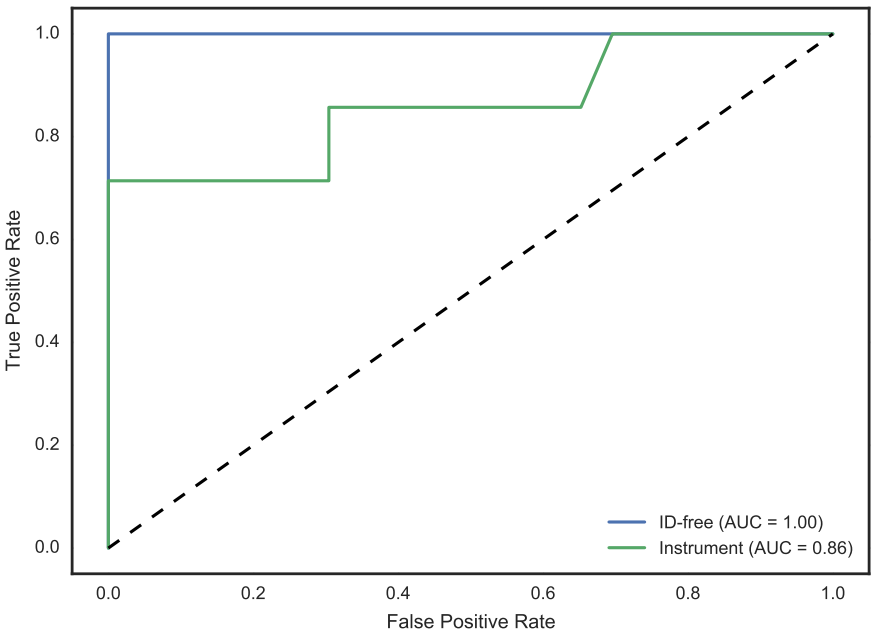
NIST MSQC metrics defined by Rudnick et al. [227]). We further used the iMonDB to compute the set of instrument metrics. To this end all experimental raw files, precomputed QC metrics, and the expert quality annotations were retrieved from the PRoteomics IDEntifications (PRIDE) database [267].

To quantify the expressiveness of these three sets of metrics, each capturing a different type of QC information, we employed a binary classifier. As the quality of the experiments was manually assessed by expert instrument operators, this labeling can be used as the ground truth to train the classifier. We used the acceptable experiments, with their quality designated as either "good" or "ok", as the positive class, and the inferior experiments, with their quality designated as "poor", as the negative class. When given an experiment represented by its QC metrics, the classification task consists of correctly predicting the experiment's quality. Prior to training the classifier we removed redundant features that have a very low variance and we rescaled the features robust to outliers by centering by the median and scaling by the interquartile range. Next, for each separate instrument type we trained a random forest classifier, for which we split the data into 65%-35% training and testing subsets that are equally stratified according to their quality labels. This classifier has been coded in Python and uses the random forest implementation from scikit-learn [207], along with functionality provided by NumPy [260] and pandas [183]. The code is available as open source at `https://bitbucket.org/proteinspector/qc-evaluation`.

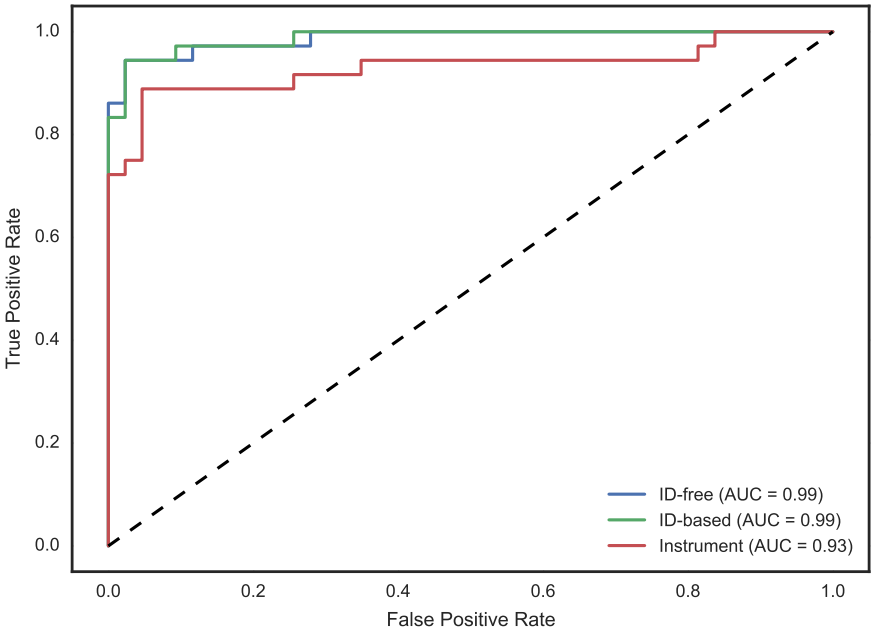
As illustrated by the receiver operator characteristic (ROC) curve in figure 3.3 all three types of QC metrics are adept at discriminating high-quality experiments from low-quality experiments. This shows that all of the different tools can give us valuable insights into the quality of an experiment, and that information captured at various different stages of the mass spectrometry process should be investigated. ID-based metrics slightly outperform ID-free metrics, most likely because the ID-based metrics can employ additional information provided by the identifications. This difference is minimal however, which is perhaps not surprising as both types of metrics take similar properties of the spectra into account. This reinforces previous research which showed that ID-based metrics are not significantly influenced by slight differences in the identifications, such as when using an alternative search engine [167]. This also shows the excellent efficacy of ID-free metrics in objectively evaluating the quality based solely on spectral information. Because ID-based metrics require additional computational steps to obtain the identifications, whereas ID-free metrics can be directly computed from the spectral results, ID-free metrics might be preferred if a speedy quality assessment is required. In contrast, instrument metrics perform a little worse at correctly identifying low-quality experiments. This is likely because they are only secondary results that are not always directly related to the data quality. Nevertheless, these metrics still have merit as they do not depend on a specific type of experiment or sample content, but are applicable on all occasions. Furthermore, by combining the individual classifiers for the various types of metrics in an ensemble classifier a further performance gain can be achieved because the different types of metrics each provide a complementary view on the quality.

3.4 Using QC metrics for decision-making

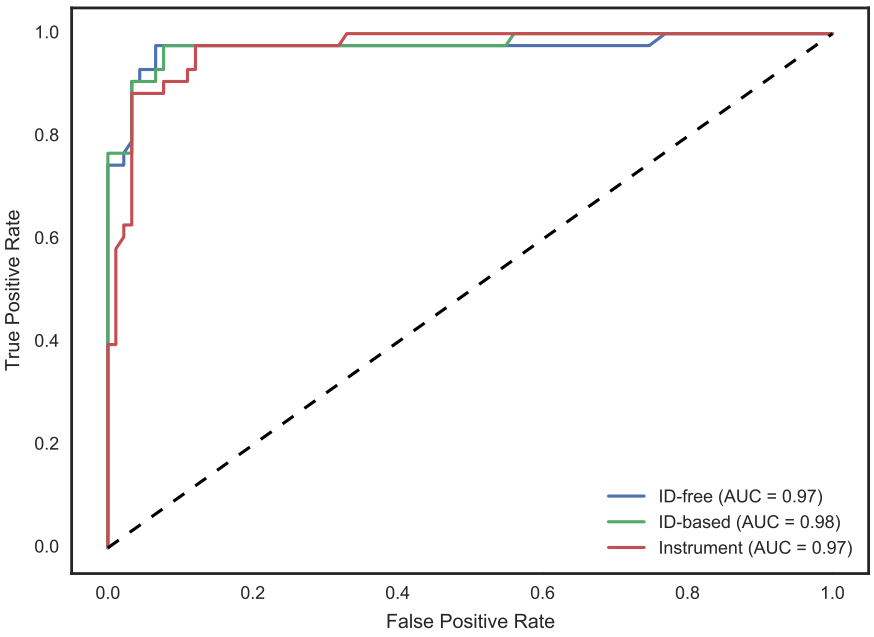
As tools for computational quality control have proliferated in recent years, the challenge in this field is now shifting from the computation of QC metrics toward informed decision-making based on these metrics. However, interpreting these metrics is not trivial. First, considerable domain knowledge is required to understand what each metric signifies. Second, the metrics form a high-dimensional data space, which complicates their analysis. Different elements in a mass spectrometry workflow do not function in isolation but instead influence each other, which has to be taken into account while analyzing metrics representing information about these elements.



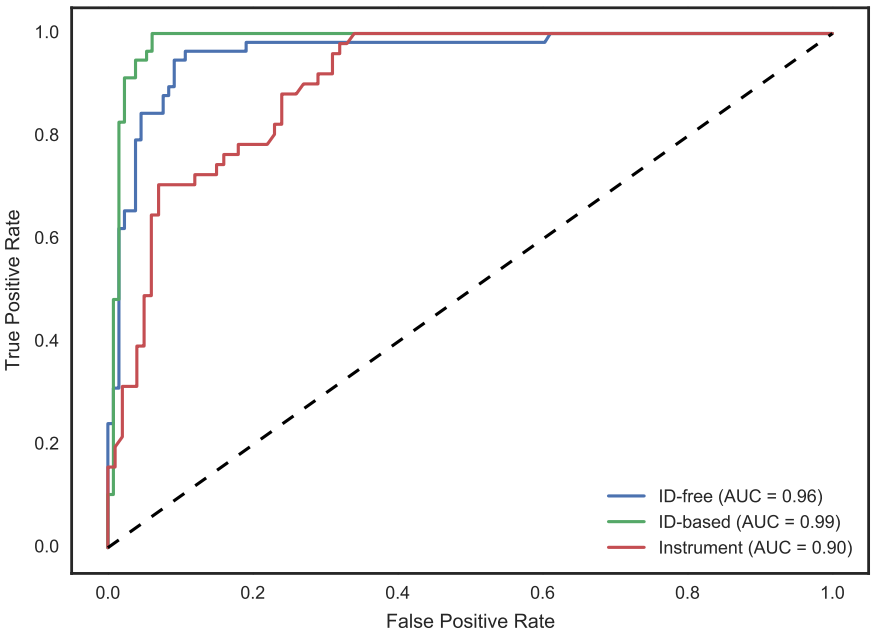
(a) Exactive



(b) LTQ-IonTrap



(c) LTQ-Orbitrap



(d) Velos-Orbitrap

Figure 3.3: ROC curves showing the classification performance of various types of QC metrics.

Therefore, univariate approaches are generally insufficient; instead multivariate approaches that can deal with the high-dimensional data space should be preferred, while also taking the curse of dimensionality into account [4].

To this end Wang et al. [273] have developed a robust multivariate statistical toolkit to interpret QC metrics. They have used a principal component analysis (PCA) transformation to reduce the data to a low-dimensional approximation, in which they were able to successfully detect outlier low-quality experiments based on pairwise dissimilarities. Furthermore, they developed an analysis of variance (ANOVA) model which enabled them to identify whether the observed variability was attributable to lab-dependent factors, batch effects, or biological variability. Such work driving the understanding of QC metrics is highly valuable, and these analyses have been applied to great effect for multiple studies. For example, it was used to assess the quality of the experimental results for various studies conducted by the National Cancer Institute (NCI) Clinical Proteomic Tumor Analysis Consortium (CPTAC) [235, 252, 288].

Similar work was done by Bittremieux et al. [30], who applied unsupervised outlier detection to identify low-quality experiments. Subsequently they used a specialized outlier interpretation technique to determine which QC metrics mostly contributed to the decrease in quality. The advantage of this approach is that all QC metrics are used to identify low-quality experiments, unlike when using a dimensionality reduction, such as PCA, which discards some of the information. Meanwhile, the advanced outlier interpretation pinpointing the most relevant QC metrics can yield actionable information for domain experts to optimize their experimental set-up.

Whereas these previous analyses used unsupervised techniques, Amidan et al. [9] trained a supervised classifier to discriminate low-quality experiments from high-quality experiments. A supervised approach will generally perform better than an unsupervised approach but will require initial training. Furthermore, a supervised classifier might have to be retrained to adapt it to data generated by a different instrument or in a different laboratory. Amidan et al. [9] have expended significant effort in manually annotating the quality of over a thousand experiments to generate training data, which allowed them to build a highly performant logistic regression classifier.

These analyses are extremely valuable, as they allow us to achieve a deeper understanding of the mass spectrometry processes and the properties of what makes a high-quality experiment. These algorithmic approaches provide a thorough quality assessment of the spectral data, which enforces informed decision-making, and which has the potential to automatically drive the spectral acquisition in the future.

3.5 Conclusion

We have given an overview of the available computational tools to generate QC metrics for mass spectrometry-based proteomics. These tools enable assessing the performance of the experimental set-up and detecting unreliable results. These are essential requirements to inspire confidence in the experimental results, which will prove to be a crucial step in the maturation of proteomics technologies, and which will allow us to for example routinely apply these technologies into a clinical setting [177, 249]. Another potential application where an accurate assessment of the data quality is paramount, is in the reuse of public data [73, 85, 141, 178]. As public data repositories keep expanding and the potential for data reuse grows, we envision that data submissions to public repositories will soon have to be accompanied by QC parameters at the time of submission, or will have a standard set of QC metrics calculated automatically after submission [178].

Finally, most current QC tools are limited to the typical use case of bottom-up data-dependent acquisition (DDA) discovery experiments, and their QC metrics often cannot be directly translated to other types of experiments. Less research has been done on QC for other types of workflows, such as data-independent acquisition (DIA) [72] or top-down proteomics [256], or even related mass spectrometry-based domains, such as metabolomics [68]. In the next few years we will likely see the efforts on QC expanded to these types of workflows as well, which will further bolster the diverse and powerful mass spectrometry ecosystem.

Chapter 4

Monitoring secondary quality control metrics

Abstract

Quality control (QC) metrics are typically derived from the mass spectral data, with secondary data often overlooked as a source of qualitative information. Nevertheless, instrument parameters, which can be extracted from the raw data files, give a detailed account on the operation of the mass instrument, and can be related to observations in the mass spectral data. The advantage of instrument information at the lowest level is its high sensitivity to detect emerging defects in a timely fashion. Furthermore, environmental variables have a profound impact upon the quality of the experimental results as well. The ambient temperature is an especially important, yet often overlooked, factor. Here we introduce the Instrument MONitoring DataBase (iMonDB) to longitudinally track secondary QC metrics. We present software tools to automatically extract, store, and manage instrument parameters from raw-data objects into the highly efficient iMonDB database structure, which enables us to monitor instrument parameters over a considerable time period. Furthermore, we show how commodity hardware can be used to systematically and affordably monitor the ambient laboratory temperature. The proposed tools foster an additional handle on quality control and are released as open source under the permissive Apache 2.0 license. The tools can be downloaded from `https://bitbucket.org/proteinspector/imondb`.

Preface

Part of this chapter was previously published as:

Wout Bittremieux et al. "iMonDB: Mass Spectrometry Quality Control through Instrument Monitoring". In: *Journal of Proteome Research* 14.5 (May 1, 2015), pp. 2360–2366. DOI: 10.1021/acs.jproteome.5b00127

The initial version of the iMonDB software was demoed at the American Society for Mass Spectrometry (ASMS) annual conference 2015 in St. Louis, MO, USA, during the "Methods and tools for intra- and inter-experiment LC MS performance tracking" workshop, and was featured again during the same workshop at the ASMS annual conference 2016 in San Antonio, TX, USA.

4.1 Introduction

As described in chapter 3, most quality control (QC) tools compute their metrics from the experimental spectral-derived data, optionally including the corresponding identification results [32]. However, besides these metrics based on the spectral data, there is a separate part of qualitative information available in the form of secondary QC metrics, consisting of mass spectrometer instrument parameters and environmental variables.

Possible instrument parameters are for example the status of the ion source, the vacuum, or a turbo pump, depending on the type of instrument. This information provides complementary insights into the operational characteristics of a mass spectrometer compared to the aforementioned QC metrics that originate from the mass spectral data. Metrics based on the experimental data can capture a wide range of possible problems during a mass spectrometry (MS) experiment; however, if these metrics are seen to deviate from normal or previously observed values, this behavior still has to be related to the malfunctioning of a particular element in the experimental workflow or defect in the MS setup. On the other hand, instrument parameters directly indicate which part of the instrument is outside its normal range of operation, and this information can subsequently be related to the interpretation of the experimental data. As such, metrics based on the spectral data and the instrument parameters offer complementary information: both are able to indicate shortcomings of the mass instrumentation, yet they operate on different layers of information. An important advantage of monitoring instrument parameters is that emerging problems can be spotted in time and remedial measures can be suggested, for example, to replace the inert gas in the collision cell. When looking at statistics based on the mass spectral data alone, an emerging problem could remain undetected while the system deteriorates further, since most instruments are cleverly constructed to compensate for the malfunctioning of a component in the MS process, for example, by increasing the ion target for fragmentation.

Furthermore, although instrument parameters provide important insights into the operation of an MS instrument, both these metrics and QC metrics derived from the spectral data still only deal with qualitative information that is internal to the operation of the instrument. Scheltema and Mann [231] have addressed this in part with their SprayQc software, which can automatically monitor the status of peripheral equipment, such as the electrospray conditions and the liquid chromatography (LC) performance. However, besides influences from the peripheral equipment, external processes that are completely unrelated to the MS instrumentation, such as environmental variables, can have a profound impact on the experimental results as well. For example, Keller et al. [136] report how a high number of sheep keratin contaminants were observed after a lab member started wearing a different sweater in the sample preparation laboratory when the weather changed. In a similar fashion, even the deodorant worn by lab members can be a source of contamination [213]. Although consistently observed contaminants can potentially be used as lock masses to recalibrate mass measurements [159], careful operational procedures are advisable to avoid excessive contamination [31], as the contaminants will interfere with the biological signal and crowd out the measurements of interest. An important, yet often overlooked, environmental factor is the ambient temperature. Most MS instruments are able to compensate for limited changes in ambient temperature, however, excessive temperature fluctuations can negatively impact measurements due to various reasons. First, certain biological compounds, such as enzymes or proteins, may not be stable at room temperature or higher temperatures, which can introduce unexpected modifications due to unanticipated chemical reactions [31]. Furthermore, importantly, LC column performance is temperature-dependent [50, 51]. The higher the temperature, the faster the exchange of the analytes between the mobile phase and the stationary phase occurs, leading to decreased peak retention times (RTs) at higher temperatures. Finally, temperature has an impact on mass accuracy as well. For example, mass measurements

on various types of instrument models, such as time-of-flight (TOF), Orbitrap, and Q-Exactive instruments, can be influenced by temperature shifts [90, 198, 201, 224].

To date, no extensive longitudinal monitoring of the mass spectrometer normal range of operation or environmental conditions is undertaken due to a lack of access to sensory data. Instrument parameters have largely been ignored for the analysis of mass spectral data and quality assessment because they are recorded in a vendor-specific raw data format, from which they are not retained after conversion to an open file format, such as the mzML format [179]. Although the instrument parameters can be read from a raw file using vendor software, this functionality is often cumbersome and rarely advertised prominently. Additionally, environmental conditions are typically measured in a separate system, if at all, and this information is rarely systematically related to the functional characteristics of the MS instruments. As a result, these important sources of qualitative information remain largely unexploited.

Here we present user-friendly software to systematically monitor secondary QC metrics such as instrument parameters and environmental conditions. First, we provide an automatic tool to extract the instrument sensor information from raw data files, along with a highly efficient database structure to store this information. Currently this functionality is limited to the extraction of instrument parameters from Thermo Scientific raw files, but a considerable part of the technical workflow is kept generalized, with the expansion toward other vendors in mind. Second, although complex and expensive systems exist to monitor the laboratory environment, temperature, a crucial variable impacting the experimental quality, can be straightforwardly monitored using commodity hardware. We show how basic networked smart sensors can be employed to systematically monitor the laboratory temperature, and how this information can provide crucial insights into the performance of an MS instrument. Using specialized processing and visualization tools, historical QC metrics can be monitored for a large number of experiments over time in order to compile the technological passport of an instrument. This unique information can be used to rapidly and reliably confirm and detect instrument failure and assess the quality of the experimental data, fostering novel methods of quality control.

4.2 Monitoring secondary QC metrics

The monitoring functionality centers around the Instrument MONitoring DataBase (iMonDB), a database optimized to store a vast amount of QC information. Figure 4.1 gives an overview of the different steps required to set up the iMonDB and how the database interfaces with external components. This central data storage enables the integration of several software tools that interact with the database, such as the iMonDB Collector, a Java application to set up and populate the database in an automated fashion, and the iMonDB Viewer, a user-friendly graphical user interface (GUI) tool to visualize longitudinal QC information. Each step in this workflow will be discussed in more detail next.

4.2.1 Instrument monitoring database

The central database, called iMonDB, is a relational database used to store historical mass spectrometer instrument parameters. Care has been taken to provide a general database structure that can accommodate a wide range of instrument types and different parameter settings. In addition, the database is optimized to perform the most common queries in a computationally efficient manner. The entity-relationship diagram of the iMonDB is shown in figure 4.2 and contains four types of information pertaining to the monitoring and interpreting of instrument parameters.

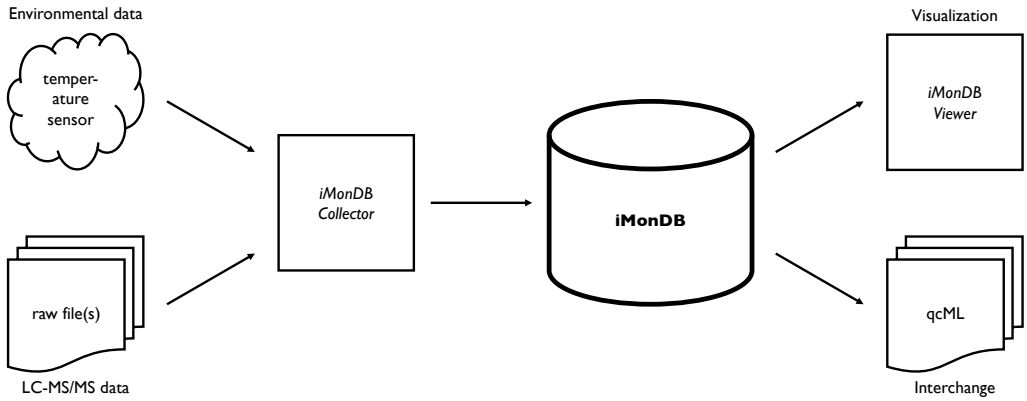


Figure 4.1: Overview of the QC monitoring functionality. Instrument parameters are extracted from experimental raw files and environmental information from a temperature sensor. All data is stored in the database, from which it is served for visualization or export to an interchange format.

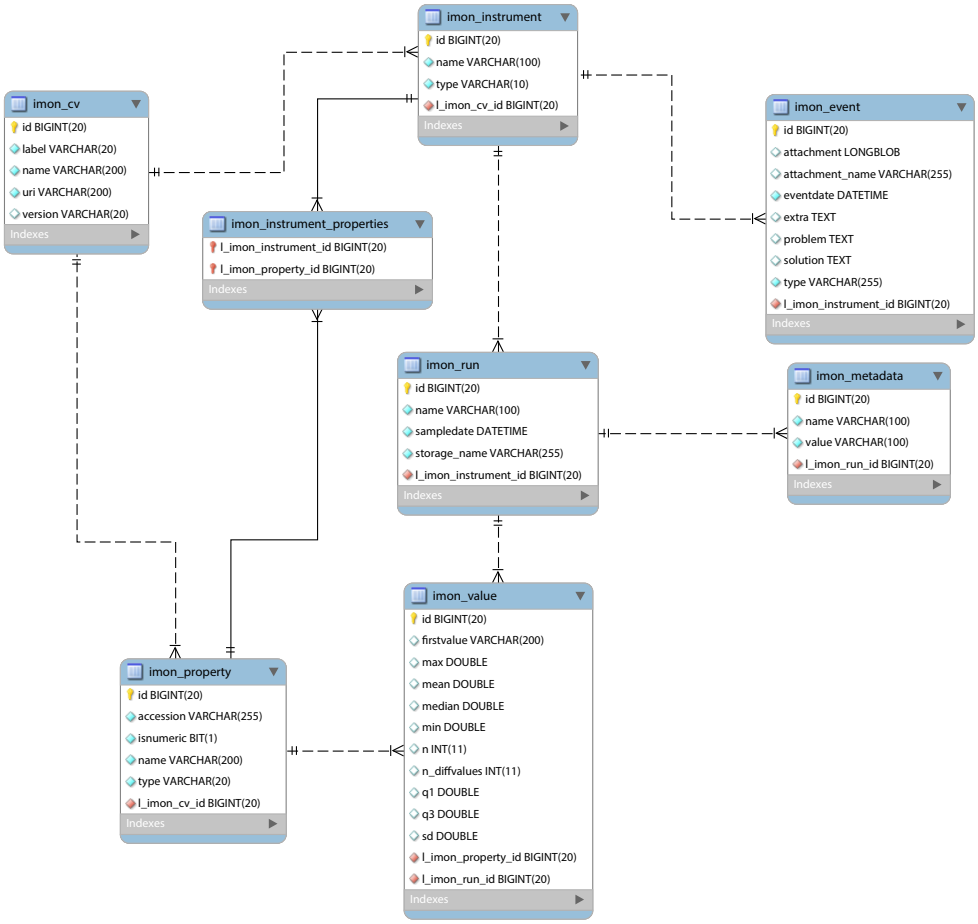


Figure 4.2: Entity-relationship model for the iMonDB version 1.0.0.

First, the main node of information is an individual instrument of a certain model. By making use of a controlled vocabulary (CV), such as the Proteomics Standards Inititative (PSI)-MS controlled vocabulary [181], an instrument can be unambiguously cataloged, while at the same time the database structure supports all instrument models that have been formally defined.

Next, associated with a particular instrument are all the runs (or a specific subset thereof) that were performed on it. A run corresponds with a sample that was analyzed on the mass spectrometer, and is represented by a single output file containing the raw measurement values. Various metadata can be associated with a run, detailing specific characteristics of the way a run was carried out, such as, for example, the sample type and content, the column features, etc. This metadata cannot be automatically retrieved, but should be provided manually by the user in a key-value format, and it can originate, for example, from an electronic lab notebook (ELN). Metadata can subsequently be used to retrieve data that adheres to a specific characteristic, for example, to visualize instrument performance on only standardized QC runs.

Third, the raw file associated with a run contains the instrument parameters that were in effect during the execution of the experiment. The parameters that are recorded over the course of the experiment obviously depend on the model type of the instrument and are explicitly specified by a property, which denominates a single parameter, and is defined in a CV. Most MS experiments have a time dimension, for example, the RT when combined with liquid chromatography. Therefore, a particular instrument parameter has multiple values that are associated with this time dimension. In fact, every scan that records a mass spectrum can be associated with an instrument parameter analogue. As the main application of the iMonDB is to monitor instrument behavior over a long time interval (i.e. years) and not to monitor instrument stability over the period of a single LC-MS run (i.e. hours), summary statistics are computed for the repeated time course observations instead of storing each observation individually in the database. The advantage of this operation is that the data volume for quality control is significantly reduced, while the summary statistics provide sufficient information about the operational characteristics of the instrument. The parameter values of a single experiment are represented by their mean and median value, the first/third quartile, the minimum/maximum value, and the standard deviation.

Finally, another type of information that is related to the operation of a mass spectrometer are external events that may occur. For example, machine calibrations, periodic maintenance events, or even unexpected incidents that an operator likes to report, such as, for example, if an unusual sound is produced by a turbo pump. Unlike the instrument parameters, this type of information cannot be automatically retrieved and instead it has to be manually provided by the user. However, this information is vital when interpreting the evolution of the instrument parameters over time.

The previous sources of information are stored in a database structure that is both general and expressive. On the one hand it does not impose any restrictions on the mass spectrometer instrument manufacturer or even the specific instrument model type, as it is conceived as a flexible framework such that new instrument parameters can be added easily. On the other hand, the iMonDB is a rigid framework as instrument parameters can be unambiguously defined by making use of specialized CVs. By optionally combining the instrument parameters with a more advanced laboratory information management system (LIMS), such as, for example, colims [52], all information relevant to a mass spectrometry experiment can be structurally stored. Combined with external events, this data integration would enable highly advanced interpretations of the instrument parameters, for example, when contrasted to the protein coverage, identification scores, etc.

Instrument name	PSI-MS CV accession number
LCQ Deca XP Plus	MS:1000169
LTQ	MS:1000447
LTQ FT	MS:1000448
LTQ Orbitrap	MS:1000449
LTQ Orbitrap Discovery	MS:1000555
LTQ Orbitrap XL	MS:1000556
LTQ FT Ultra	MS:1000557
LTQ Velos	MS:1000855
TSQ Vantage	MS:1001510
LTQ Orbitrap Velos	MS:1001742
LTQ Orbitrap Elite	MS:1001910
Q-Exactive	MS:1001911
Orbitrap Fusion	MS:1002416

Table 4.1: List of explicitly supported instruments, along with their accession number in the PSI-MS controlled vocabulary [181].

4.2.2 Software implementations

In order to set up and populate the iMonDB several Java tools have been developed. First, there is a Java application program interface (API) geared toward developers. Based on this API there are processing and visualization tools available. All tools are released as open source and are available from the project website (`https://bitbucket.org/proteinspector/imondb`).

API

The iMonDB Java API features a high-level API for use by bioinformaticians and developers. The API provides all required functionality to extract instrument parameters from experimental raw files, and to store and retrieve this extracted data from the iMonDB. This API allows other developers to easily provide full iMonDB support in their own applications or to rapidly prototype powerful new tools.

The extraction of the instrument parameters from experimental raw files is based on functionality provided by the ProteoWizard library [48]. Making use of the abstraction layer ProteoWizard provides for experimental raw files, lightweight extraction tools have been implemented. Currently this functionality is limited to the extraction of instrument parameters from Thermo Scientific raw files. Table 4.1 lists all instrument models that are presently supported, however, other Thermo Scientific instruments might be supported implicitly, and explicit support for additional instruments is added regularly.

Specifically, two sources of instrument parameters are extracted from the raw files: the tune method information and the status log information. The tune method contains several configurable values for numerous instrument parameters, such as, for example, the flow of the sheath gas and the auxiliary gas for an electrospray ionization (ESI) probe, the voltage and temperature of the ESI capillary, the voltage of the multipole and various different lenses, etc. Obviously, the set values of these tune parameters depend on the application of the experiment. Furthermore, the status log contains information on all the sensors in the mass spectrometer, for which the values are recorded at the same time intervals as the mass spectral data. The status log includes the actual values of some parameters specified in the tune method, as well as additional

instrument parameters. For example, figure 4.3 shows the temperature of the iMonDB capillary in degrees Celsius, as configured in the tune method and as measured over the course of the experiments in the status log. The measured temperature roughly equals the set value in the tune method, however, it slightly deviates during the course of the experiments, as evidenced by the quartile values and the extreme values, which are used to summarize the repeated sensor readouts. Additional status log values include, for example, the pressure of the vacuum; the speed, power, and temperature of several turbo pumps; the flow rate of the syringe pump; etc.

To interact with the iMonDB, either to store extracted instrument parameters or to retrieve previously stored data, the iMonDB API makes use of the Java Persistance API (JPA) [125], and specifically the Hibernate implementation [113], to convert between records in the database and Java objects. This allows the user to work with high-level Java objects, instead of having to manage the various tables in the database and other technicalities. Retrieving or storing of instrument parameters in the iMonDB is implemented by executing the corresponding JPA persistence methods, while more advanced queries can easily be performed using the Java Persistence Query Language (JPQL) [77].

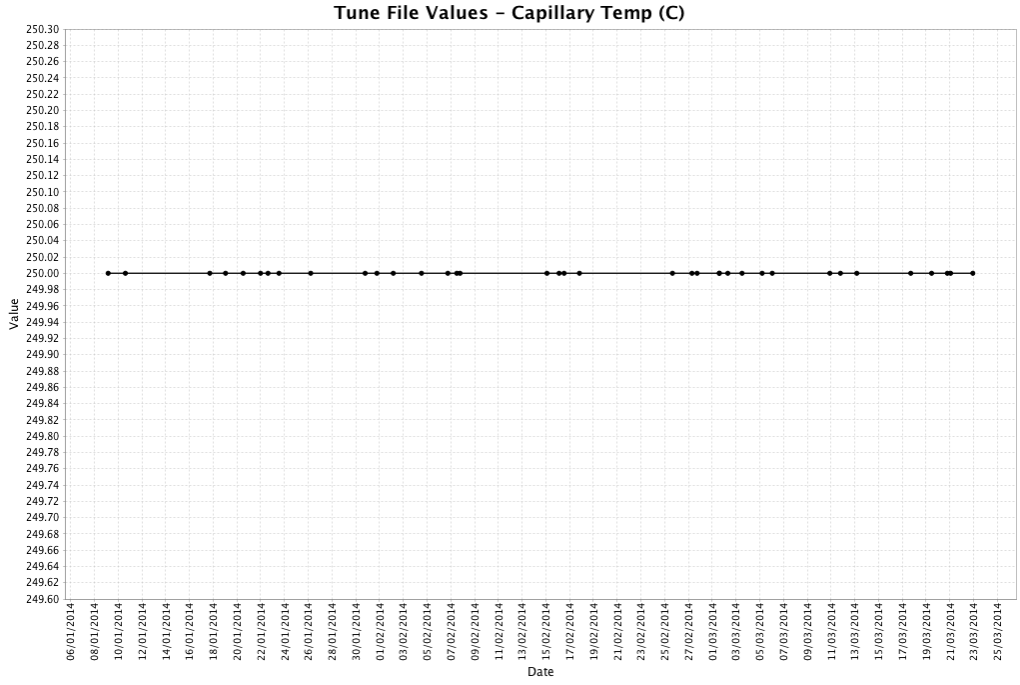
Data collection

The iMonDB Collector is a computer program that can be scheduled to keep the information in the iMonDB up to date. Experimental raw files for which the instrument parameters are not present in the iMonDB yet are gathered, after which these parameters are extracted from the raw files and stored in the database. Additionally, linking and mapping of external sensor information to the iMonDB can be done through the iMonDB Collector. The Collector can be run on a daily or weekly basis, and is fully configurable. For example, based on a lab's particular file naming scheme, specific metadata can be extracted from the file names and directory structures by making use of user-specified regular expressions. Extensive documentation for the full capabilities of the data collector are available on the project website.

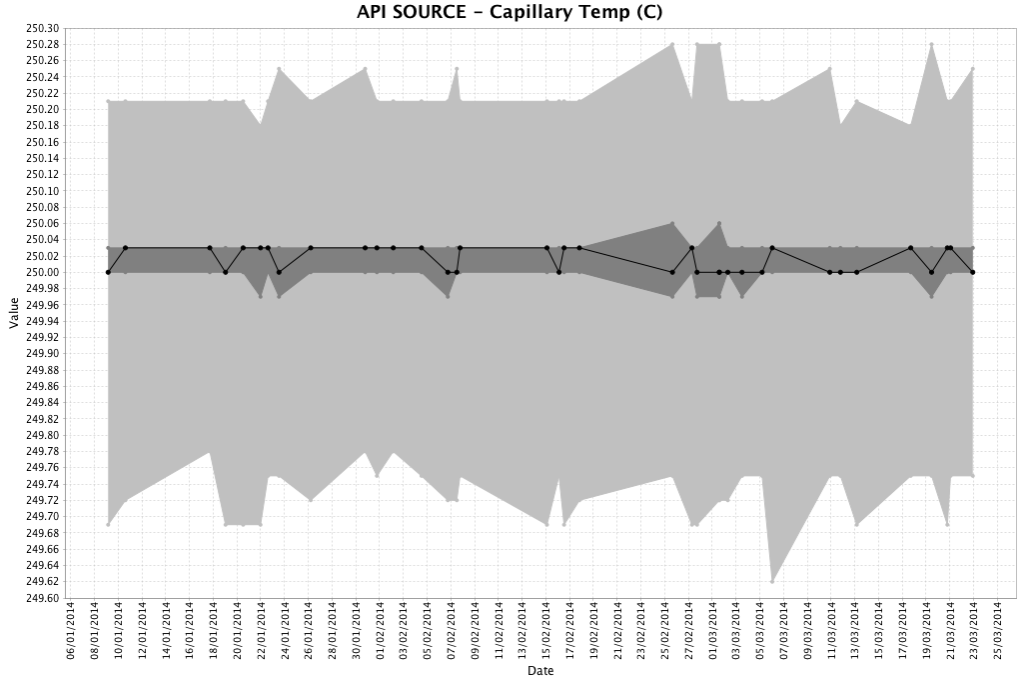
Data visualization

The iMonDB Viewer consists of a basic GUI application to visualize instrument parameters retrieved from the iMonDB. Each instrument parameter can be visualized over time, as is shown in figure 4.4. For each experiment the median value is shown in black, while the dark gray band depicts the first quartile and the third quartile, and the light gray band depicts the minimum and maximum values. This type of graph visualizes several elements of the summary statistics in an intuitive manner. Additionally, it is possible to select only the experiments that are associated with specific metadata to visualize only a subset of the experiments that are associated with a particular instrument.

Next, the vertical lines in figure 4.4 illustrate the external events submitted by the operator. It should be noticed that there are four different colors that represent predefined event categories. A user is able to add comments on instrument calibration, maintenance and service, downtime and errors, and an undefined category to store information on miscellaneous events, such as, for example, operator intuition. The causal effects of these events can be visually related to the trends in the instrument parameters. Obviously, these events cannot be extracted from the experimental raw files, therefore the iMonDB Viewer offers a reporting tool to add these events to the database and visualize them, as is shown in figure 4.5. Several event types can occur at a specific date, and can be augmented with customizable information on the observed problem, such as service reports or links to ELNs, and the actions that have been undertaken to solve the problem. This simple tool allows for a structured recording of all the events that occurred



(a) The set capillary temperature in the tune method.



(b) The actual capillary temperature in the status log.

Figure 4.3: Values for the temperature of the ESI capillary for a range of experiments. The median value is represented by the black line, the first and third quartile values are represented by the dark gray band, and the minimum and maximum values are represented by the light gray band.

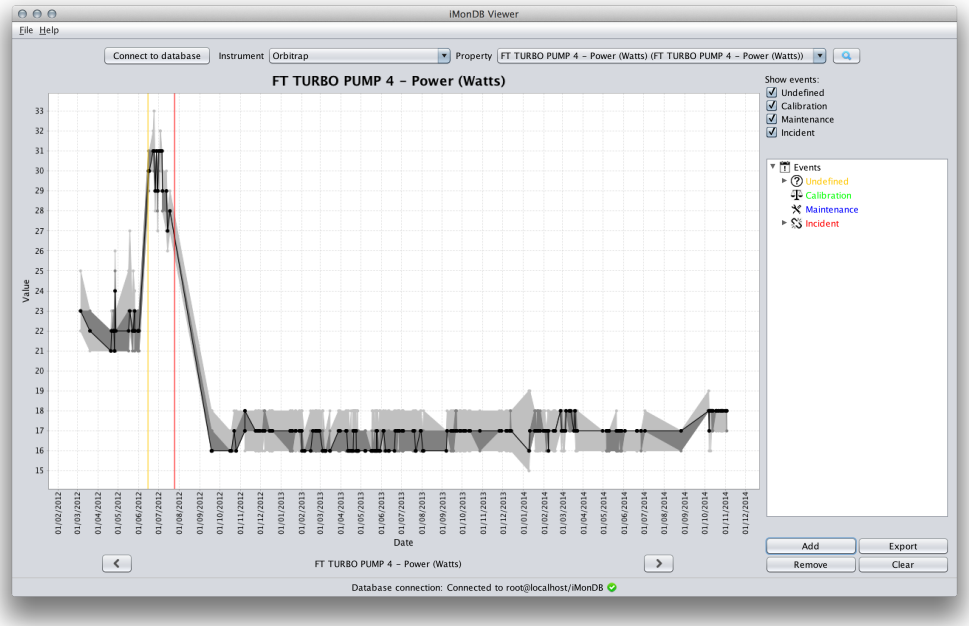


Figure 4.4: The viewer can be used to visualize how instrument parameters retrieved from the iMonDB progress over time.

on an instrument, which yields invaluable information for the downstream interpretation and analysis of the instrument parameters. Furthermore, the events are stored in the centralized iMonDB, which forms a single point of updated information, while users can easily retrieve this information on their personal computer by connecting to the iMonDB.

4.2.3 Case study

Instrument failure detection

A case study is presented to illustrate how the instrument parameters can be employed to detect/predict instrument failure. For this case study the application of the iMonDB was restricted to incorporate only the LC-MS runs originating from standard QC samples, and more particularly bovine serum albumin (BSA) samples. The advantage of such QC runs is that they are measured on a daily basis and that they easily allow us to assess the operation of a mass spectrometer because of their low sample complexity and controlled sample content [31]. It should be noted that in general the computation of QC metrics is not restricted to a specific sample type, but a homogeneous sample list may be required when interpreting some of the instrument characteristics. If LC-MS runs vary over time due to a change in the organism or a change in the wet lab protocol then it is difficult to link detected anomalies to a particular instrument artifact.

The graph displayed by the iMonDB Viewer in figure 4.4 shows the FT turbo pump 4 of a Thermo Scientific Orbitrap Velos on standard BSA LC-MS runs. From the graph, it can be noticed that the power consumption started to increase at the beginning of June 2012. The reason for this abnormal incline is that the turbo pump was deteriorating. In this case, intelligent electronics

Figure 4.5: Event information can be manually added and visualized through the iMonDB Viewer.

tried to increase the power consumption to remain fully functional. However, the increased power consumption was insufficient to continuously achieve the required speed and overcome the friction, after which the instrument finally broke down (down time: red line). After replacement of the turbo pump, the power consumption returned to its original value, and correct operation of the instrument could be resumed. Note that the instrument operator observed a high-pitched noise before the turbo pump finally broke down and reported this in the lab notebook (undefined event: yellow line).

While this is a clear-cut example of instrument malfunctioning, which eventually will lead to an intervention, by consistently monitoring the instrument parameters, this malfunctioning could have been predicted as the increase in power consumption serves as a proxy for turbo pump retrogression. Hence, by closely monitoring the instrument parameters and defining a normal range of operation, a suspected malfunctioning can be diagnosed early on and be reported. Next, based on this diagnosis, a timely intervention can prevent a highly undesirable loss of precious sample, analysis time, and effort.

Temperature monitoring

Basic smart sensors, purchased from Sen.se (`https://sen.se/`), were used to monitor the ambient laboratory temperature. These sensors are small hardware components that can measure both temperature and motion, and they can be unobtrusively attached to any laboratory equipment. Temperature monitoring was set up alongside a Thermo Scientific Q-Exactive mass spectrometer. Both temperature data and instrument parameters were collected over a period of several months in the second half of 2016, as shown in figure 4.6. Besides the explicit monitoring of the ambient

```
QcML qcml = new QcML();
qcml.setFileName(run.getName() + ".qcML");
QualityAssessment qualityAssessment = new QualityAssessment(run.getName());
qcml.addRunQuality(qualityAssessment);

int i = 0;
for(Iterator<Value> valIt = run.getValueIterator(); valIt.hasNext(); ) {
    Value value = valIt.next();

    Cv cv = convertAndAddCv(value.getDefiningProperty().getCv(), qcml);

    QualityParameter param = new QualityParameter(
    value.getDefiningProperty().getName(), cv, "value_" + (i++));
    param.setAccession(value.getDefiningProperty().getAccession());
    param.setValue(value.getFirstValue());
    qualityAssessment.addQualityParameter(param);
}

QcMLWriter writer = new QcMLFileWriter();
writer.writeQcML(qcml);
```

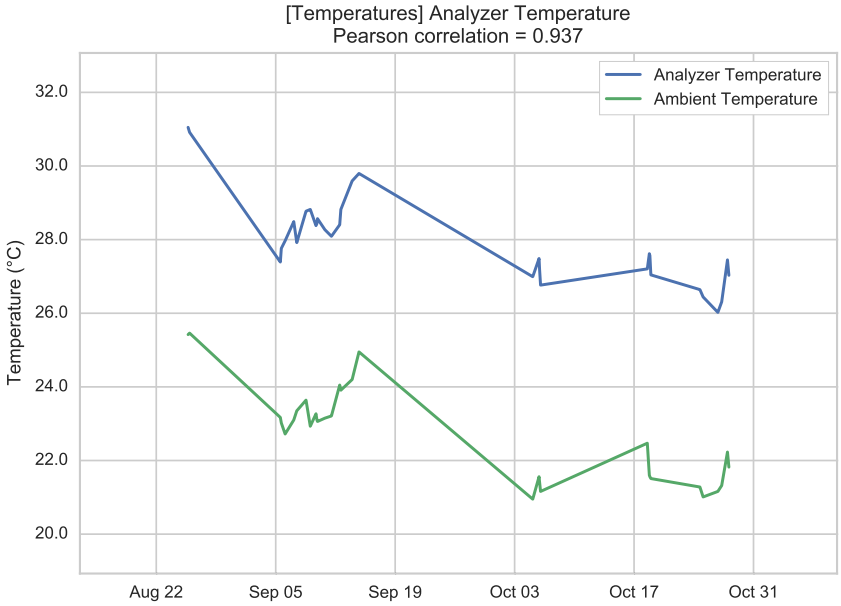
Listing 4.1: Exporting of instrument parameters to the qcML interchange format. The input `run` is a run retrieved from the iMonDB or extracted from an experimental raw file.

temperature, specific instrument elements measure the temperature internal to the instrument as well (figure 4.6a). Not surprisingly, the internal temperature is highly correlated to the ambient temperature, the latter which can therefore potentially influence the functioning of temperature-sensitive instrument elements. Furthermore, figure 4.6b shows an inverse correlation between the ambient temperature and the temperature measured by a cooling element, indicating that more cooling is performed at a higher ambient temperature. Correlations of other instrument parameters with the ambient temperature can be seen as well. For example, in figure 4.6c the pressure of the vacuum slightly changes according to the ambient temperature. Indeed, it is known that turbo pumps generating the vacuum frequently malfunction, notably when exposed to sustained and excessive heat. Additionally, in figure 4.6d an inverse correlation can be seen between the ambient temperature and the electrospray voltage.

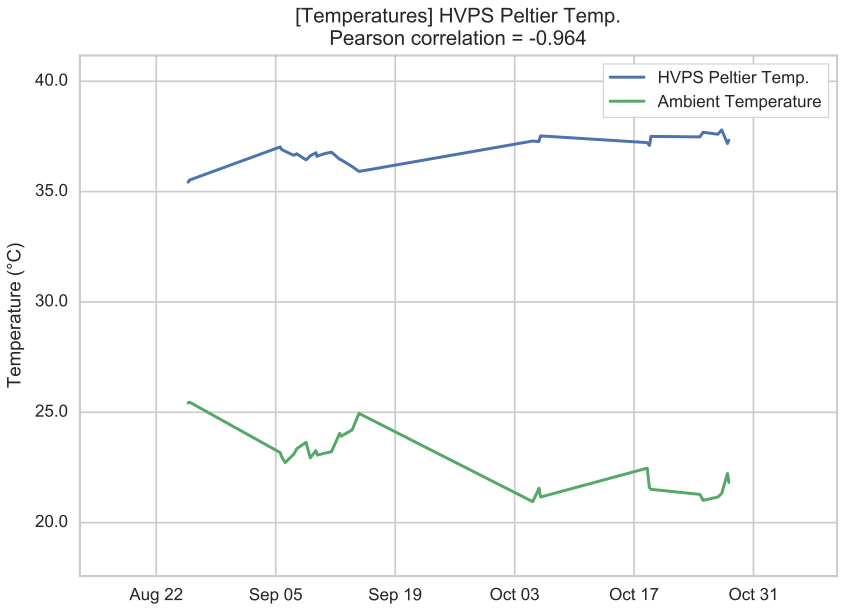
qcML export

Recently the qcML format [272] has been proposed as a standard data format for mass spectrometry QC information. Even though the primary purpose of the iMonDB is to provide longitudinal data storage, the qcML format can be used to easily interchange instrument parameters.

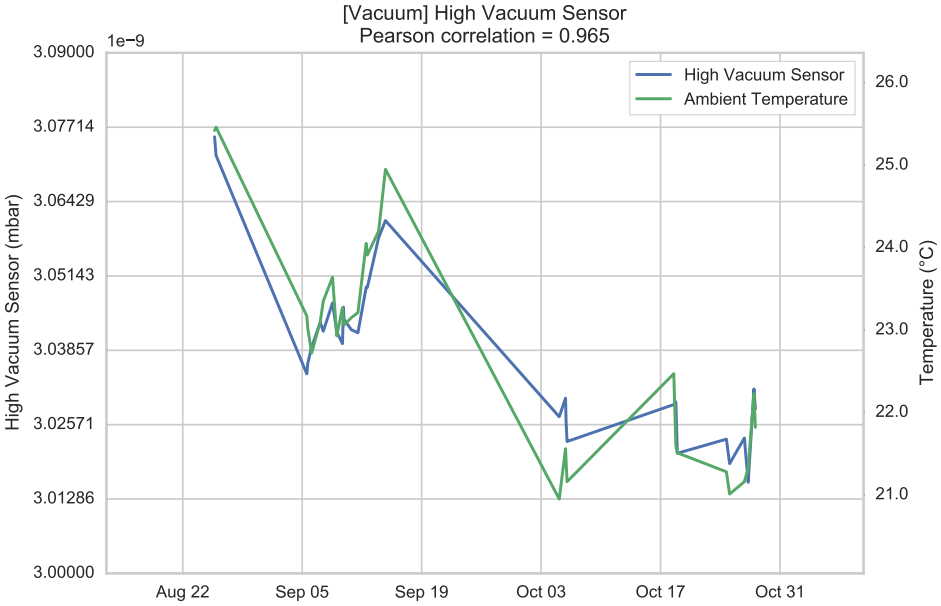
By making use of the iMonDB API and the equivalent jqcML API [29], it is straightforward to export data from the iMonDB to the eXtensible Markup Language (XML)-based qcML format. Listing 4.1 contains the full Java code that is required to store a specific run, extracted from the iMonDB, in a qcML file. Optionally the iMonDB can be forgone as well, and the instrument parameters can be extracted from an experimental raw file and directly stored in a qcML file by making use of the appropriate API functions.



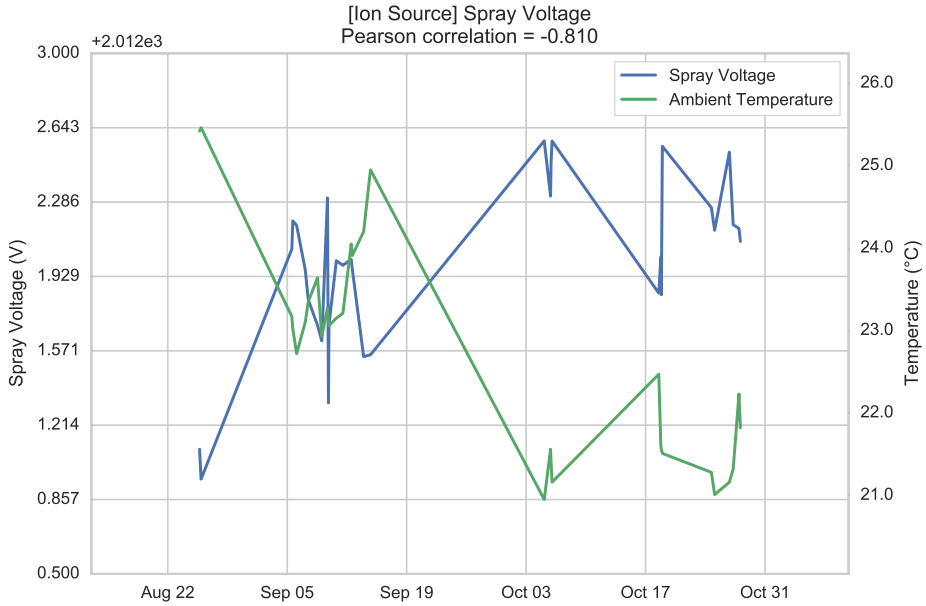
(a) Analyzer temperature.



(b) HVPS Peltier temperature.



(c) High vacuum sensor.



(d) Spray voltage.

Figure 4.6: Examples of instrument parameters that are affected by the ambient temperature. The strength of the relationship between the instrument parameters and the ambient temperature is indicated by the Pearson correlation.

4.3 Conclusions

Although quality control has recently been rightly identified as a vital element of an MS experiment, instrument parameters and environmental variables have so far been largely ignored. In contrast, this data forms an invaluable source of QC information. Instrument parameters are often directly related to the performance of the mass spectrometer and are able to identify possible problems in an early stage. And although the ambient temperature is a highly important factor influencing the experimental results, it is often overlooked and not systematically monitored even though commodity hardware suffices to measure the ambient temperature in an affordable fashion.

Here we have presented the iMonDB and associated tools to easily track and visualize instrument parameters and temperature information. Furthermore, an open-source Java API is provided to facilitate the handling of instrument parameters by other developers. Currently this API is only able to extract instrument parameters from Thermo Scientific experimental raw files, other vendors are not (yet) supported. Nevertheless, the iMonDB database structure is kept general but expressive, and allows for integration with other mass spectrometer vendors as well.

The highlighted case studies have illustrated the importance and potential benefits that can be gained by carefully monitoring secondary QC metrics. Tracking this information can help to assure a consistent and high-level quality of the experimental results. Furthermore, instrument downtime and sample loss can be avoided by scheduling a targeted maintenance when the metrics indicate that specific parts of the instrument are outside their normal range of operation, but have not completely broken down yet. Monitoring secondary QC metrics has a big potential to contribute to the stability of MS experiments, which is of vital importance for the development of robust proteome analysis workflows. Furthermore, for industry, the iMonDB can be supportive to manage and automate data provenance.

The iMonDB, its accompanying Java API, and various tools are freely available and are released as open source under the permissive Apache 2.0 license. The binaries, source code, and extensive documentation can be accessed at the project website at `https://bitbucket.org/proteinspector/imondb`.

Chapter 5

Making quality control more accessible

Abstract

In order to be confident of the results acquired during biological mass spectrometry experiments, a systematic approach to quality control is of vital importance. Nonetheless, until now only scattered initiatives have been undertaken to this end, and these individual efforts have often not been complementary. To address this issue, the Human Proteome Organization (HUPO) – Proteomics Standards Inititative (PSI) has established a new working group on quality control at its meeting in the spring of 2016. The goal of this working group is to provide a unifying framework for quality control data. The initial focus will be on providing a community-driven standardized file format for quality control. For this purpose the previously proposed qcML format will be adapted to support a variety of use cases for both proteomics and metabolomics applications, and it will be established as an official PSI format. An important consideration is to avoid enforcing restrictive requirements on quality control, but instead provide the basic technical necessities required to support extensive quality control for any type of mass spectrometry-based workflow.

Preface

This chapter combines two previously published papers on providing the necessary technical infrastructure to support quality control methods for biological mass spectrometry:

1. Wout Bittremieux et al. "jqcML: An Open-Source Java API for Mass Spectrometry Quality Control Data in the qcML Format". In: *Journal of Proteome Research* 13.7 (July 3, 2014), pp. 3484–3487. DOI: 10.1021/pr401274z
2. Wout Bittremieux et al. "The HUPO-PSI Quality Control Working Group: Making Quality Control More Accessible for Biological Mass Spectrometry". In: *Analytical Chemistry* (In revision)

Note that the jqcML application program interface (API) was first developed prior to the establishment of the HUPO-PSI Quality Control working group. The jqcML API comprises an essential element of the ecosystem around the qcML file format and it has been updated to incorporate the changes made to the standard based on feedback from the HUPO-PSI Quality Control working group.

5.1 Introduction

As mass spectrometry (MS) proteomics and metabolomics have matured over the past few years, a growing emphasis has been placed on quality assurance (QA) and quality control (QC), which are of crucial importance to endorse the generated experimental results. Mass spectrometry is a highly complex technique, and because its results can be subject to significant variability [251], suitable quality control is necessary to model the influence of this variability on experimental results. Potential sources of variability can include the introduction of unexpected modifications during sample preparation [133], limited stability of proteolytic digestion [262], the presence of contaminants [136], variability in the chromatography [15] and mass measurements [227], etc. A systematic approach to quality control makes it possible to quantify the technical variability within experimental results, which can inform subsequent data analysis steps and can be fed back to optimize the MS set-up. For example, Slebos et al. [235] have combined a meticulous experimental design with advanced computational quality control procedures to detect deviating measurements in a high-profile proteogenomic cancer study. This allowed them to trace the variability in the experimental results to both batch effects from instrument drift and biological variability, which would not have been possible by only examining high-level identification performance. Quality control plays an increasingly important role in such large-scale multi-site projects [46, 235, 252, 288] to enable an intra- and inter-laboratory comparison of experimental results. Additionally, although suitable quality control procedures are beneficial for any biological mass spectrometry application, a formal approach to quality control is of particular importance in a clinical setting [177, 249].

As a result, computationally derived QC metrics have been defined to objectively assess the quality of MS experiments [227]. These QC metrics capture quantitative information that may be related to the performance of the various processes in an MS experiment. The importance of quality control is exemplified by the recent proliferation of tools that can compute such metrics [21, 32]. However, most of these tools require specialized and non-standard experimental and software environments, which significantly hinders their evaluation and universal applicability. This problem is further exacerbated by the fact that each tool extracts different types of metrics from mass spectral data, and uses different frameworks to store, visualize, interpret, and communicate these metrics. In other words, the interoperability and comparability of these tools is essentially non-existent. These problems significantly hinder the systematic adoption of existing QC tools in well-established informatic pipelines. Consequently these tools are not yet adopted as a standard in the field, hindering biological mass spectrometry from reaching its full potential.

To address these issues a unifying framework for QC data is required, which would make it possible to bring these scattered initiatives together in a concerted approach, enabling a long-term strategy for quality control in proteomics. Moreover, we envision that in the future QC data will accompany MS data and associated results in repositories such as those coordinated by the ProteomeXchange Consortium [266] and MetaboLights [111]. This will, for instance, enable scientists interested in the reuse of these data to easily assess the quality of heterogeneous datasets in the increasingly extensive catalog of publicly available MS data. Therefore, the Human Proteome Organization (HUPO) – Proteomics Standards Inititative (PSI) [64] and members of the Metabolomics Standards Initiative (MSI) Data Standards task group [229] have established a new working group on quality control at the HUPO-PSI meeting during April of 2016 in Ghent, Belgium. This working group consists of a wide range of stakeholders from the proteomics and metabolomics communities and is composed of academic, government, and industry researchers, software developers, journal representatives, and instrument manufacturers. Its main goal is to define a community standard format for QC data and associated controlled vocabulary (CV) terms, in order to facilitate the use of QC metrics more broadly in the proteomics and metabolomics communities and to enable the data exchange and archiving of mass spectrometry-derived

QC metrics. It is important to emphasize that the working group does not seek to impose restrictive requirements on how quality control should be performed and how QC metrics should be interpreted. Instead, its aim is to provide the basic technical necessities to support extensive quality control practices over the whole course of an MS experiment and to bolster a strong community-driven ecosystem of quality control tools and methodologies.

This aim fits into the overall objective of the HUPO-PSI to define community standards for proteomics data to facilitate data comparison, exchange, and verification [64]. At the same time a strong emphasis is placed on full interoperability with metabolomics approaches. The new working group will exclusively address applications related to quality control, with previously established HUPO-PSI working groups focusing on mass spectrometry, proteomics informatics, molecular interactions, and protein separations. These working groups have previously established several standard data formats [54, 112, 219, 222, 282], CVs [180, 181], and minimum information guidelines [188, 253], which have significantly contributed to the maturation and unification of proteomics research.

5.2 Quality control for biological mass spectrometry

Myriad applications of mass spectrometry have been used in biology, each with specific properties and considerations. Therefore, no single strategy for performing quality control will be appropriate in all scenarios. Both external factors, such as sample preparation and environmental conditions, and instrumental factors, from autosampler to LC pump to column to MS method, contribute to the overall performance and should be measured in an appropriate quality control regime. For example, in shotgun proteomics, the total number of identified tandem mass spectra is a high-level QC metric that is often used to assess the performance of an experiment. However, this is mostly useful for discovery experiments, where the aim is to identify as many proteins as possible. It would not be reasonable, however, to count identified spectra in a selected reaction monitoring (SRM) experiment, since tandem mass spectra are only sometimes collected in this process, and they are generally not used as an input to database search. This implies that different types of QC metrics are needed to suit a wide variety of experiment types.

QC metrics can vary in the time scale of assessment, spanning the retention time (RT) of a single mass spectrometry liquid chromatography (LC) gradient or comparing among multiple experiments over the lifetime of an instrument. QC metrics may reveal information about different aspects of the experimental apparatus [227]. The metrics can be identification-free metrics that are computed from raw spectral data [273], which can be applied to some extent to both proteomics and metabolomics use cases. Alternatively, the metrics may depend upon application-specific results, such as identification performance, to draw inferences (for example, the extent of oxidation or carbamylation observed in a sample). Additionally, metrics that are closely tied to data from a particular class of instrument may retrieve information directly from the control software, such as column temperature or back pressure [231].

Different types of QC samples of varying complexity can be used. In proteomics the QC samples can range from a simple peptide mixture or a single protein digest, such as BSA, to a complex whole-cell lysate, such as a yeast or HeLa cell lysate. Complementary information can be provided by spiking synthetic mixtures into the experimental samples, which enables monitoring specific peptides of interest to measure the dynamic range (as one example). In metabolomics the QC samples can similarly exhibit different levels of complexity: they can be composed of either pooled samples (combining a small aliquot of each biological sample), or of mixtures of different compounds or chemical standards [100]. Pooled QC samples characterize the entire collection of samples included in the study qualitatively and quantitatively by providing an

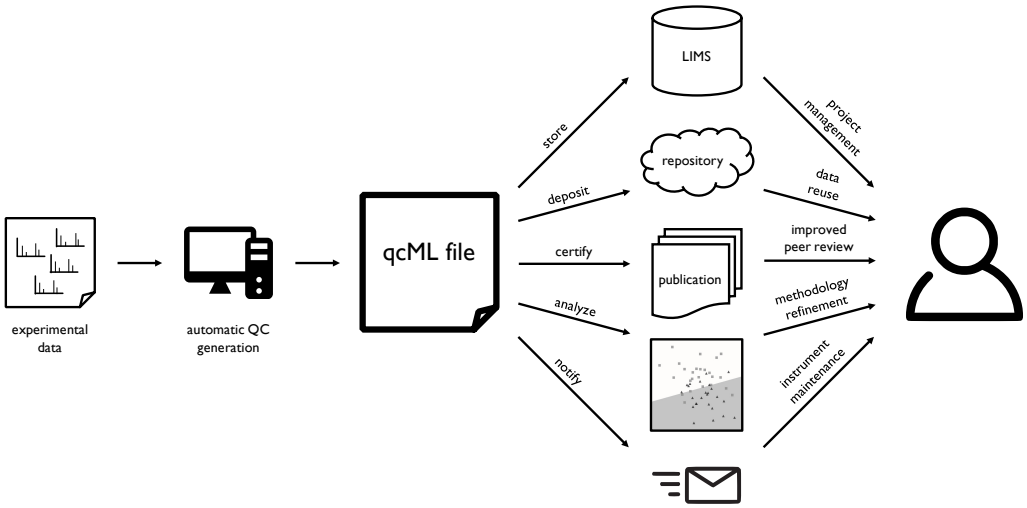


Figure 5.1: The qcML format is intended as the focal point for all QC applications. QC metrics are automatically generated as experiments are performed, whereupon the qcML data can be stored locally in a laboratory information management system (LIMS) where it can be used to support project management and data provenance. Additionally, the qcML data can be submitted to a public data repository, empowering data reuse, and it can facilitate peer review by certifying the data quality upon publication. Advanced analyses can aid informed decision-making to drive methodology refinements and automatically notify the instrument operator when a malfunction is detected.

average metabolome representation. As an alternative to pooled QC samples, predefined mixtures of certain biological fluids such as serum, plasma, or urine are commercially available (e.g. via National Institute of Standards and Technology (NIST)). Additionally, synthetic QC samples prepared under identical conditions and consisting of a mixture of compounds representing the different classes of metabolites expected to be present in the study samples can be used.

These different types of QC samples are not mutually exclusive; instead how they are used is closely linked to the experimental design [169], as they are each able to measure specific performance characteristics, and they should be used in combination. One consideration is how many QC samples of each type should be used; another is how to interleave them with experimental samples [21]. Further, besides these dedicated QC samples, other commonly used sample types to assess the quality of a run or an instrument in proteomics and/or metabolomics include: blanks, used to monitor or control instrument contamination; calibration curve samples, spanning a wide dynamic range and consisting of pure reference compounds; and internal standards, usually consisting of synthetic peptides or of a single metabolic compound or a mixture of compounds that is added to all samples to monitor the reproducibility of the analytical methods. Another source of qualitative information comes from replicate measurements; based on the experimental design these replicates provide identical inputs for each injection and are typically used to compensate for variance across the analytical study [68].

5.3 A community-driven standard file format for QC data

As there is such a wide variety in the composition of mass spectrometry workflows and corresponding quality control methodologies, it is impossible to define a single fixed QC directive.

Instead, the HUPO-PSI Quality Control working group wants to facilitate quality control for a wide variety of configurations by providing the basic technical foundation. To successfully communicate and interpret advanced quality control information a unified frame of reference is required. To this end a community-driven standardized format for the archival, transmission, analysis, and visualization of QC metrics derived from mass spectrometry will function as the focal point supporting various advanced tasks, as represented in figure 5.1. The definition of an unambiguous and expressive standard file format supported by powerful application program interfaces (APIs) and robust tools will allow bioinformaticians to focus on uncovering novel biological knowledge instead of being encumbered by low-level implementation details. This requisite technical infrastructure will be developed in the context of the HUPO-PSI Quality Control working group. Crucially, the file format constituting the centerpiece of the QC ecosystem will support metrics of an arbitrary type to accommodate QC information relevant for all kinds of experimental configurations. For this we will adopt the previously proposed qcML file format [272] as the starting point. Although this format is not an official PSI standard yet, it has been developed according to the same philosophy, and it has received considerable feedback at the 2016 HUPO-PSI meeting. Based on this feedback, an updated version of the qcML format will be developed, which will subsequently be submitted to the PSI formal document process [268] to establish it as a PSI standard format. A key part of this work will connect the qcML format and a CV in accordance with the previously established PSI CV [181] to enable direct interpretability of the collected metrics and addition of further metrics without the need to update the standard format to a new version. When metrics are defined in CV terms, they are more comprehensible, even as QC pipelines gain complexity and cover thousands of mass spectrometry acquisitions. This will hopefully also elevate the ease of informed decision-making during analysis processes [9, 30, 273] and contribute to the prevalence of applied QC in biological mass spectrometry. An important aspect that ties in with this semantic interpretability is the development of a Minimum Information About a Proteomics Experiment (MIAPE)-like document for quality control, as suggested earlier [73]. The information specified in the MIAPE-QC document will present opportunities for extensive linking of the QC data to the experimental results. For metabolomics, some earlier work on capturing the minimum information for reporting on quality control already exists [83, 230], but this will need to be revisited in coordination with the Metabolomics Society Data Quality and Data Standards task groups [17, 229].

To make the qcML format more attractive to the authors of quality metric generators, we will create a software library that is able to import and export information in the qcML format, which will enable developers to easily create new qcML files and extract information from existing qcML files. This will mainly be mediated in the form of the jqcML Java API [29]. Although at present the jqcML API only structurally validates qcML files against the schema definition, additional functionality to semantically validate the qcML files based on the terms defined in the CVs and the MIAPE-QC specification will be added [188]. Similar to APIs for other standard formats [54, 112, 219, 222, 282], the availability of the jqcML API will assist developers to support the qcML format, fostering the interoperability of QC tools. For example, this will enable the construction of a custom tool workflow where a first tool generates various QC metrics, a second tool applies advanced algorithms to draw inferences from the data, and a third tool provides long-term storage and visualization. Instead of a rigid, monolithic framework, the qcML format and the jqcML API will support the construction of modular, highly customizable QC pipelines. Furthermore, besides updates to existing support for the qcML format in OpenMS [226] and SimpatiQCo [215], native support will be added to other tools developed by members of the working group as well, such as QuaMeter [167] and iMonDB [34]. We will develop a user-friendly graphical user interface (GUI) tool for visualization of quality control data in the qcML format, including established outlier detection techniques to automatically identify low-quality experiments [30, 273]. This tool will enable visual data exploration and easy downstream processing of QC data. These steps

will foster broader interest in and adoption for the qcML format, both from developers and end users.

5.3.1 The jqcML Java API for the qcML standard

The success of any standard file format rests in large part on the existence of software libraries in popular programming languages that allow easy read and write access to this format. jqcML [29], a fully operational, production-grade Java API, aims to fill this role for the qcML data standard [272]. Written in Java, the jqcML API is inherently platform independent. Built to be used in demanding circumstances, jqcML provides a complete object model to interpret and manipulate qcML data while retaining a minimal memory footprint and without sacrificing the overall speed of data access. Furthermore, jqcML is able to interact in a uniform and transparent way with QC data in eXtensible Markup Language (XML)-based qcML files as well as the qcDB relational database [272]. This dual access enables the user to transparently work with qcML data from both sources, without requiring any changes in the data processing code.

jqcML will prove to be useful for both producers and users of qcML data. Producers can use jqcML to directly export their data in compliance with the qcML format, or convert between their existing output results and the qcML format. Meanwhile, end users will have out-of-the-box access to qcML data, without having to worry about implementation details. This universal applicability is reinforced by the fact that, as a pure Java API, jqcML is completely platform-independent.

jqcML is freely available, and is released as open source under the permissive Apache 2.0 license. The binaries, source code, and documentation can be downloaded from the project website at `https://bitbucket.org/proteinspector/jqcml`.

Object model

The data defined by the qcML data format is represented by a complete object model consisting of simple Java classes. It is of note that the qcML data format is less complex than related standard formats, such as mzML [179]. The object model intends to reflect this relative simplicity, while at the same time providing expressive access to the data. The information specified by the qcML standard can be represented by the object model, and includes basic quality metrics, as well as tabular attachments or more complex attachments (such as binary strings). Figure 5.2 highlights the key components of the jqcML object model.

Data processing

As mentioned earlier, jqcML is capable of reading and writing XML-based qcML files, but also fully supports reading from and writing to the qcDB relational database. The interaction with both types of data sources is generalized through common interfaces, allowing the underlying data access implementation to be abstracted from the user. This generalized object model thus allows the user to handle the different data sources in a uniform manner, making it trivial to switch between several different data sources. A simplified schema of the workflow is shown in figure 5.3.

Similar to other XML-based proteomics standards, the XML-based qcML file structure is intended to exchange QC data between multiple users or organizations, while the qcDB relational database will most likely be used within a single organization in order to store large amounts of QC data over time, and to perform data retrieval and analysis operations across extensive collections of

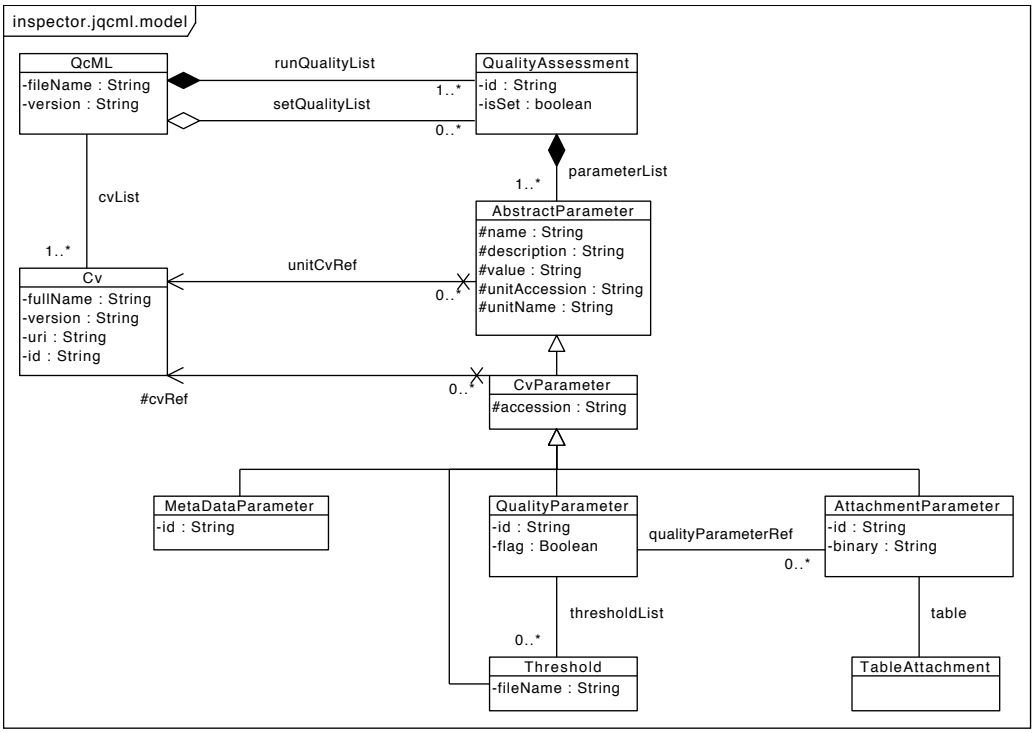


Figure 5.2: Class diagram highlighting the key components of the jqcML object model for qcML version 0.0.8.

qcML data. In order to ease the translation of QC data from one format to the other, jqcML also supports the easy conversion of data between both representations. The key implementation aspects of jqcML for the XML and relational database formats are discussed next.

XML-based file format To interact with the XML-based qcML files, the Java Architecture for XML Binding (JAXB) [124] is used in the form of the EclipseLink MOXy implementation [70]. The interaction between the object model and the XML-based file format is handled through the mapping of the formal XML schema definition of qcML to specific elements of the jqcML object model. This mapping allows for an automatic translation between the object model and the XML structure, which enables both reading of qcML files into jqcML Java objects, as well as exporting jqcML objects to qcML files.

Because a qcML file can contain data from several different runs, it can in principle become arbitrarily large. When working with such very large files, precautions should be taken to guarantee that there is sufficient memory available to process the file. Therefore, jqcML is designed to elegantly handle very large qcML files while controlling the memory requirements. For this purpose, jqcML exploits an XML indexer component, xxIndex [188, 283], that allows the qcML file to be accessed like an indexed random access file. This fragmented data access prevents the reading of a complete qcML file into memory, allowing jqcML to read even very large qcML files while retaining a minimal memory footprint. The XML indexer approach also enables the user to retrieve only the data of a single experiment as an iterative procedure, such that only a subset of the full qcML file is kept in main memory at once.

The XML-based file structure makes use of internal references to crosslink various elements, for example to link an annotation parameter to the corresponding entry in a controlled vocabulary.

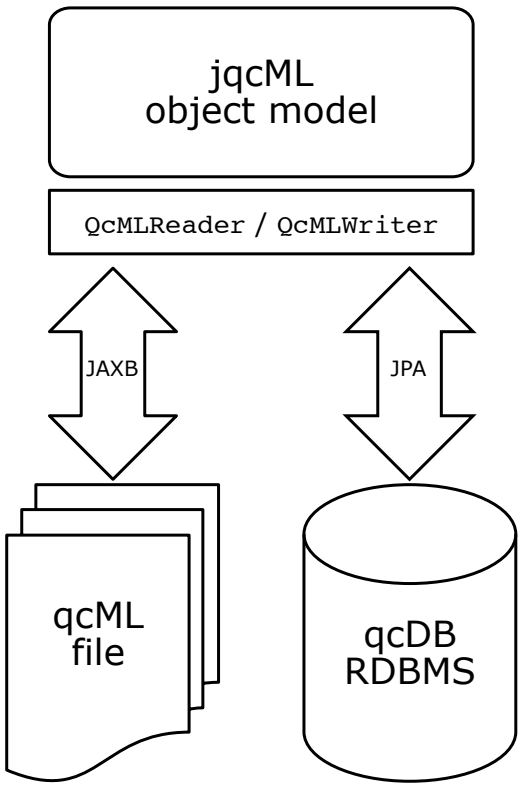


Figure 5.3: Simplified representation of the jqcML architecture. Through the use of the common `QcMLReader` and `QcMLWriter` interfaces, jqcML is able to work with qcML data from several sources in a uniform way.

When only retrieving an individual element, as explained previously, these references may initially not refer to existing Java objects, because only part of the qcML file is read into main memory at once. However, the use of xxIndex enables jqcML to resolve these references automatically as well. A complete object model is thus always accessible, even when reading only a section of a qcML file.

Relational database structure In addition to the XML-based file format, the qcML specification includes an entity-relationship model as a reference implementation of a qcDB relational database to store qcML data. In order to interact with a qcDB, the Java Persistance API (JPA) [125] is used, and in particular the EclipseLink JPA implementation [69]. In analogy to the way JAXB is used for interacting with XML-based qcML files, JPA is used to interact with a relational database by creating a mapping between the object model and the database model.

By using JPA, the need for low-level operations, such as Structured Query Language (SQL) queries, can be avoided in the code. Instead of having to manage the various tables in the database and other technicalities, operations can be defined on the level of the object model. An additional advantage of defining operations on a higher level is that the user does not need to focus on implementation details, such as adapting to the specific SQL syntax of the chosen relational database engine.

Retrieving or storing of quality control data in a qcDB is implemented by executing the corresponding JPA persistence methods, while more advanced queries can easily be performed using the Java Persistence Query Language (JPQL) [77]. Equivalent to the JAXB access layer discussed

above, the JPA implementation supports fragmented data retrieval of the QC information based on an iterative procedure.

jqcML currently fully supports both the MySQL and SQLite database engines. However, other relational database systems can easily be included as well. The only requirement is the availability of a suitable Java driver for the specific type of database. When initializing a new system, it is even possible to let jqcML automatically create the required tables in the database.

5.4 Broadening the applicability of quality control

Although shotgun LC tandem mass spectrometry (MS/MS) proteomics and metabolomics can assuredly benefit from wider adoption and automation of quality control, other classes of data in proteomics and metabolomics can also benefit from these tools. The working group particularly emphasizes the use of quality control in quantitative proteomics methods, from isobaric tags for relative and absolute quantitation (iTRAQ) to data-independent acquisition (DIA) datasets, where details of the experimental procedures can be employed to compute further relevant and focused QC metrics. For example, for iTRAQ experiments the isobaric tagging reagents can be a source of variability, influenced by the protein abundances, and fold changes are biased towards $1:1$ ratios [171]. The labeling efficiency can broadly be determined by verifying the fraction of MS/MS spectra for which reporter ions are observed. Additionally, as iTRAQ reagents bond to primary amines both at the N-terminus and on lysine residues the labeling efficiency can be determined in full detail by evaluating the extent to which the labels are present on one or both of these sites. Furthermore, the labeling stability can be evaluated based on the evolution of the reporter ion intensity over the course of an experiment and their signal to noise ratios.

In contrast to during a data-dependent acquisition (DDA) experiment, during a DIA experiment MS/MS scans are measured with wide isolation windows that do not target any particular peptide precursor [98]. This way all analytes within the desired precursor mass range can be measured in an unbiased fashion, potentially leading to an increase in reproducibility. To evaluate the performance of a DIA experiment general QC metrics from the DDA setting can likewise be applied. In both cases the consistency of the LC and MS performance can be evaluated using well-characterized standard QC samples [78], which can for example be visualized using the powerful Skyline software tool [72]. Furthermore, specialized QC metrics can be defined based on the characteristics of a DIA experiment. For example, the isolation window size can be evaluated based on the rate of ion interference [291]. This is sample-dependent, as complex samples will lead to a lower precursor selectivity resulting in highly complex chimeric MS/MS spectra. Because all analytes are reproducibly measured during a DIA experiment consecutive MS/MS scans can be compared to each other to further evaluate the LC and MS performance. Measurements of the same analyte over repeat scans can be used to assess the mass accuracy, while successive scans covering the same isolation window (separated by the duty cycle) can provide information on the chromatographic sample rate. An important step during DIA spectrum identification is the correlation of a precursor ion with its corresponding product ions. Although a high rate of cofragmentation complicates an accurate precursor–product correlation this is essential for obtaining correct peptide identifications. Some QC metrics that can be used to evaluate whether precursor and product ions are accurately correlated are for example the fraction of MS isotopic packets that match product ions [95], the RT variability of associated ions [40], and whether or not the elution profiles of corresponding precursor and product ions match a similar exponentially modified Gaussian peak shape [128].

Matrix-assisted laser desorption/ionization (MALDI) imaging mass spectrometry is an increasingly popular technique for molecular imaging, yet a standardized approach to quality control

has not been established thus far. Currently, quality assessments are typically done manually through visual inspection of the spectra. Additionally, simple plots can be employed to evaluate the variation in peak intensities among different measurements [132]. This can, for example, be complemented by an ion intensity histogram to highlight specific regions of poor signal. However, these QC methods for MALDI imaging usually disregard the spatial information present in the data. Conversely, important qualitative measures can be defined by comparing nearby measurements as, for example, spatial proximity has an influence on the intensity similarity. A recent result has shown how image analysis measures based on sliding windows of increasing sizes, which consider the spatial information implicitly, can be used to accurately replicate manual expert quality assessments [204].

These are but a few examples of how quality control can be expanded to further mass spectrometry technologies. To adequately cover all these different workflows the novel qcML standard will have to be flexible enough to support MS data ranging from shotgun proteomics to SRM to MALDI imaging data. This effort will require a broader perspective than has dominated QC software to date. The HUPO-PSI QC working group has members from both the proteomics and the metabolomics communities and welcomes any contributions, ensuring wide applicability of the qcML open data standard in a variety of mass spectrometry-based settings.

5.5 Conclusions

Quality control will indubitably play a growing role in aiding the maturation of biological mass spectrometry as a field. A systematic approach to quality control will aid researchers to assess their workflows over time or to compare data among different laboratories. Furthermore, it can stimulate public data reuse to harvest new knowledge [178]. We envision that the inclusion of QC metrics will become an integral component when submitting datasets to public data repositories or for peer-reviewed publication, similar to the situation for three-dimensional structures in the Worldwide Protein Data Bank (wwPDB) [25].

The HUPO-PSI seeks to broaden the conversation surrounding quality control in our community. This effort will provide a format definition along with examples and software infrastructure that will enable new research in the interpretation of QC metrics. We emphasize that this is very much a community effort, and any and all contributions are welcome. Our group charter is available on the HUPO-PSI website (`http://psidev.info/groups/quality-control`), including a summary of the milestones the working group wants to achieve. You can connect with us through our GitHub repository (`https://github.com/HUPO-PSI/qcML-development`) and through our mailing list (`Psidev-qc-dev@lists.sourceforge.net`). These online sources contain further detailed instructions on how to get started and how to contribute.

Chapter 6

Unsupervised quality assessment of experiments

Abstract

Despite the availability of various sets of quality control metrics which can be used to understand and evaluate how technical variability affects the results of an experiment, a systematic approach to quality control is often still lacking because the metrics are not fully understood and hard to interpret. Here we present a toolkit of powerful techniques to analyze and interpret multivariate quality control metrics to assess the quality of mass spectrometry proteomics experiments. We show how unsupervised techniques applied to these quality control metrics can provide an initial discrimination between low-quality experiments and high-quality experiments prior to manual investigation. Furthermore, we provide a technique to obtain detailed information on the quality control metrics that are related to the decreased performance, which can be used as actionable information to improve the experimental setup. Our toolkit is released as open source and can be downloaded from `https://bitbucket.org/proteinspector/qc_analysis/`.

Preface

This chapter was previously published as:

Wout Bittremieux et al. "Unsupervised Quality Assessment of Mass Spectrometry Proteomics Experiments by Multivariate Quality Control Metrics". In: *Journal of Proteome Research* 15.4 (Apr. 1, 2016), pp. 1300–1307. DOI: 10.1021/acs.jproteome.6b00028

This work was presented at the Benelux Bioinformatics Conference 2015 in Antwerp, Belgium, where it won the "outstanding oral presentation" award.

After publication it garnered multiple mentions on social media and blogs, such as the well-known *News in Proteomics Research* blog by Ben Orsburn [202]. This illustrates the vital importance of such techniques and reinforces the conclusions which state that there is a significant and urgent need for user-friendly bioinformatics tools to support a systematic approach to quality control.

6.1 Introduction

The variability exhibited by mass spectrometry experiments can originate from different sources, such as steps in the experimental setup that are not (yet) fully understood, stochastic processes, or the bioinformatics processing workflow, and it impedes achieving reproducible results across multiple experiments [46, 251]. To understand and evaluate how technical variability affects the results of an experiment, several quality control (QC) and performance metrics have been introduced [32]. These metrics are computationally derived from the output of a mass spectrometry experiment and aim to capture the important operational characteristics of a mass spectrometer to provide an objective evaluation of the quality of the experiment.

However, despite the availability of QC metrics covering a wide range of qualitative information, a systematic approach to quality control is still lacking. Instead, quality control practices frequently involve only monitoring a few simplified performance measures in a spreadsheet, or sometimes are even limited to only checking QC metrics retrospectively in an ad hoc fashion when a malfunction is suspected.

One of the barriers impeding the adoption of a systematic quality control workflow is the lack of knowledge about what specific QC metrics signify. It is often unclear which metrics can be applied to detect which problems, and the acceptable variability within metrics is ill defined [21]. Most aforementioned tools illustrate their applicability by highlighting a few particular use cases where they were able to detect inferior experiments, however, this often does not translate to a more general setting. Specific metrics might only be relevant for highly exceptional situations, or the metrics might be hard to interpret and link to actionable solutions even for a domain expert.

Another issue that is generally overlooked when interpreting QC metrics is that interdependencies have to be taken into account. During a mass spectrometry (MS) experiment the different steps do not function in isolation, instead they influence each other. As most tools only look at a single metric at the same time in a univariate fashion, these dependencies are ignored, which can lead to erroneous results [21]. To accommodate for these dependencies a multivariate approach can be employed, as has been done by Wang et al. [273], who analyzed the variability present in samples originating from multiple National Cancer Institute (NCI) Clinical Proteomic Tumor Analysis Consortium (CPTAC) studies using multivariate statistics, such as principal component analysis (PCA). Whereas Wang et al. [273] mainly applied unsupervised techniques, a different approach was employed by Amidan et al. [9]. Here, first the quality of a multitude of experiments was manually reviewed by expert instrument operators, after which this labeling was used to train a supervised classifier to distinguish good experiments from poor experiments. In general, such multivariate approaches can protect against false positives in the event of a high number of variables and can allow to detect patterns that are invisible when evaluating each metric individually. An additional advantage is that certain multivariate techniques provide insight on how some of the variables are related to each other, which can improve the understanding of the QC metrics and their interpretation.

Here we will illustrate how computationally derived QC metrics can be used to provide an initial discrimination between low-quality and high-quality experiments in an unsupervised fashion prior to manual investigation. The presented techniques will take into account the multidimensional feature space exhibited by the QC metrics, while also prioritizing the most relevant QC metrics. A special emphasis will be laid on the interpretability of the obtained results, as in order to integrate a systematic approach to quality assurance (QA) into existing mass spectrometry proteomics workflows, the interpretability of the qualitative information, even to

non-expert users, is paramount. Finally, we will unify these different steps to present an open-source toolkit of powerful techniques for the analysis and interpretation of mass spectrometry proteomics quality control metrics.

6.2 Quality control metrics

Our goal is to discriminate the low-quality experiments from the high-quality experiments when considering multiple experiments. Suitable data for this kind of analysis are for example standard quality control samples, such as the simple bovine serum albumin (BSA) samples that are used by several labs. The advantage of such QC runs is that they are measured on a very frequent basis and that their low sample complexity and controlled sample content allow to easily asses the operation of a mass spectrometer. Although the computation of quality control metrics is not restricted to such samples, a controlled and consistent sample content and operating procedure facilitate interpreting the analysis results. If liquid chromatography (LC)-MS runs vary over time due to biological or technical changes in the sample itself, such as a different wet lab protocol or a different biological condition, this may complicate discerning the origin of potentially detected anomalies.

6.2.1 Experimental data

We used two public datasets to investigate the experiment quality. The first dataset consists of a number of standard QC LC-MS runs performed on several different instruments at the Pacific Northwest National Laboratory (PNNL) [9]. Each sample had an identical content (whole cell lysate of *Shewanella oneidensis*), and the quality of the various runs has been manually annotated by expert instrument operators as being either "good", "ok", or "poor". We split up the various runs depending on the instrument type, with each instrument group consisting of multiple individual instruments. Both the experimental raw files and the expert annotations have been retrieved from the PRoteomics IDEntifications (PRIDE) database [267]. Please see the original publication for further information on the experimental procedures [9].

The second dataset was generated as part of The Cancer Genome Atlas (TCGA) project where the aim was to perform a proteogenomic characterization of human colon and rectal cancer [288]. For this study 95 samples from 90 patients were obtained, with each sample fractionated into 15 concatenated peptide fractions before being subjected to an LC-MS analysis. This resulted in 1425 raw files, which were retrieved from the CPTAC data portal [71]. For a more detailed exposition of the sample content and preparation, please refer to the original publications [235, 288].

Table 6.1 shows an overview of how the data from these two sources has been split into four datasets. The PNNL datasets consist of a unique public data resource, as they not only contain high-quality measurements, as is common, but explicitly also include low-quality measurements. Furthermore, the expert annotations provide crucial information to validate the detection of low-quality experiments. The TCGA dataset can be used to highlight how instrument performance evolves over time when working with complex sample contents, because all raw files have been obtained on a single Orbitrap Velos mass spectrometer using the same operating procedure over an extended time period.

Reference	Denomination	Instrument model (accession)	Number of raw files	Expert annotation?
Amidan et al. [9]	PNNL LTQ-IonTrap	Thermo LTQ MS (MS:1000447)	225	yes
Amidan et al. [9]	PNNL LTQ-Orbitrap	Thermo LTQ Orbitrap (MS:1000449), Thermo LTQ Orbitrap XL (MS:1000556)	379	yes
Amidan et al. [9]	PNNL Velos-Orbitrap	Thermo LTQ Orbitrap Velos (MS:1001742)	538	yes
Zhang et al. [288]	TCGA	Thermo LTQ Orbitrap Velos (MS:1001742)	1425	no

Table 6.1: Overview of the characteristics of the various datasets. The instrument accession numbers refer to their identifiers in the PSI-MS controlled vocabulary [181].

6.2.2 Metrics generation

Over the past few years multiple sets of quality control metrics have been defined [32]. Here we focus on so-called identification (ID)-free metrics, which have the advantage that they are directly derived from the raw data and that they do not depend on identification results. This enables generating the metrics as soon as the experimental raw data is available, instead of having to analyze that raw data in a potentially computation-heavy peptide and protein identification workflow. Furthermore, the lack of dependency on the identification results eliminates possible sources of computational variability and prevents suboptimal settings in the various bioinformatics steps from influencing the quality control [19].

Specifically, we used the ID-free metrics computed by QuaMeter [273], which are listed in table 6.2. These metrics are derived from the raw spectral data and provide information on various stages of a mass spectrometry experiment: they include information on the chromatography, the MS and tandem mass spectrometry (MS/MS) performance, and the charge distribution. For the PNNL data, several different sets of QC metrics were already computed, and we restricted the metrics under consideration to the QuaMeter metrics. For the TCGA data, we used QuaMeter version 1.1.91 to produce the QC metrics.

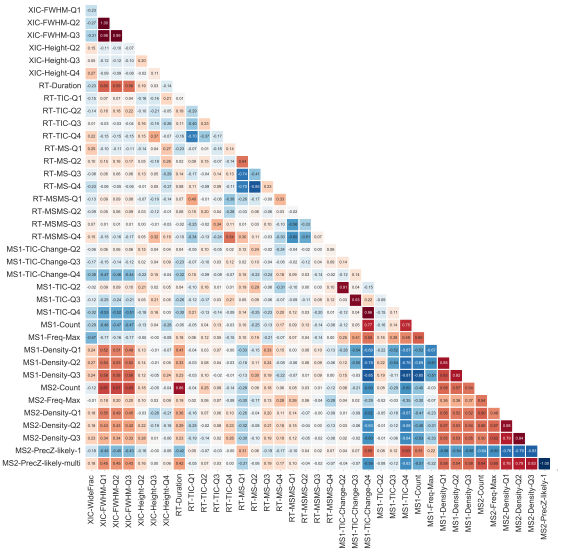
6.2.3 Preprocessing

To prepare the QC metrics for analysis several preprocessing steps have to be performed. First invariant metrics with a low information content are removed, as they have the same value for each experiment and needlessly increase the dimensionality without adding additional information. Furthermore, mutually dependent metrics can be derived from each other, so the duplicated metrics can be omitted without any information loss. Removing the low-variance and correlated metrics decreases the dimensionality while retaining all embedded information, and prevents irrelevant metrics from deteriorating the subsequent analyses.

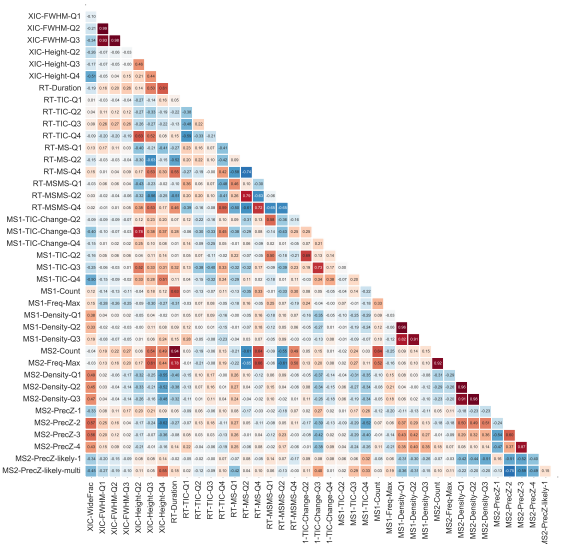
Although the occurrence of these extraneous metrics depends on the operational characteristics, and varies depending on the experimental setup, for the 44 ID-free metrics computed by QuaMeter [273] and listed in table 6.2, some specific metrics generally seem less expressive than others. For example, some charge states only occur rarely, resulting in a very low variance for the metrics denoting the uncommon charge states. Furthermore, some metrics are very often highly correlated (Pearson correlation above 0.90). The chromatography metrics denoting the full width at half maximum (FWHM) (XIC-FWHM-Q1, XIC-FWHM-Q2, XIC-FWHM-Q3) can generally be represented by only a single observation, as do the metrics denoting the MS and MS/MS spectral counts (MS1-Density-Q1, MS1-Density-Q2, MS1-Density-Q3, and MS2-Density-Q1, MS2-Density-Q2, MS2-Density-Q3). Figure 6.1 shows the correlation matrices for all four different datasets.

Metric	Category	Metric Group	Units	Description
XIC_WideFrac	Chromatography	Peak width variability	Ratio	Fraction of precursor ions accounting for the top half of all peak width
XIC_FWHM_Q1	Chromatography	Peak width variability	seconds	25%ile of peak widths for the wide XICs
XIC_FWHM_Q2	Chromatography	Peak width variability	seconds	50%ile of peak widths for the wide XICs
XIC_FWHM_Q3	Chromatography	Peak width variability	seconds	75%ile of peak widths for the wide XICs
XIC_Height_Q2	Chromatography	Peak height variability		The log ratio for 50%ile of wide XIC heights over 25%ile of heights.
XIC_Height_Q3	Chromatography	Peak height variability		The log ratio for 75%ile of wide XIC heights over 50%ile of heights.
XIC_Height_Q4	Chromatography	Peak height variability		The log ratio for maximum of wide XIC heights over 75%ile of heights.
RT_Duration	Chromatography	Interquartile retention time period	seconds	Highest scan time observed minus the lowest scan time observed
RT_TIC_Q1	Chromatography	Intensity distribution vs. time	Fraction	The interval when the first 25% of TIC accumulates divided by RT-Duration
RT_TIC_Q2	Chromatography	Intensity distribution vs. time	Fraction	The interval when the second 25% of TIC accumulates divided by RT-Duration
RT_TIC_Q3	Chromatography	Intensity distribution vs. time	Fraction	The interval when the third 25% of TIC accumulates divided by RT-Duration
RT_TIC_Q4	Chromatography	Intensity distribution vs. time	Fraction	The interval when the fourth 25% of TIC accumulates divided by RT-Duration
RT_MS_Q1	Chromatography	MS events vs. time	Fraction	The interval for the first 25% of all MS events divided by RT-Duration
RT_MS_Q2	Chromatography	MS events vs. time	Fraction	The interval for the second 25% of all MS events divided by RT-Duration
RT_MS_Q3	Chromatography	MS events vs. time	Fraction	The interval for the third 25% of all MS events divided by RT-Duration
RT_MS_Q4	Chromatography	MS events vs. time	Fraction	The interval for the fourth 25% of all MS events divided by RT-Duration
RT_MSMS_Q1	Chromatography	MS/MS events vs. time	Fraction	The interval for the first 25% of all MS/MS events divided by RT-Duration
RT_MSMS_Q2	Chromatography	MS/MS events vs. time	Fraction	The interval for the second 25% of all MS/MS events divided by RT-Duration
RT_MSMS_Q3	Chromatography	MS/MS events vs. time	Fraction	The interval for the third 25% of all MS/MS events divided by RT-Duration
RT_MSMS_Q4	Chromatography	MS/MS events vs. time	Fraction	The interval for the fourth 25% of all MS/MS events divided by RT-Duration
MS1_TIC_Change_Q2	MS1 Signal	ESI Stability	Ratio	The log ratio for 50%ile of TIC changes over 25%ile of TIC changes
MS1_TIC_Change_Q3	MS1 Signal	ESI Stability	Ratio	The log ratio for 75%ile of TIC changes over 50%ile of TIC changes
MS1_TIC_Change_Q4	MS1 Signal	ESI Stability	Ratio	The log ratio for largest TIC change over 75%ile of TIC changes
MS1_TIC_Q2	MS1 Signal	Dynamic Range	Ratio	The log ratio for 50%ile of TIC over 25%ile of TIC
MS1_TIC_Q3	MS1 Signal	Dynamic Range	Ratio	The log ratio for 75%ile of TIC over 50%ile of TIC
MS1_TIC_Q4	MS1 Signal	Dynamic Range	Ratio	The log ratio for largest TIC over 75%ile TIC
MS1_Count	Acquisition Stats	Spectrum counts	Count	Number of MS spectra collected
MS1_Freq_Max	Acquisition Stats	Acquisition Rate	Hz	Fastest frequency for MS collection in any minute
MS1_Density_Q1	Acquisition Stats	Spectrum counts	Count	25%ile of MS scan peak counts
MS1_Density_Q2	Acquisition Stats	Spectrum counts	Count	50%ile of MS scan peak counts
MS1_Density_Q3	Acquisition Stats	Spectrum counts	Count	75%ile of MS scan peak counts
MS2_Count	Acquisition Stats	Spectrum counts	Count	Number of MS/MS spectra collected
MS2_Freq_Max	Acquisition Stats	Acquisition Rate	Hz	Fastest frequency for MS/MS collection in any minute
MS2_Density_Q1	Acquisition Stats	Spectrum counts	Count	25%ile of MS/MS scan peak counts
MS2_Density_Q2	Acquisition Stats	Spectrum counts	Count	50%ile of MS/MS scan peak counts
MS2_Density_Q3	Acquisition Stats	Spectrum counts	Count	75%ile of MS/MS scan peak counts
MS2_PrecZ_1	Mass Precision	Charge distribution	Fraction	Fraction of MS/MS precursors that are singly charged
MS2_PrecZ_2	Mass Precision	Charge distribution	Fraction	Fraction of MS/MS precursors that are doubly charged
MS2_PrecZ_3	Mass Precision	Charge distribution	Fraction	Fraction of MS/MS precursors that are triply charged
MS2_PrecZ_4	Mass Precision	Charge distribution	Fraction	Fraction of MS/MS precursors that are quadruply charged
MS2_PrecZ_5	Mass Precision	Charge distribution	Fraction	Fraction of MS/MS precursors that are quintuply charged
MS2_PrecZ_more	Mass Precision	Charge distribution	Fraction	Fraction of MS/MS precursors that are charged higher than +5
MS2_PrecZ_likely_1	Mass Precision	Charge distribution	Fraction	Fraction of MS/MS precursors lack known charge but look like 1+
MS2_PrecZ_likely_multi	Mass Precision	Charge distribution	Fraction	Fraction of MS/MS precursors lack known charge but look like 2+ or higher

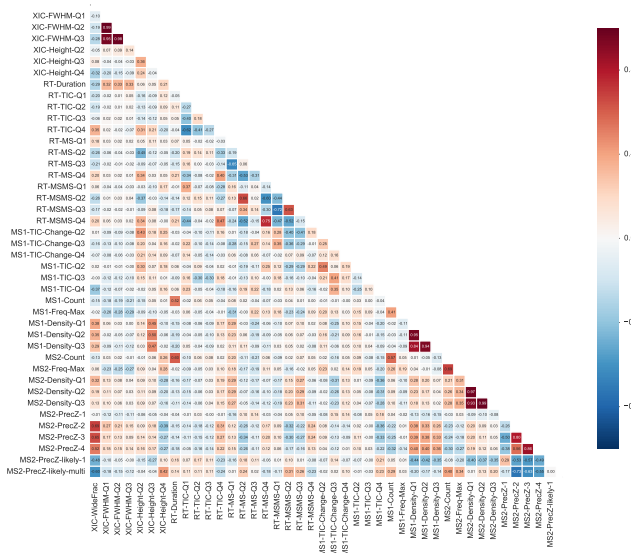
Table 6.2: The ID-free QC metrics computed by QuaMeter [273], as categorized by Amidan et al. [9].



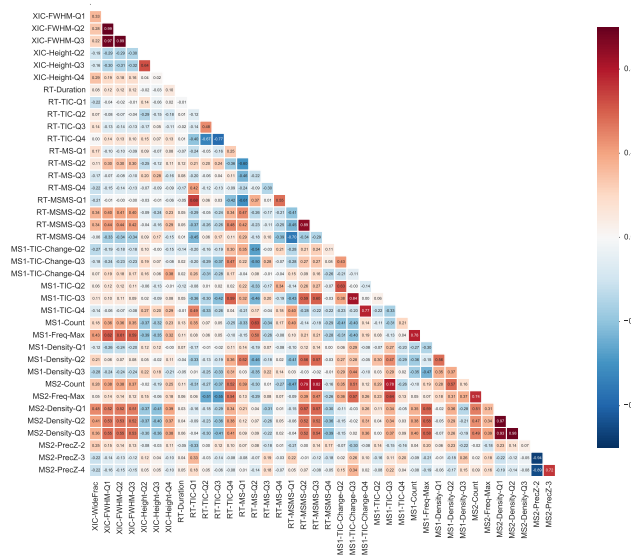
(a) PNNL LTQ-IonTrap



(b) PNNL LTQ-Orbitrap



(c) PNNL Velos-Orbitrap



(d) TCGA

Figure 6.1: Correlation matrices for all four datasets showing the dependencies between the different QC metrics (after the removal of metrics with a low variance).

Additionally, after removing the irrelevant features, the values for each of the metrics separately were scaled by removing the median and scaling according to the interquartile range (IQR). A common approach to scale data is by removing the mean instead of the median, and by scaling to unit variance instead of according to the IQR. However, by using the median and the IQR, the scaling is more robust towards outliers, which can have a large effect on the mean and the standard deviation. Scaling is done to ensure that a limited number of features does not dominate the subsequent analyses, which can be the case if the features have a different scale.

6.2.4 Visualization

Data visualization and visual data exploration is often a crucial first step as it provides an initial approach to understand the data and acts as a driver for the subsequent data analysis [61]. Especially in a high-throughput analytical discipline, such as mass spectrometry proteomics, suitable visualization techniques are paramount to facilitate the understanding and interpretation of the data [203].

Visualizing using a reduced PCA dimensionality

Multiple approaches can be used to visualize quality control metrics. On the one hand each metric can be visualized individually in for example a scatter plot to assess its behavior over time, or for example using a box plot to evaluate the data distribution. However, as has been mentioned before, evaluating each metric individually has certain drawbacks. On the other hand, high-dimensional data cannot be directly visualized. A conventional approach is to first perform a dimensionality reduction to convert the original data to a lower dimensionality, e.g. to two or three dimensions, and then visualize the low-dimensional approximation. A commonly used dimensionality reduction technique is PCA. Very briefly, PCA performs a linear transformation of the original data into its principal components, which are a set of orthogonal variables accounting for the largest possible variance. The data is then reduced by only retaining the first pair of principal components, which can be visualized straightforwardly. A transformation of the QC metrics for the TCGA dataset onto its first two principal components is shown in figure 6.2a, with the first principal component explaining 22.2 % of the total variance and the second principal component explaining 15.6 % of the total variance. Furthermore, the loadings of all metrics for the first two principal components are available in table 6.3. Figure 6.2a indicates that experiments that were performed within a short period of time of each other, as indicated by the color coding, are generally nearby in the principal component space. However, the first two principal components together amount only for about 38 % of the total variance, indicating that other combinations of the metrics can explain different sources of variability, and by increasing the total number of principal components under consideration additional information can be gleaned. Likewise, using the first two principal components to visualize the metrics is bound to show only a part of the underlying information.

Using t-SNE to optimally visualize high-dimensional data

PCA is a technique that is ubiquitously used to achieve a dimensionality reduction, and often visualization efforts do not extend beyond this. However, alternative visualizations can highlight different properties of the data.A powerful recent visualization technique is called t-Distributed Stochastic Neighbor Embedding (t-SNE) [259]. Whereas PCA is a general-purpose dimensionality reduction technique, t-SNE specifically focuses on providing a dimensionality reduction optimized

Metric	PC 1 (22.2 %)	PC 2 (15.6 %)
XIC-WideFrac	0.02260	−0.29913
XIC-FWHM-Q1	−0.00131	−0.26797
XIC-Height-Q2	−0.09709	0.18803
XIC-Height-Q3	−0.02067	0.16791
XIC-Height-Q4	−0.02408	−0.17960
RT-Duration	0.00541	−0.05206
RT-TIC-Q1	0.10094	0.01544
RT-TIC-Q2	0.20395	0.04479
RT-TIC-Q3	0.42124	0.12968
RT-TIC-Q4	−0.40794	−0.10295
RT-MS-Q1	−0.13735	−0.02242
RT-MS-Q2	0.21124	−0.13638
RT-MS-Q3	−0.03770	0.15018
RT-MS-Q4	0.02087	0.03377
RT-MSMS-Q1	0.18828	0.02967
RT-MSMS-Q2	−0.15782	−0.18267
RT-MSMS-Q3	−0.18827	−0.20177
RT-MSMS-Q4	−0.01714	0.14618
MS1-TIC-Change-Q2	−0.23319	0.13199
MS1-TIC-Change-Q3	−0.23875	0.10587
MS1-TIC-Change-Q4	−0.03089	−0.13564
MS1-TIC-Q2	−0.09301	−0.05094
MS1-TIC-Q3	−0.23209	−0.04442
MS1-TIC-Q4	0.03271	−0.01462
MS1-Count	0.16658	−0.24925
MS1-Freq-Max	0.14916	−0.49492
MS1-Density-Q1	−0.10492	0.08909
MS1-Density-Q2	−0.19927	−0.05020
MS1-Density-Q3	−0.21194	0.22557
MS2-Count	−0.23238	−0.17535
MS2-Freq-Max	−0.22226	−0.09512
MS2-Density-Q1	−0.07309	−0.29534
MS2-PrecZ-2	0.02298	−0.11661
MS2-PrecZ-4	−0.03345	0.10291

Table 6.3: The PCA loadings indicate which QC metrics for the TCGA dataset are most relevant for the first two principal components, with the first principal component explaining 22.2 % of the variance and the second principal component explaining 15.6 % of the variance. The loadings with the highest absolute values are highlighted.

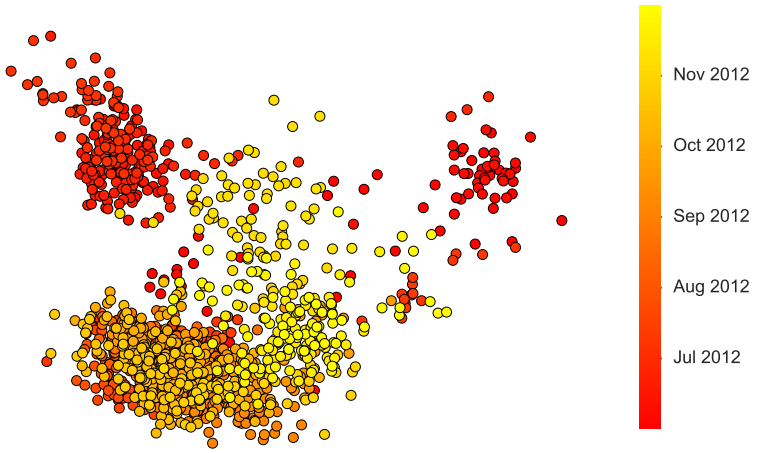
for visualization purposes. Unlike PCA, which is a linear transformation, t-SNE achieves a non-linear transformation to create a mapping from the high-dimensional data to a low-dimensional representation that aims to reveal structure at many different scales. Intuitively, t-SNE creates a low-dimensional representation that tries to model similar objects by nearby points and dissimilar objects by distant points. This is done by converting both the high-dimensional distances and low-dimensional mapped distances between objects to conditional probabilities that represent similarities. Next, it finds the low-dimensional data representation that optimizes the agreement between both probability distributions by minimizing the Kullback-Leibler divergences between the two distributions. Figure 6.2b shows the t-SNE visualization of the QC metrics for the TCGA dataset, which clearly exhibits different data groupings. Moreover, almost all data points clustered together represent experiments that were performed in close succession (as indicated by the color coding), which might indicate batch effects. t-SNE has been shown to be a very powerful data visualization technique for a multitude of data sources [259], and here as well it is able to show underlying similarities between different experiments.

The previous example clearly demonstrates that different visualization techniques can provide complementary views, leading to a more profound understanding of the data characteristics. In this manner, a suitable data visualization can act as a powerful tool to support the explorative data analysis. Furthermore, when visualizing data, seemingly trivial choices can greatly aid interpretability. For example, using a suitable color scheme to add a third dimension of information to a two-dimensional plot can provide valuable additional insights in the data. In figure 6.2 a sequential color scheme was used to represent consecutive dates, which adds clear visual information to discern which experiments were conducted in close succession.

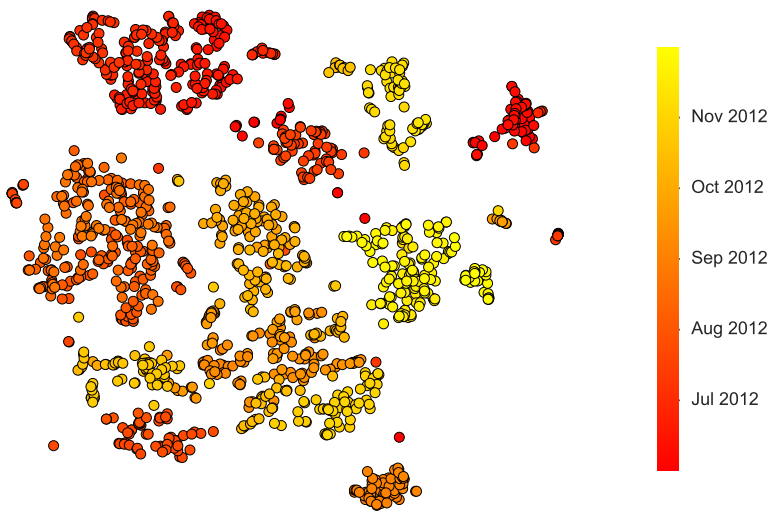
6.3 Quality analysis

As most quality control tools are able to generate (at least) several dozens of metrics, any single experiment can be characterized by multiple QC metrics. Therefore it is often not clear which metrics are most interesting in general, or even which metrics are relevant in a specific situation. The numerous metrics form a multidimensional data space, which results in several challenges during the analysis. For example, the measurement for a single metric might slightly deviate while all other metrics are firmly within the normal range of operation. In such cases, the deviating measurement might simply be due to random fluctuations and not actually due to an abnormal performance. When evaluating the metrics individually, a multiple test correction is an often overlooked necessity to avoid spurious results. Furthermore, as the different stages of a mass spectrometry experiment do not function in isolation, but instead influence each other, likewise some metrics will be correlated. For example, a problem during ionization will lead to different charge state proportions, might influence the number of MS/MS scans, and will have an impact on the number of successfully identified spectra. Again, it is inadequate to only look at a single metric, which might lead to incomplete or even false conclusions. These simple examples illustrate that analyzing each metric individually is often insufficient, and instead multivariate techniques that take into account all metrics simultaneously should be used.

On the other hand, not all multivariate techniques are always applicable. For example, when a multivariate approach using a dimensionality reduction, such as PCA, is applied, part of the data is lost, which can likewise lead to faulty or incomplete results. Furthermore, an additional disadvantage of using PCA-based and related techniques is that the principal components are formed by linear combinations of the original features, which complicates their interpretation. Nevertheless, applying a dimensionality reduction can still be useful in specific situations, for example when combining multiple sources of quality control metrics. Merging multiple sets of QC



(a) Visualization of the QC metrics using the first two principal components.



(b) Visualization of the QC metrics using t-SNE.

Figure 6.2: Various visualization techniques can present different views on the data (TCGA dataset).

metrics will drastically increase the data dimensionality, which can have a profound effect on the subsequent analysis. Namely, for high-dimensional data, various detrimental effects commonly subsumed under the term "curse of dimensionality" pose challenges for algorithms that make use of the full feature space [292]. Therefore, in some cases applying a prior dimensionality reduction, or using algorithms that are optimized for a high-dimensional search space, might achieve a superior performance. Because in our analysis we have limited the QC metrics to only the ID-free metrics that are computed by QuaMeter, the dimensionality of the dataset is not overly large to impede taking all features into account. Furthermore, using the full dimensionality has certain advantages, such as the availability of more information compared to when a dimensionality reduction technique would be applied.

We will next show how optimized techniques that take into account the multidimensional feature space can be used to assess the quality of MS experiments and to discriminate low-quality from high-quality experiments.

6.3.1 Outlier detection

Outlier detection can be used to detect deviating experiments with a low performance or a high level of (unexplained) variability. These outlying experiments can subsequently be analyzed to discover the source of the reduced performance to enhance the quality of future experiments. Additionally, outlier detection can be a vital step to remove invalid measurements ahead of further processing, such as, for example, sample identification, to ensure that these low-quality experiments do not unduly influence the output results.

Local outlier detection

We have used the Local Outlier Probability (LoOP) [146] algorithm to detect outlying experiments, as it has a few beneficial properties. Most importantly, LoOP identifies outliers based on their local neighborhood [39]. This approach is more sensitive than global outlier detection methods, as outliers are identified based on the density of the neighbors in their immediate vicinity, as opposed to the global data distribution. For example, when analyzing a sizable number of experiments performed over an extended time period, as we have done, it is conceivable that there will occur small environmental changes over time, which will have an influence on the experimental results. Because these effects might be more or less pronounced at certain times, this prohibits the use of a single global outlier measure. Instead, when the outlier measure is restricted to the local neighborhood, outliers will be identified based on (excessive) differences with their closest matching experiments. Another advantage is that the LoOP outlier scores are normalized and can be expressed as a probability, whereas most other outlier detection algorithms report scores with an arbitrary scale, with scores often incomparable between different datasets or different parameter values, even when using the same algorithm. LoOP, however, consistently uses probabilities, which ensures that these outlier scores can readily be compared and straightforwardly be interpreted [146].

Figure 6.3 shows a histogram of the outlier scores that have been assigned to all of the experiments in the PNNL LTQ-IonTrap dataset. As can be seen, most experiments have a (relatively) low outlier score, with the bulk of the experiments having a score close to 0 %. Other experiments have a higher outlier score, with some of them being marked as clear outliers. As each experiment has been assigned a numeric score, this enables ranking the various experiments by their gradation of being an outlier. Furthermore, generally a score threshold is set to distinguish outliers from non-outliers. Although setting such a score threshold can sometimes be quite subjective, there are

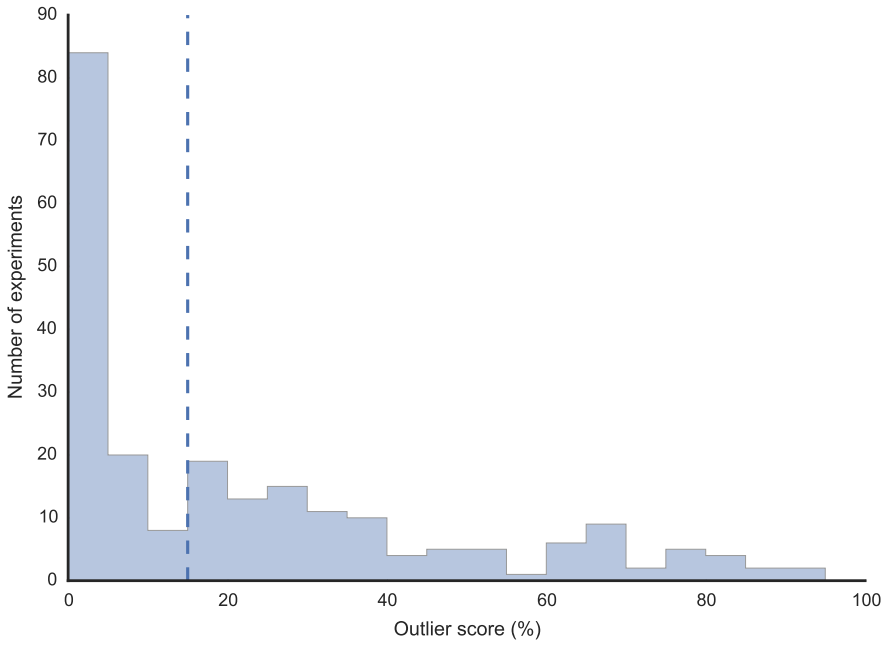


Figure 6.3: Histogram of the LoOP outlier scores for the PNNL LTQ-IonTrap dataset. The score threshold of 15 % is indicated by the dashed line.

a few considerations that can be taken into account. First, we expect that most of the experiments are non-outliers with a score close to zero, while outliers have a higher score over a wider range. Second, we aim to have a high sensitivity to detect most if not all of the low-quality experiments, so the threshold should be set quite conservatively to ensure that all the low-quality experiments are detected, at the expense of some false positives. As the outlier detection strategy is an unsupervised method some number of false positives is unavoidable, which should be filtered out in a subsequent manual evaluation step. Keeping in mind these considerations, for example for the outlier histogram in figure 6.3 a good choice for the score threshold would be 15 %.

Outlier validation

Because outlier detection is an unsupervised method it is not straightforward to validate the results when using real-life datasets, as the ground truth is often unknown. However, for the PNNL data the quality of the experiments was assessed by expert instrument operators, whose labeling can be used as the ground truth for validating the detected outliers, as each experiment was assessed as having either “good”, “ok”, or “poor” quality.

Using these quality assignments, the obtained outlier scores can be validated. Figure 6.4 shows the receiver operator characteristic (ROC) curves for the PNNL datasets. The blue ROC curves show the outlier detection performance when considering the “good” and “ok” experiments as the positive class, i.e., the acceptable experiments, and the “poor” experiments as the negative class, i.e., the unacceptable, low-quality experiments. Additionally, the green curves show the performance when considering only the “good” experiments as the positive class, while the red curves show the performance when considering only the “ok” experiments as the positive class.

These ROC curves clearly indicate that the outlier detection technique successfully manages to discriminate high-quality experiments from low-quality outlying experiments. Furthermore, they show that there is an optimal distinction between the experiments with the highest quality (labeled "good") and the low-quality experiments (labeled "poor"). Meanwhile, experiments with a slightly diminished but still sufficient quality (labeled "ok") can still be discriminated quite successfully from the low-quality experiments.

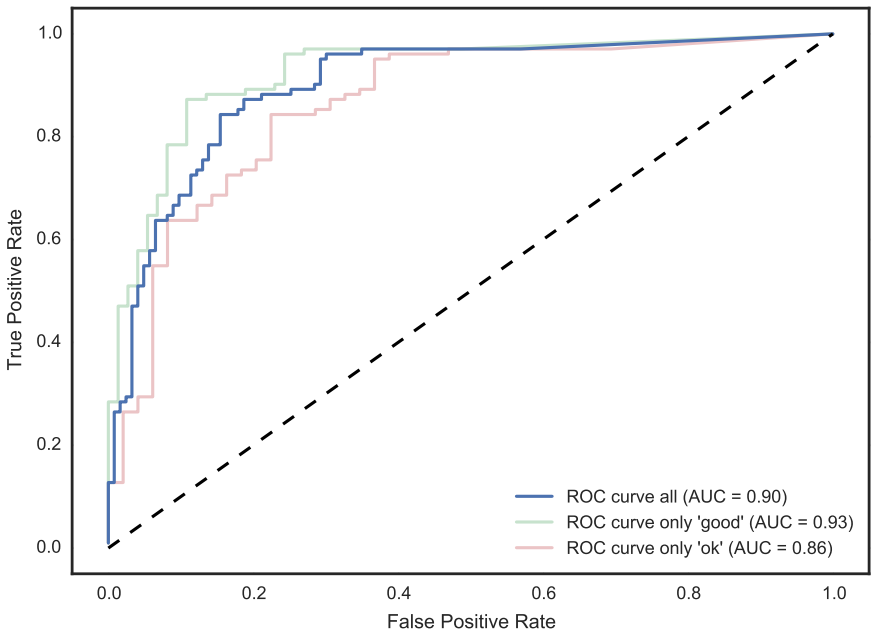
This is also indicated by figure 6.5, which shows the density of the outlier scores for the various quality labels. It shows that the high-quality experiments indeed have a very low outlier score, while the low-quality experiments have a higher outlier score. As illustrated by figure 6.5a, for the case of the PNNL LTQ-IonTrap dataset our previous assumptions stated to determine a score threshold when no validation information is available hold true, and the choice of 15 % as threshold provides an adequate separation of high-quality and low-quality experiments.

Note that, although the previous results indicate that a very good performance detecting the low-quality experiments is achieved, we are still outperformed by the original results by Amidan et al. [9]. However, contrary to their approach, which entails a supervised classifier, our approach is fully unsupervised. For a supervised approach training data is required, which in this case means that the quality of a considerable number of experiments needs to be annotated manually to provide the ground truth. On the other hand, our approach does not require a training phase, but can be applied directly on a set of experiments of unknown quality. Therefore, the time-consuming manual curation can be forgone, while still being able to successfully identify low-quality experiments. Furthermore, a supervised classifier has to be retrained for different instruments or even for different operating procedures on the same instrument, with each situation again incurring the need for manual curation to provide training data. Contrary to this, our outlier detection strategy can directly be applied to diverse datasets with widely varying characteristics, as evidenced by the consistent performance across data generated on different instrument types. However, because an unsupervised technique will inherently result in more false positives than a supervised technique, our outlier detection strategy is mostly suited as a filtering step to quickly provide an initial discrimination between high-quality and low-quality experiments. By conservatively setting the outlier score threshold, the experiments marked as outliers can subsequently be inspected manually in full detail to exclude any false positives. Because our outlier detection strategy is already quite sensitive despite being an unsupervised technique, the effort required for the manual evaluation will be significantly reduced. Furthermore, by using the local neighborhood to determine whether an experiment is an outlier, optionally dissimilar data sources can even be combined while still achieving a similar performance (data not shown). Finally, our approach requires only a few simple and intuitive parameters, which can easily be understood and set, as is detailed next.

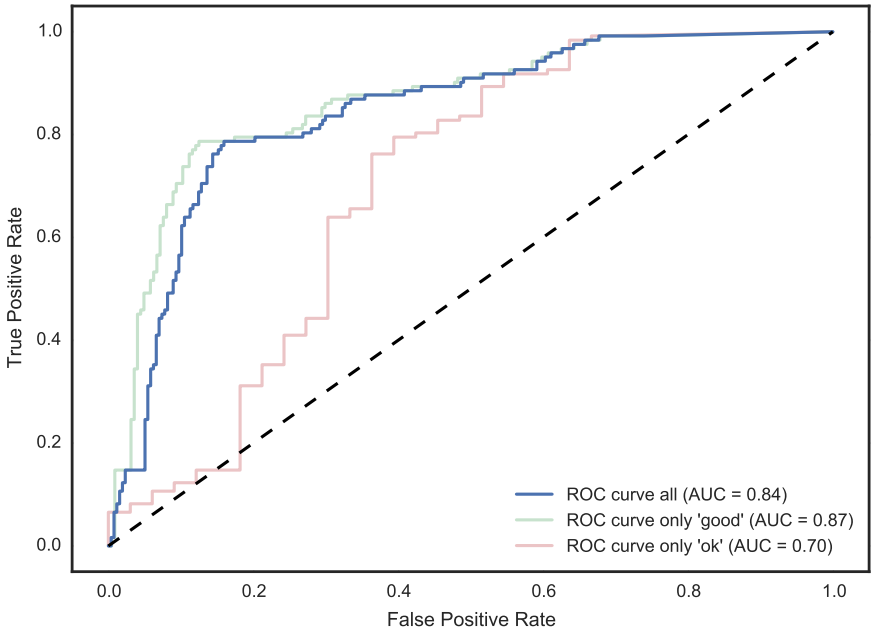
Parameter configuration

A common problem when applying data mining techniques is that advanced algorithms often depend on multiple parameters that have to be manually set and can be hard to understand. An important advantage of the presented outlier detection strategy is that there are only a limited number of parameters which can easily be understood intuitively. The outlier detection strategy mainly depends on two different parameters: the size of the local neighborhood, and the outlier score threshold. Based on the annotated PNNL data we can determine the optimal values for these parameters.

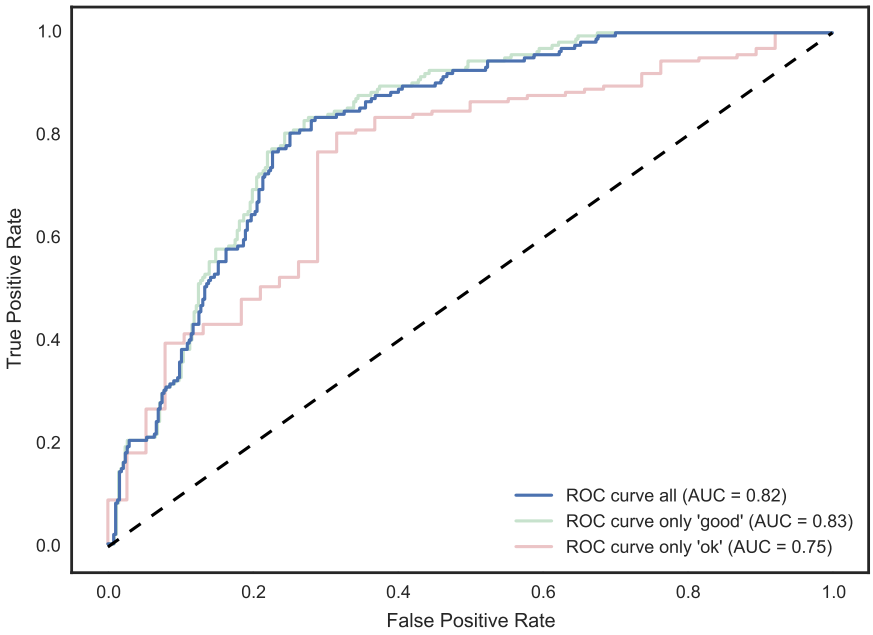
We have previously presented the results obtained by the outlier detection technique using ROC curves. The earlier presented PNNL curves are those with the highest area under the curve (AUC) for varying values of the local neighborhood size. By considering more or fewer neighboring



(a) PNNL LTQ-IonTrap



(b) PNNL LTQ-Orbitrap

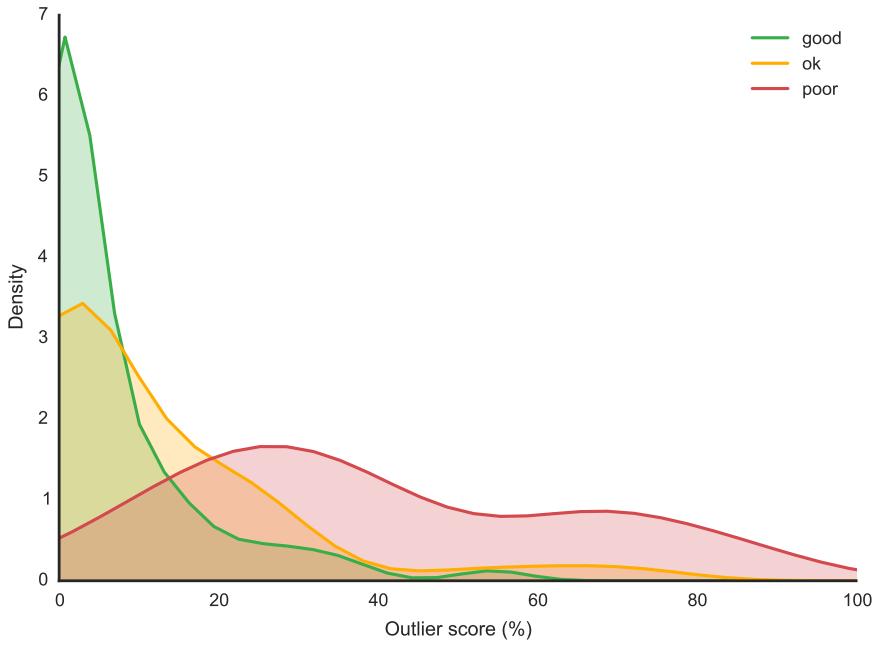


(c) PNNL Velos-Orbitrap

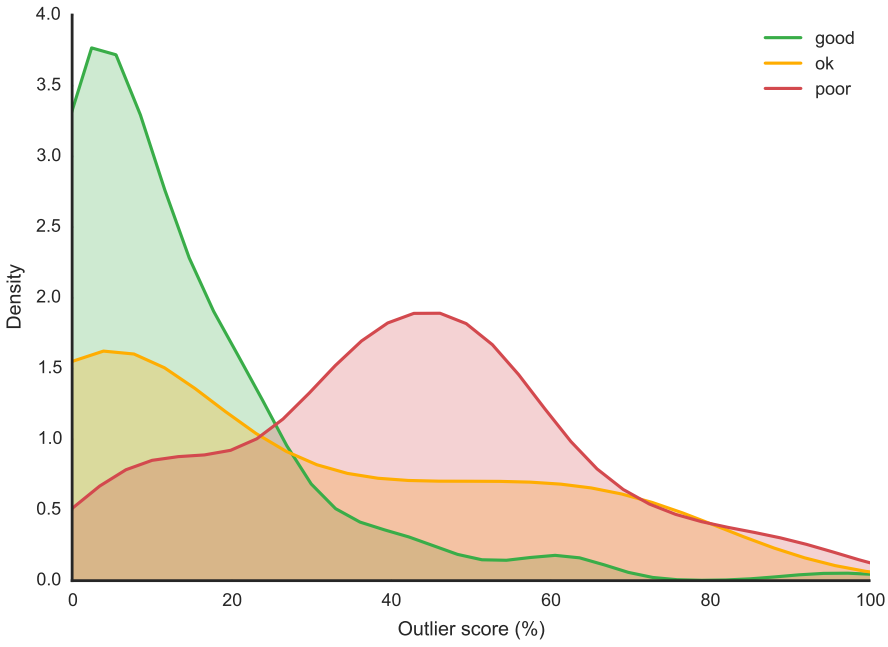
Figure 6.4: ROC curves show the performance of the outlier detection strategy for the PNNL datasets.

experiments to determine the outlier score for each experiment, we can find the optimal size of the local neighborhood to achieve the best discrimination between low-quality and high-quality experiments. As can be seen in figure 6.6, maximal AUCs are achieved when taking a considerable number of nearest neighbors in account. This is because most good experiments do not differ much from each other, so they all have a lot of similar experiments in their local neighborhood. On the other hand, some outlying experiments might share a few close neighbors, but when increasing the neighborhood size the outlying experiments will clearly start to differ from the bulk of experiments. Figure 6.6 indicates that the AUC rises for increasing sizes of the local neighborhood, and levels off after the optimal neighborhood size. In general, optimal AUCs are achieved for local neighborhoods consisting of 50 to 60 experiments, more or less irrespective of the total number of experiments. Therefore, it is recommended to consider larger rather than smaller neighborhoods when performing outlier detection. This also hints at a limitation of our technique: a sufficient number of observations needs to be available in order to establish optimally discriminating local neighborhoods. Therefore, our approach is mostly suited when a higher number of experiments is available, for example during a longitudinal analysis, as opposed to within a single study.

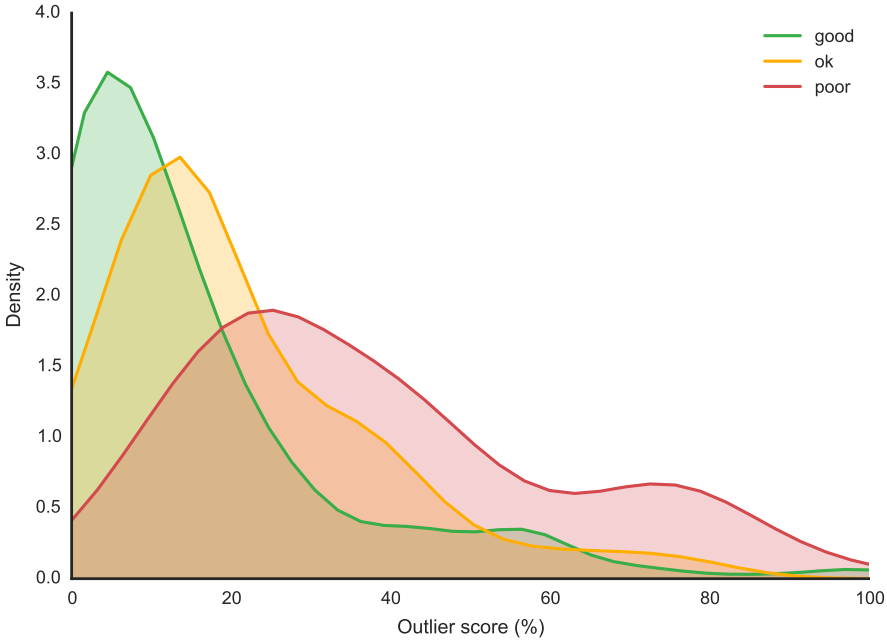
The other significant parameter is the outlier score threshold, which is used as a cut-off value to separate the experiments with an acceptable outlier score from those with an excessive outlier score. The advantage of the LoOP outlier scores is that these scores are normalized between 0 and 1, and are comparable between different executions. To evaluate the outlier score threshold we can make a trade-off between the sensitivity (true positive rate) and the specificity (true negative rate, or 1 − false positive rate), which can be derived from the axes of the ROC curves in figure 6.4. Because it is most important to successfully identify all the low-quality experiments,



(a) PNNL LTQ-IonTrap



(b) PNNL LTQ-Orbitrap



(c) PNNL Velos-Orbitrap

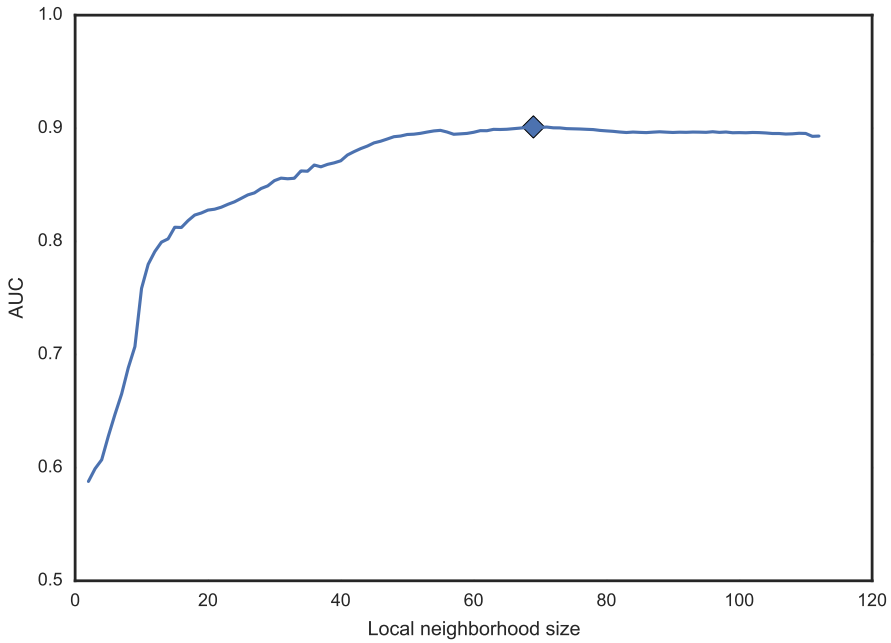
Figure 6.5: Outlier score densities for the various quality levels for the PNNL datasets.

at the expense of a few misclassified high-quality experiments that might have to be checked manually, we prefer a high sensitivity. The ROC curves in figure 6.4 and figure 6.7, which shows the sensitivity and specificity compared to the outlier score threshold, can provide a guideline to determine a suitable, conservative, value for the outlier score threshold.

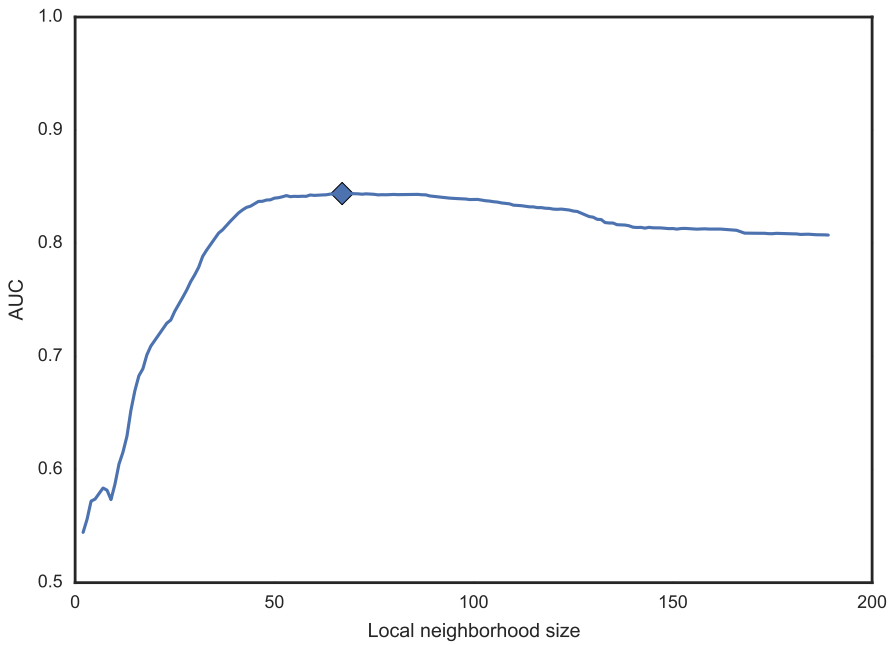
6.3.2 Outlier interpretation

Although we have shown that we can successfully differentiate low-quality from high-quality experiments, it is insufficient to only know that a specific experiment is an outlier, it is also of vital importance to know why the experiment is an outlier. For this purpose the outlier score is only useful to a limited extent: it indicates how significantly an experiment is an outlier, but it does not provide an explanation as to what causes the experiment to be an outlier. To provide an explanation why an experiment is an outlier, the subspace in which the outlying experiment can be differentiated from the other experiments can be used [186]. Here a subspace is formed by one or more attributes, which correspond to the various QC metrics. Thus, this subspace can be used to interpret the outlier by indicating which QC metrics best explain the outlying behavior. The relevant subspaces for an outlier can be used by domain experts to increase interpretability and investigate the performance of the experiment.

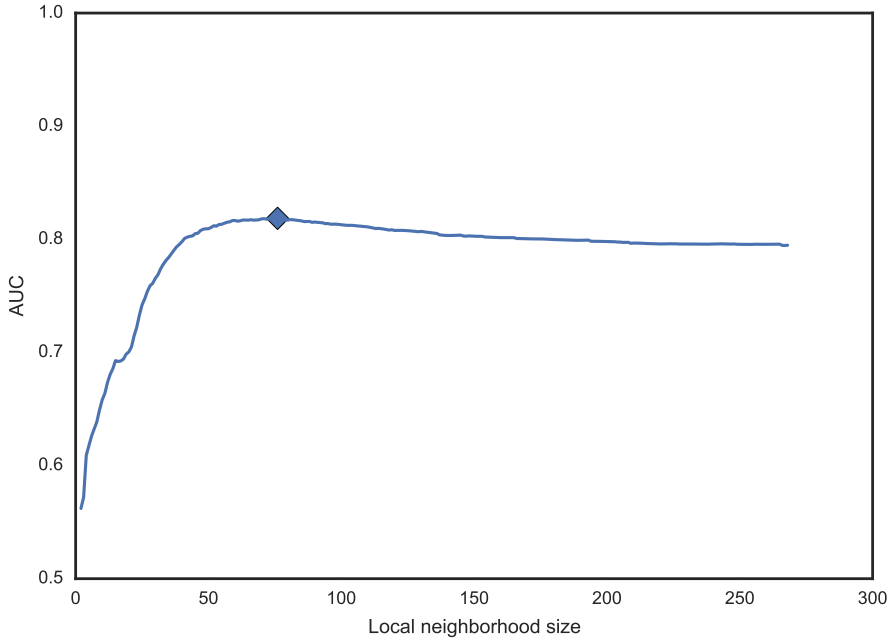
For multiple experiments, each with specific values for the various QC metrics, the experiments can be considered as the observations and the QC metrics as the attributes for each observation. To interpret an outlier, the subspace of attributes for which the outlier shows separability from the inliers is identified based on a procedure detailed by Micenková et al. [186]. The relevant subspace for each outlier is determined by performing a classification between the outlier and



(a) PNNL LTQ-IonTrap



(b) PNNL LTQ-Orbitrap



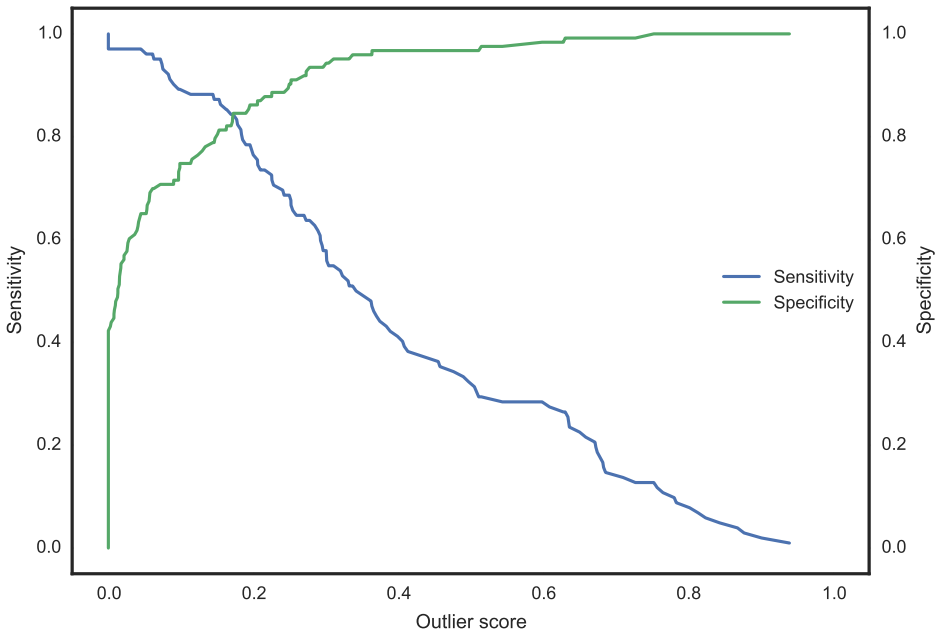
(c) PNNL Velos-Orbitrap

Figure 6.6: ROC curve AUCs in function of the size of the local neighborhood during outlier detection. The maximal obtained AUC value is highlighted by the diamond marker.

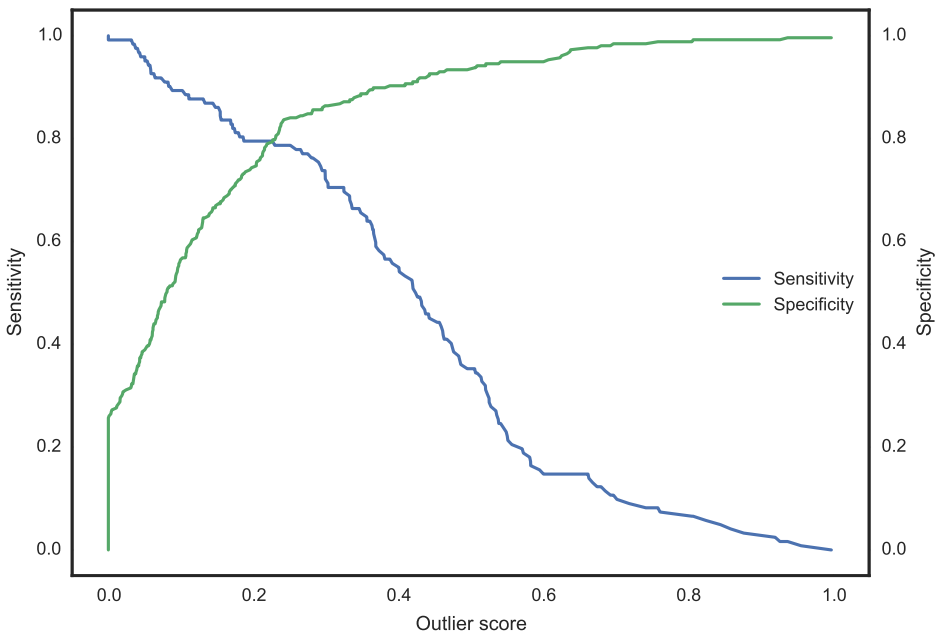
the inliers. Because a single outlying observation cannot be correctly classified, a balanced classification problem is created by supersampling the outlier and subsampling the inliers. The outlier is supersampled by generating multiple points from a multivariate normal distribution centered around the outlier. The subsampling of the inliers consists of two elements: a specified number of nearest neighbors to the outlier are selected, and additional points are randomly selected from among the other inliers.

Next, a classification approach can be used to differentiate between the (equal-sized) sets of simulated outliers and the selected inliers. After the classification, the relevant features for which the set of outliers is most separable from the inliers can be extracted, which forms the subspace explaining the outlier. We have employed a random forest classification approach to differentiate the outlier from the inliers. The advantage of a random forest in terms of feature importance is that it is very simple to determine the weight of each feature by considering the number of times the feature in question was used during the classification. This results in a straightforward numerical measure of importance for each feature which can be easily interpreted. To minimize the random effects of supersampling the outlier, subsampling the inliers, and performing the random forest classification, this procedure can be repeated multiple times, after which the averages of the feature weights are used to determine the final feature importances.

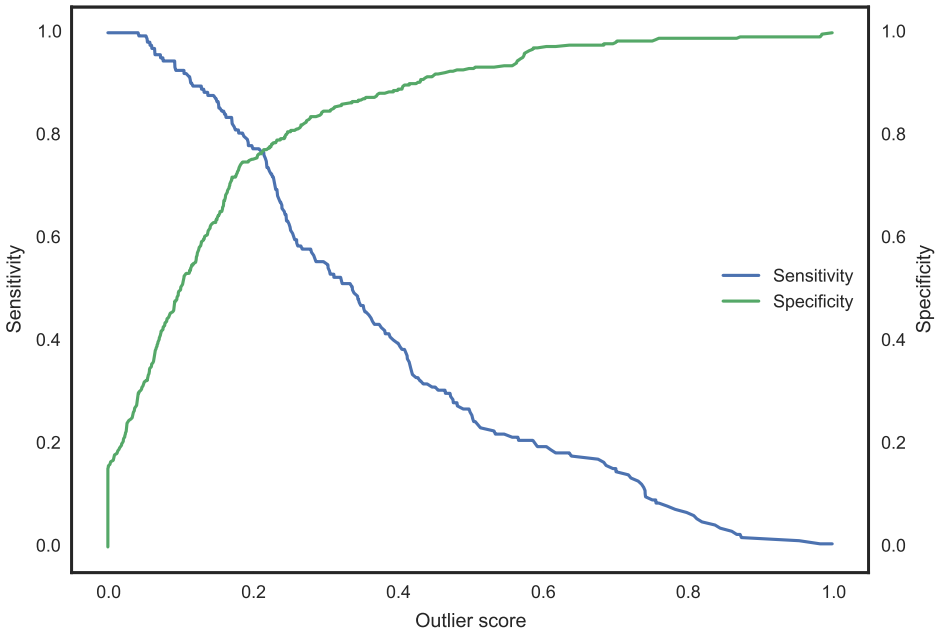
After the feature importances have been computed, the relevant subspace is extracted based on a few rules of thumb. Specifically, features are selected in descending feature importance order until the cumulative feature importance explains at least 50 % of the total importance, or until the importance of the additional feature to select is less than two-thirds of the largest feature importance. These simple rules of thumb mostly result in the selection of intuitive subspaces



(a) PNNL LTQ-IonTrap



(b) PNNL LTQ-Orbitrap



(c) PNNL Velos-Orbitrap

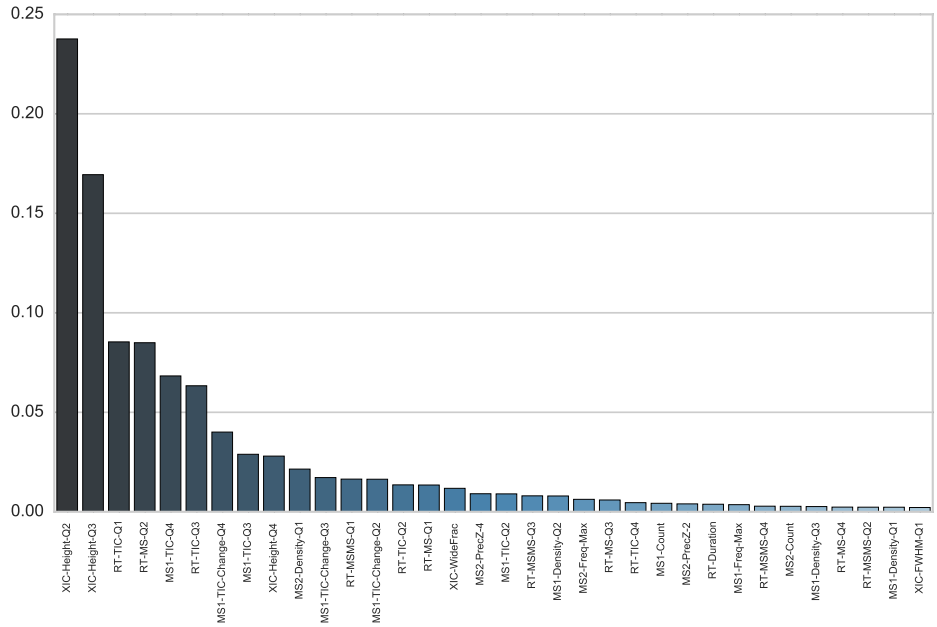
Figure 6.7: Outlier score threshold values required to obtain a specific sensitivity and specificity for the PNNL datasets.

consisting of the most important features while avoiding the inclusion of unnecessary features.

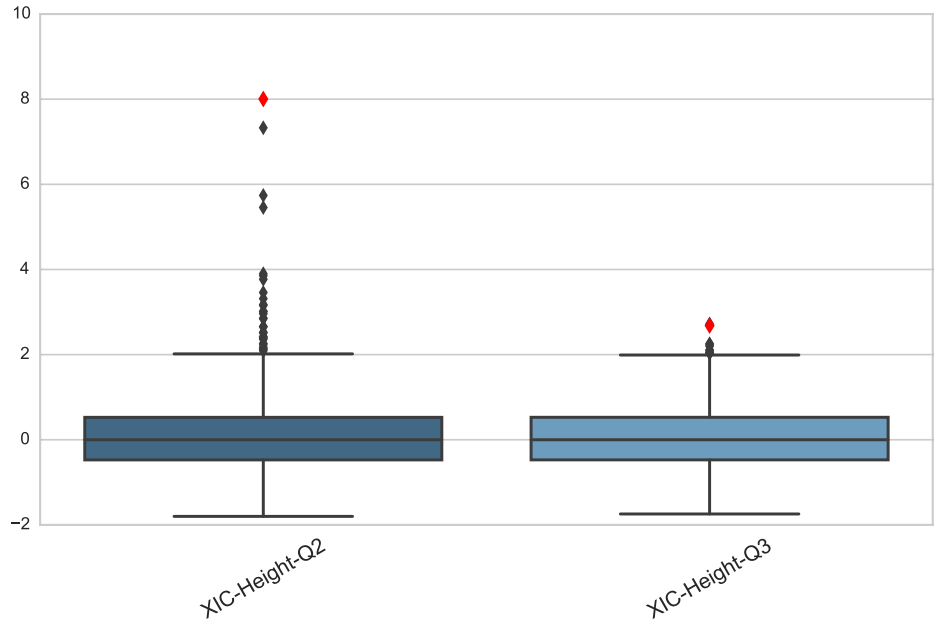
Although the data consists of dozens of features, using the above approach provides an explanation for the outliers using only a few relevant features (for the TCGA dataset between 1 and 7 features are used per outlier, with an average of 2.3 features). This limited subspace dimensionality hugely facilitates outlier interpretability compared to the full high-dimensional feature set, and pinpoints a few highly relevant features for closer examination by domain experts.

Figure 6.8 shows an example of interpreting a specific outlying experiment from the TCGA dataset. As detailed previously, to retrieve the relevant subspace for an outlier, first the feature importances for the various QC metrics are computed, as is shown in figure 6.8a. Next, the subspace formed by the relevant QC metrics is extracted, as is shown in figure 6.8b. The feature subspace explaining the outlier can be interpreted by domain experts, and can provide insights in relationships between various QC metrics. For example, figure 6.8b shows that this particular outlier exhibited an exceptionally high variance in peak heights, which may indicate problems with the chromatography.

This example also shows the advantage of using a multivariate approach, instead of only looking at single QC metrics individually. In figure 6.8 this advantage is somewhat less pronounced, as the values in the outlier's subspace are individually both already significant outliers. However, the interplay of various metrics can be crucial to detect outliers, as is shown in figure 6.9. Here, the values in the outlier's subspace are well within 1.5 times the interquartile range, as denoted by the whiskers of the box plot, so this outlier cannot be detected using univariate techniques. However, by comparing against the local neighborhood and by analyzing all metrics simultaneously the



(a) The feature importances for the various QC metrics to discriminate the experiment from the non-outlying experiments.



(b) The subspace formed by the most relevant QC metrics. The outlier's values in the subspace are highlighted in red.

Figure 6.8: Feature importances and subspace for experiment 'TCGA-AA-A02O-01A-23_W_VU_20130206_A0218_10A_R_FR07', which has an outlier score of 98.06 %.

aberrant proportion of the metrics still results in a high outlier score. Meanwhile, prominent outliers for a single metric will still be detected as well, with explanatory subspaces consisting of only this single metric, as is shown in figure 6.10.

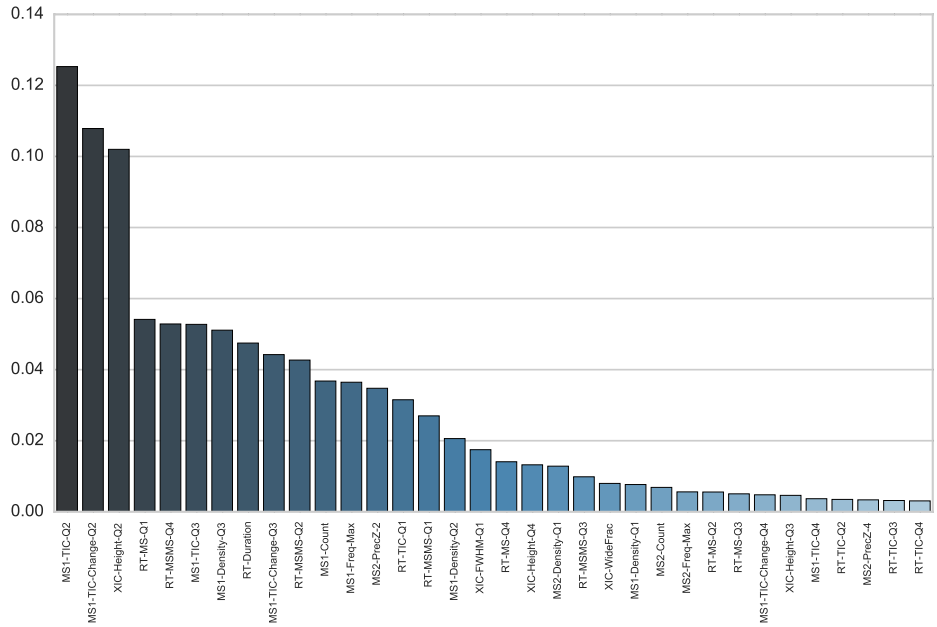
Furthermore, it is worth highlighting that taking the full set of metrics into account while detecting outliers has advantages over multivariate approaches where a dimensionality reduction technique is used as well because all metrics are taken into account, instead of only a lower-dimensional approximation. For example, when applying PCA, only the most significant principal components are retained to achieve a dimensionality reduction. In this case, only the metrics with the highest variance contribute significantly to the first few principal components. Therefore, using such an approach, a prominent outlier such as shown in figure 6.8 cannot be detected successfully because the metrics describing the peak heights only have a limited contribution to the first two principal components, as can be verified in table 6.3.

Frequent outlier subspaces

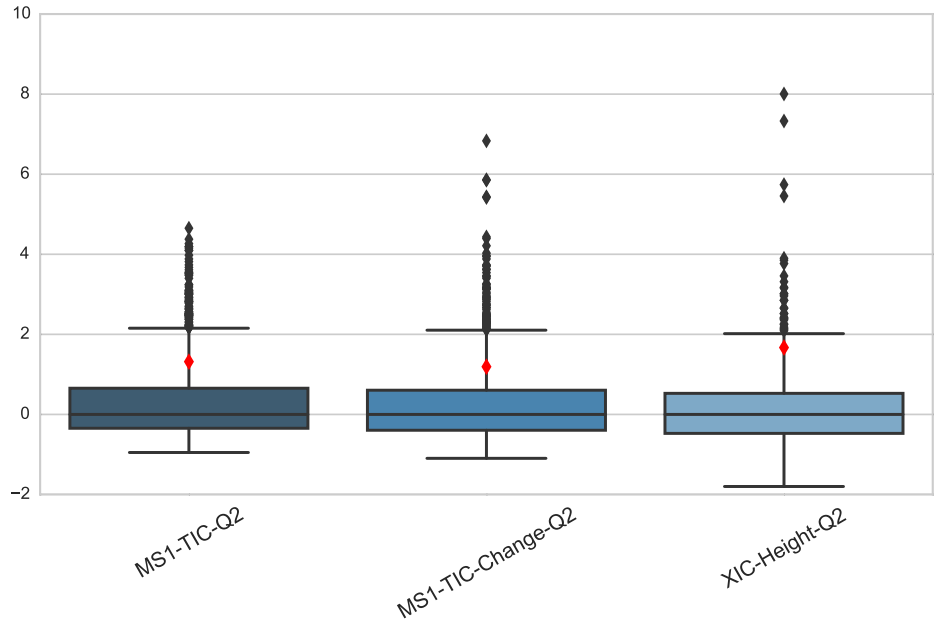
Next, by combining the explanatory subspaces for all individual outliers, it is possible to get a general view on which QC metrics are most relevant when detecting deviating experiments. Most outliers can be interpreted by a limited number of QC metrics: for the TCGA dataset, when setting the outlier score threshold at a conservative 25 % which results in 508 outliers (figure 6.11), on average each subspace consists of about 2.3 QC metrics, with a minimum of one metric and a maximum of seven metrics. By considering each outlying experiment as a transaction and the QC metrics that form the subspace of the experiment as items in the transaction, frequent itemset mining [193] can be applied to detect QC metrics that often co-occur in the outliers' subspaces. Table 6.4 shows the QC metrics that form frequent itemsets with a minimum support of 5 %, i.e., the QC metrics are present in the subspace of at least 5 % of all outliers. Here, the higher the support, the more often QC metrics co-occur as important explanatory variables in the outliers' subspaces. Table 6.4 indicates that some QC metrics are more useful than others to detect outliers, as they are present more often. Furthermore, some sets of multiple metrics occur often as well, such as in the case of the itemset consisting of MS1-TIC-Q4 and MS1-TIC-Change-Q4, which has a total support of 6 %. Indeed, it seems logical that these two metrics are related, as excessive changes in the total ion current (TIC) between the third and the fourth quartile (MS1-TIC-Change-Q4) will influence the total amount of TIC near the end of the experiment (MS1-TIC-Q4). Other pairs of co-occurring metrics were observed as well, such as the combination of metrics XIC-Height-Q2 and XIC-Height-Q3, concerning the chromatographic peak height, and metrics MS2-PrecZ-2 and MS2-PrecZ-4, concerning the precursor charge states, although these and other combinations have a slightly lower support value and are not included in table 6.4.

Spectral identification performance

We were able to show the accuracy of the outlier detection method using manually curated datasets. However, in general such an approach is not possible because it takes too much effort to manually assess the quality of a large collection of experiments. Therefore, the number of peptide-spectrum matches (PSMs) is often used as a stand-in quality measure, as one would expect that the low-quality experiments result in a lower number of identified spectra due to a diminished performance. When comparing the number of valid PSMs between the outlying experiments and the non-outlying experiments for the TCGA dataset, the latter result in a slightly higher number of PSMs, as is shown in figure 6.12a, although the difference is not very pronounced. The outliers seem to contain both experiments that have a lower number of PSMs, and experiments that have an average or even an above-average number of PSMs. This observation confirms prior findings

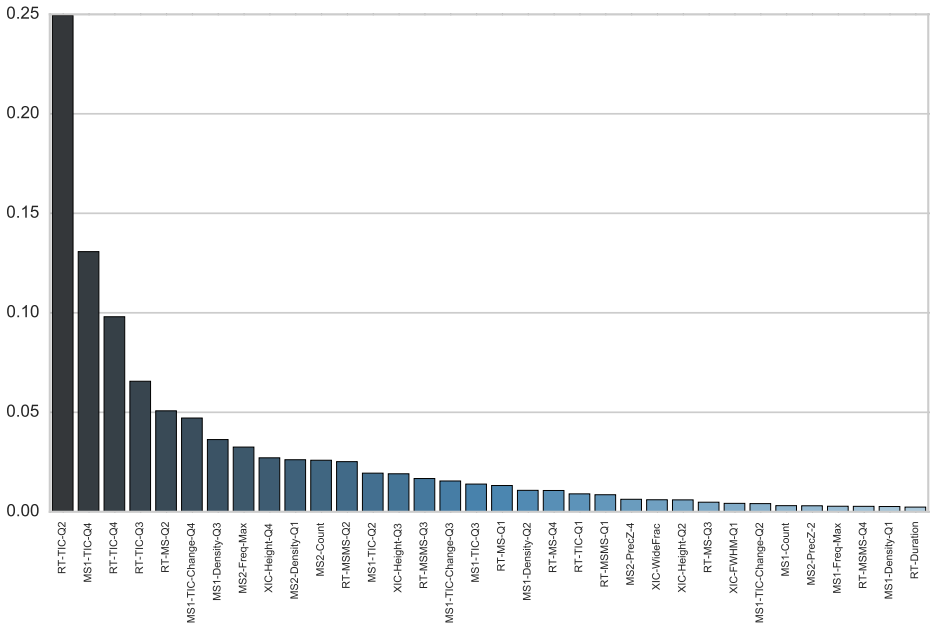


(a) The feature importances for the various QC metrics to discriminate the experiment from the non-outlying experiments.

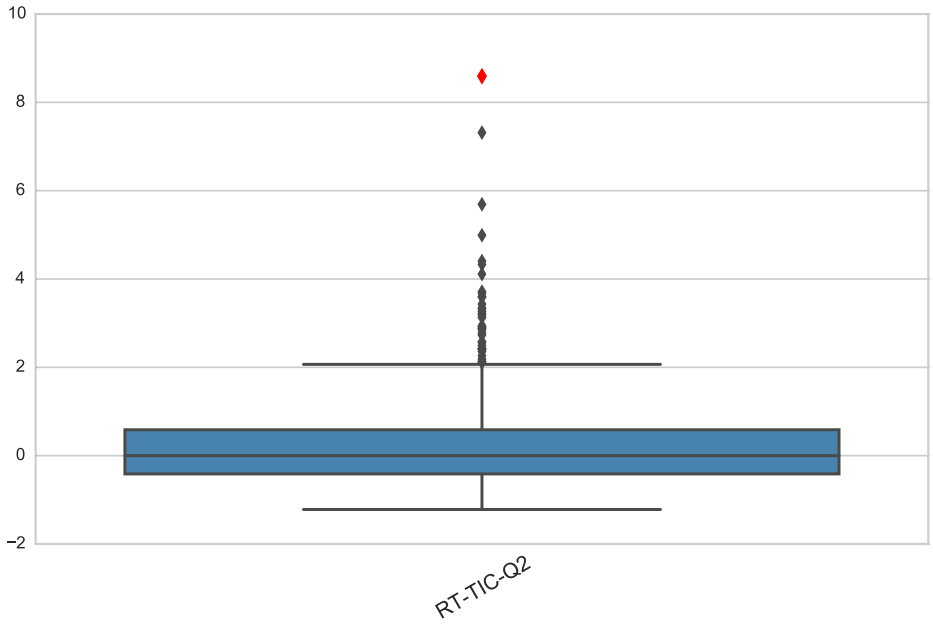


(b) The subspace formed by the most relevant QC metrics. The outlier's values in the subspace are highlighted in red.

Figure 6.9: Feature importances and subspace for experiment 'TCGA-AA-A01X-01A-23_W_VU_20120807_A0218_1H_R_FR03', which has an outlier score of 75.68 %.



(a) The feature importances for the various QC metrics to discriminate the experiment from the non-outlying experiments.



(b) The subspace formed by the most relevant QC metrics. The outlier's value in the subspace is highlighted in red.

Figure 6.10: Feature importances and subspace for experiment 'TCGA-AG-A016-01A-23_W_VU_20120731_A0218_1D_R_FR12', which has an outlier score of 92.37 %.

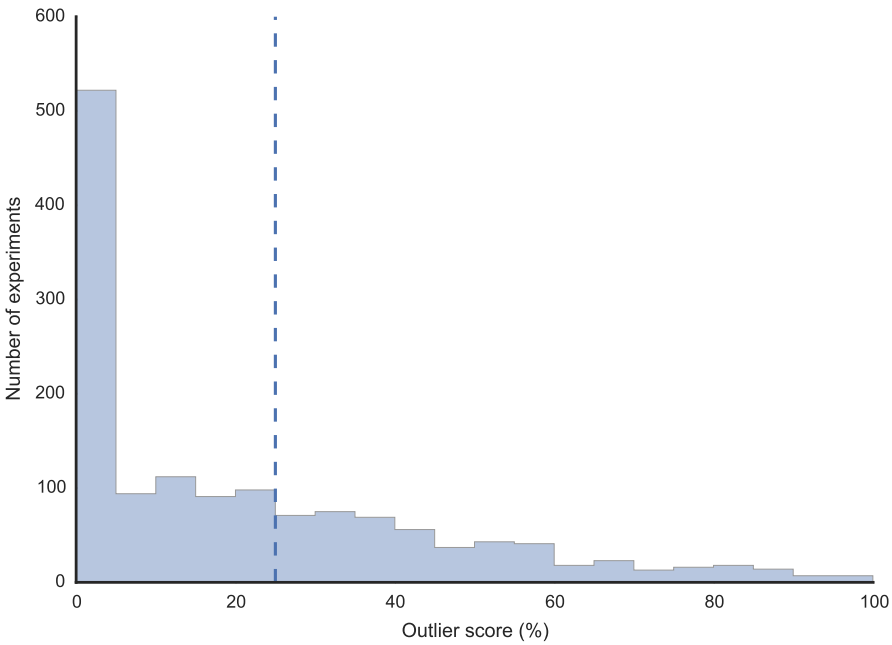


Figure 6.11: Histogram of the LoOP outlier scores for the TCGA dataset. The score threshold of 25 % is indicated by the dashed line.

Outlier subspace QC metric(s)	Support (%)
MS1-TIC-Change-Q4	16
XIC-Height-Q4	15
RT-Duration	14
RT-TIC-Q3	14
XIC-Height-Q2	12
XIC-Height-Q3	11
XIC-WideFrac	11
MS1-TIC-Q4	10
MS1-TIC-Change-Q2	9
RT-TIC-Q2	8
MS2-PrecZ-4	8
MS1-TIC-Q2	8
RT-TIC-Q4	7
MS2-PrecZ-2	7
RT-MSMS-Q4	7
MS1-TIC-Q4, MS1-TIC-Change-Q4	6
MS1-Freq-Max	6
RT-MS-Q3	6
RT-MSMS-Q3	6
RT-MSMS-Q1	5
MS1-Density-Q3	5

Table 6.4: Overview of the QC metrics that frequently occur in the outliers' explanatory subspaces for the TCGA dataset. Exact support values can differ slightly between separate executions due to some variable effects while computing the explanatory subspaces.

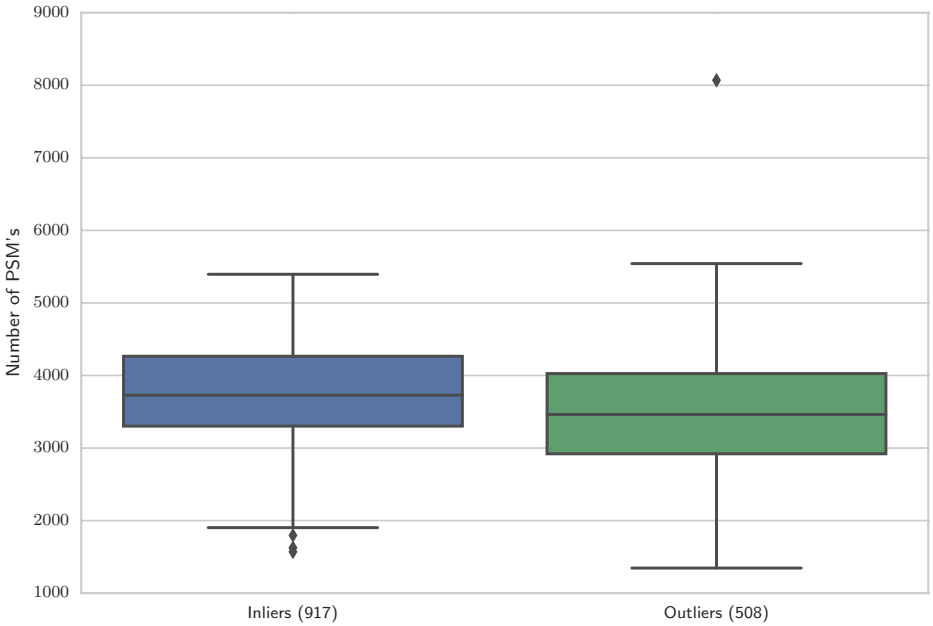
Metric(s)	Support (%)	p-value
MS1-TIC-Change-Q4	16	0.01851
XIC-Height-Q4	15	0.00002
RT-Duration	14	0.08282
RT-TIC-Q3	14	0.00001
XIC-Height-Q2	12	0.02115
XIC-Height-Q3	11	0.00048
XIC-WideFrac	11	0.00000
MS1-TIC-Q4	10	0.91656
MS1-TIC-Change-Q2	9	0.00582
RT-TIC-Q2	8	0.00008
MS2-PrecZ-4	8	0.00144
MS1-TIC-Q2	8	0.04482
RT-TIC-Q4	7	0.00000
MS2-PrecZ-2	7	0.00000
RT-MSMS-Q4	7	0.65808
MS1-TIC-Change-Q4, MS1-TIC-Q4	6	0.17899
MS1-Freq-Max	6	0.07700
RT-MS-Q3	6	0.27253
RT-MSMS-Q3	6	0.59691
RT-MSMS-Q1	5	0.02163
MS1-Density-Q3	5	0.02578

Table 6.5: Evaluation of the identification performance for the most frequent QC metrics in the outlier subspaces for the TCGA dataset. The p-value is computed by a two-tailed t-test and indicates whether the number of PSM's for the outlying experiments for each metric is dissimilar from the number of PSMs for the non-outlying experiments, with p-values lower than 0.05 highlighted.

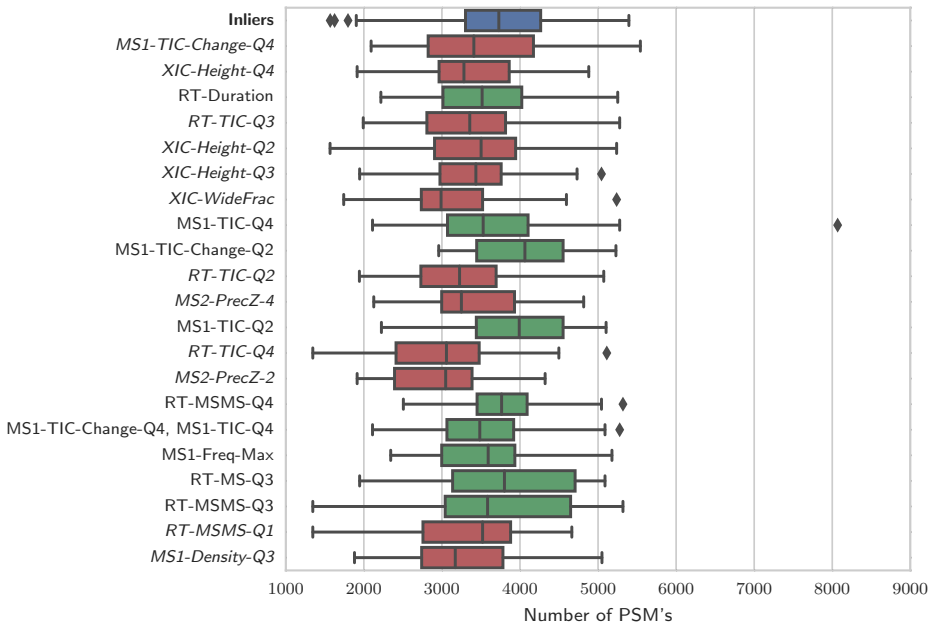
that outliers for this dataset do not necessarily arise due to sources impeding successful spectrum identifications, but can possibly be attributed to a significant biological diversity between the various samples [235].

However, when the explanatory subspaces for the outliers are taken into account, a distinction between several of the outliers can be made. As can be seen in figure 6.12b, for some specific QC metrics the number of PSMs for the outliers is notably lower than for the non-outlying experiments. Conversely, for a few other QC metrics the number of PSMs for the outliers is very similar to those of the non-outlying experiments. Table 6.5 confirms the difference in terms of PSMs between the non-outlying experiments and the outlying experiments for each of the frequently occurring QC metrics by computing a t-test with null hypothesis that they have identical expected values. The reported p-values show that this null hypothesis can be safely rejected in some instances, with significantly lower numbers of PSMs for some sets of outliers, which have also been highlighted in figure 6.12b. We can hypothesize that the QC metrics for which the number of PSMs is comparable to that of the non-outlying experiments mainly capture sources of variability that do not necessarily impede spectral identifications, such as, for example, biological differences. Meanwhile, the QC metrics that show a clear discrepancy in terms of valid PSMs compared to the non-outlying experiments warrant a closer look, as they might indicate potential problems during the experiment resulting in a diminished performance.

As monitoring a large number of QC metrics on a regular basis is often unpractical, it might be more convenient to limit the general analysis to a small number of user-friendly, well-understood, and discriminating metrics [46]. From table 6.4 and figure 6.12b it can be deduced that, for



(a) For all outliers combined there is a major overlap between the outlying and the non-outlying experiments.



(b) When considering the relevant QC metrics for the outliers, a notable difference can be observed between various QC metrics. The metrics considered are those that most frequently occur in the explanatory subspaces, listed in table 6.4.

Figure 6.12: Comparison of the number of PSMs between the non-outlying and the outlying experiments for the TCGA dataset.

example, metrics detailing the chromatographic performance, the TIC accumulation, and the precursor ionization are suitable candidates. Because these metrics occur frequently in the explanatory subspaces, they can be used to detect a wide range of outliers. Furthermore, they seem to indicate a significant decrease in identification performance, highlighting the outliers that are most likely to have a negative influence on the eventual output results. The efficacy of these QC metrics has independently been noted before as well, with the chromatographic peak width and the electrospray ionization selected to monitor the experimental quality during a previous multicenter performance study [46].

6.4 Software availability

To aid users in the quality exploration of their own experiments, we have bundled the presented analysis techniques into a software toolkit tuned for mass spectrometry quality control. This software, which has been fully coded in Python, making use of scientific and machine learning libraries such as NumPy [260], pandas [183], and scikit-learn [207], allows users to run the presented workflow to detect and interpret outlying experiments. The software can be used as a command-line application and exports the outlier analysis as a qcML file [272], which can be viewed in any web browser.

All code is released as open source and is available at `https://bitbucket.org/proteinspector/qc_analysis`.

6.5 Conclusions

A recent informal poll conducted by the Netherlands Protemic Center (NPC) highlighted that proteomics researchers regard quality control as a crucial component during data analysis, but it also revealed that the majority of researchers do not incorporate systematic quality control approaches in their day-to-day workflows because applying currently existing quality control tools is perceived to be too hard. This clearly shows that there is still significant room for improvement to make quality control an ubiquitous step during MS proteomics experiments, something that is vital for mass spectrometry-based methods to mature into analytical, transparent, and reproducible disciplines [177].

In this paper, we have presented a powerful technique to perform an initial filtering of low-quality mass spectrometry experiments based on computationally derived quality control metrics, with a strong focus on providing easily interpretable results. After all, when identifying low-performance experiments, it is insufficient to just know that an experiment has failed, it is also crucial to understand why the experiment in question exhibits a decreased performance to be able to remedy the problems that caused the failure. Furthermore, we have shown that our approach can be successfully applied across different instruments and instrument types, and for sample contents with varying complexity. As such, the methodology we have presented can play an important role in investigating the performance of experiments, as it is able to detect outlying experiments, as well as provide an explanation about the outlying behavior, which can be interpreted by domain experts. A potential disadvantage, however, is that to able to successfully detect low-quality experiments, it must be possible to establish a positive baseline, which requires that there are a sufficient number of experiments available and that the number of low-quality experiments does not exceed the number of high-quality experiments. As in normal conditions only a few low-quality experiments are present, and because our approach does not require manual setting of this positive baseline, but instead is able to automatically infer

it, we believe that this is not a limiting factor. Additionally, because we employ an unsupervised technique, some false positives are unavoidable. However, to determine which experiments exhibit a decreased performance the presented technique can be used as an initial filtering step prior to a detailed manual inspection, which will drastically decrease the number of experiments that need to be checked manually, resulting in significant time savings.

However, the quality control analyses presented here should not stand on their own, instead they need to be integrated with the experimental results and they should be closely linked to all operational information and relevant events pertaining to the mass spectrometry instruments. For example, environmental conditions, instrument maintenance schedules, etc., all can have a significant influence on the experimental results [20, 34]. This information should ideally be structurally recorded in a sort of electronic lab notebook (ELN), and should then be related to the experimental results and the quality analysis.

Furthermore, novel analyses or algorithmic approaches are insufficient on their own. Ideally there should be a consolidation of the developed quality control methods to date, and these methods should be made available to a wide audience in intuitive and user-friendly tools, something that is still severely lacking in the community at large.

Finally, although here we explicitly focused on quality control for proteomics, our approach has potential applications in other mass spectrometry-based domains, such as metabolomics, as well. Current quality control practices in metabolomics mainly involve the direct comparison of features measured in specialized QC samples to subject samples [68, 97, 100], whereas the QC metrics that were considered here form an additional level of abstraction as they are derived from the experimental results. However, because we employed identification-free metrics, which do not directly depend on domain-specific identification procedures, but instead capture the general properties of a mass spectrometry experiment, most of these metrics can be straightforwardly applied outside of the proteomics setting as well. Therefore, to conclude, extending such computational QC approaches to related fields such as metabolomics can be a very interesting avenue of future research.

Chapter 7

Optimized open modification spectral library searching

Abstract

Open modification searching is a search strategy that enables the identification of modified peptides without having to explicitly specify the modifications under consideration, leading to record numbers of identified spectra. However, open modification searching suffers from a high computational burden because the search space is considerably expanded by opening up the precursor mass window to the order of several hundreds of Dalton. We here present how approximate nearest neighbors indexing techniques can be applied to spectral library searching to speed up open modification searching. Approximate nearest neighbors indexing allows to retrieve only the most relevant candidates from a spectral library to avoid expensive and undue similarity computations. We show how approximate nearest neighbors indexing achieves a speed-up of several orders of magnitude during open modification searching, rendering this a viable identification strategy.

Preface

This work was selected for an oral presentation at the renowned American Society for Mass Spectrometry (ASMS) annual conference 2016 in San Antonio, TX, USA, where it received highly positive feedback.

7.1 Introduction

Although mass spectrometry (MS) is a very powerful technique to characterize proteins in complex biological samples, a significant portion of spectra often still remains unidentified. One of the reasons for this is that some of the unidentified spectra represent peptides containing post-translational modifications (PTMs) [49]. PTMs arise from covalent changes to the proteins after their synthesis in the ribosome, and they play a key role in many cellular processes [271]. As such, the identification of PTMs can give us crucial insights into various protein activities. However, although MS techniques have become quite mature, a comprehensive identification of modified proteins in complex samples remains challenging [139, 173].

An unknown tandem mass spectrum is typically identified by calculating a score between the spectrum and all its potential matches, after which the highest scoring match is selected as the

identification. However, it has been observed that numerous spectra remain unidentified [49] or are misidentified [36] because they correspond to peptides containing one or more unexpected PTMs. To correctly identify these spectra all occurring modifications have to be explicitly specified in the search settings. However, because for each peptide that contains a potential modification both the modified and unmodified variant(s) have to be considered this leads to a significant increase in search space, resulting in a higher computational load and a loss in sensitivity. Consequently, generally only a limited selection of the most prevalent modifications are considered. Whereas typically only candidates that fall within a limited mass window around the query spectrum's precursor mass are considered, with the width of the mass window depending on the instrument's mass accuracy, an alternative strategy to identify modified spectra, called open modification searching (OMS), consists of opening up the precursor mass window [5, 192]. By allowing a wide precursor mass window modified query spectra can be compared to their standard, unmodified, variant. This implicitly considers all possible PTMs, after which the presence and type of the modifications can be derived from the mass difference between the observed precursor mass and the mass of the unmodified peptide.

Although it is possible to identify a wide range of spectra containing diverse PTMs through OMS, resulting in an unparalleled number of identifications, this approach suffers from a drastically increased search space because a very wide precursor mass window has to be used. A popular approach to keep OMS computationally feasible is to use spectral libraries [6, 118, 165, 285]. Because spectral libraries only contain previously observed peptides the search space is severely restricted compared to sequence database searching strategies where all theoretically possible peptides are considered [105]. An additional advantage of spectral libraries is that they contain 'real' spectra, i.e. spectra that have been confidently identified in previous experiments (mostly processed to form consensus spectra [153]), as opposed to *in silico* generated, theoretical, spectra in the case of sequence database searching. This enables the calculation of a more sensitive score [290], which is especially beneficial in the case of OMS because identifications of modified query spectra are derived from partial matches between these modified spectra and the unmodified, commonly confidently identified, library spectra.

However, despite the reduced search space exhibited by spectral libraries compared to sequence databases, due to the increased availability of high-quality datasets in public data repositories [211] the size of spectral libraries has significantly risen over the past few years, as indicated in figure 7.1. Spurred by the significant increase in computational requirements caused by the continuous expansion of spectral libraries a few approaches have been proposed to speed up spectral library searching ranging from parallelizing spectral matching using graphics processing units (GPUs) [16], to utilizing the Apache Spark big data processing engine to search spectral libraries [118], to using a filtering approach based on a shared peak count to reduce the number of candidates for which a spectrum-spectrum match (SSM) with a query spectrum has to be computed [269].

In the case of OMS though the computational burden is increased further by orders of magnitude due to the expansion of the search space. To counter this we propose a general indexing technique to reduce the number of candidates matches that have to be evaluated for each query spectrum. By making use of an (approximate) nearest neighbor indexing technique it is possible to quickly retrieve the most similar library spectra to a query spectrum without depending on the precursor mass, therefore inherently supporting open modification searching. Because traditional multidimensional indexing techniques break down in high dimensions because of the curse of dimensionality [27], considering the very high-dimensional nature of spectra we make use of approximate nearest neighbors (ANN) techniques instead. These indexing methods are able to cope with higher dimensionalities by loosening the guarantee that an exact solution is provided at all times; instead sometimes only an approximate solution is achieved [122]. Here we

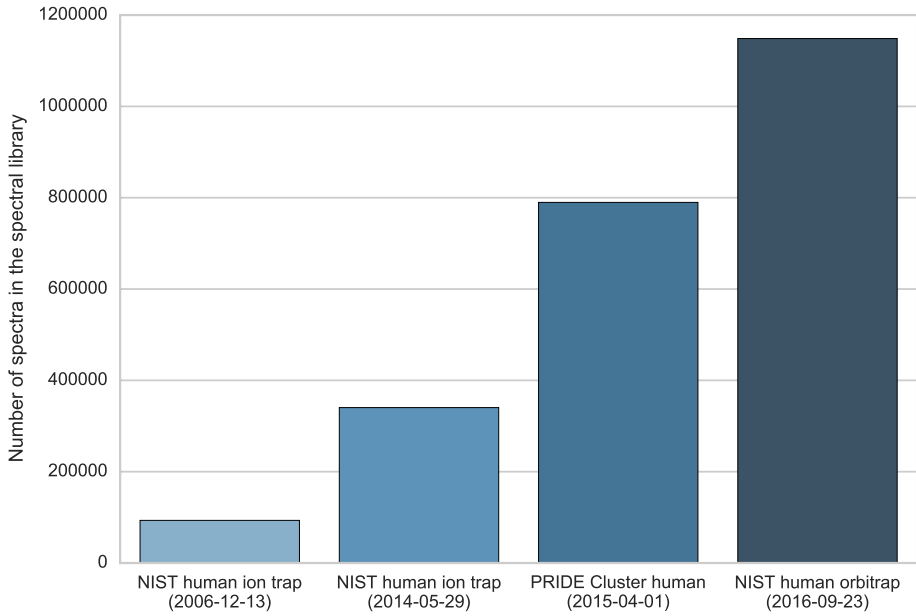


Figure 7.1: Spectral library sizes increase as more high-quality public datasets become available. Representative human spectral libraries from the National Institute of Standards and Technology (NIST) and the PRoteomics IDEntifications (PRIDE) repository [267] were used. The NIST spectral libraries are curated by NIST, whereas the PRIDE Cluster spectral library is generated by clustering spectra deposited to PRIDE [106, 107].

will show how ANN techniques can be used to index spectral libraries. Notably, because of its independence on the precursor mass to filter the candidates in the spectral library, we show that this approach is ideally suited for OMS. Using these techniques we achieve significant speed-ups while being able to identify spectra containing unexpected modifications, making OMS a viable identification strategy.

7.2 Spectral library indexing

7.2.1 Approximate nearest neighbor indexing

Thanks to advances in terms of recording and storing data, the amount of available data has tremendously increased in recent years. This wealth of data has prompted the need for so-called 'big data' approaches, with advanced algorithmic techniques necessary to uncover new insights. A common task is nearest neighbor searching, which aim is to find the most similar known items in a database compared to an unknown query item. Applications of nearest neighbor searching are for example finding similar products to generate recommendations for (e-)commerce applications, performing image recognition by identifying similar pictures, recognizing text by comparing similar characters when performing optical character recognition (OCR), etc. A naive approach to find the nearest neighbors in a database for a query item is to loop through the entire database and compute the similarity between the query item and all database items. However, because of

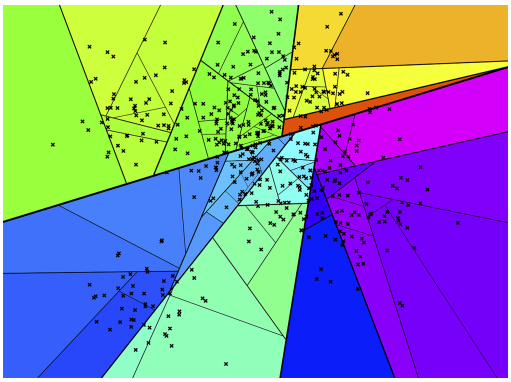
the massive sizes of the available data such an exhaustive, brute-force, similarity search is often computationally unfeasible and would take an excessive amount of time. Instead, specialized indexing structures can be used to quickly find the most similar items. These indexing structures function in such a way that instead of having to exhaustively look at all items in the database, the most similar neighbors for a query item can be found in sublinear time in function of the database size. For dealing with multidimensional items, several multidimensional indexing techniques have been used in a variety of domains, for example such as to deal with spatial data in a geographic information system (GIS). However, traditional multidimensional indexing techniques break down when faced with a medium to high number of dimensions due to the curse of dimensionality [27]. Instead, for such high-dimensional data, often ANN indexing techniques are used [122]. These techniques relax the requirement to provide the exact nearest neighbors in all cases, and sometimes provide only an approximate solution, in favor of significant performance gains.

Although multiple algorithms for ANN indexing exist, a common property of most of these algorithms is that they repeatedly partition the data space into smaller subspaces that are locality sensitive. This means that these subspaces are more likely to hold items that are close to each other than items that are far away. We have employed the Approximate Nearest Neighbors Oh Yeah (Annoy) library for ANN indexing, which is based on random projections to divide the data space. The random projections form split hyperplanes to recursively subdivide the data space into two subspaces. In this fashion a binary search tree can be constructed, as is shown in figures 7.2a and 7.2c. This search tree can be used to efficiently find the data subspace containing the nearest neighbors to a query item in sublinear time, as is shown in figures 7.2b and 7.2c. However, because indexing in high-dimensional spaces is a hard problem and due to the randomization employed no single search tree will likely be optimal. In contrast, by exploiting the random behavior complementary index trees can be constructed if alternative split hyperplanes are used. The index trees can be combined in an ensemble, or 'forest' of index trees, with additional trees providing complementary views on the data, which serves to offset the random effects and the curse of dimensionality. By using multiple trees the correct set of nearest neighbors for a query item can be approximated more accurately. For example, if a query item falls very close to the border of a data subspace, its nearest neighbors might be present in an adjacent subspace. However, by merging the candidate nearest neighbors from all index trees in the forest to construct a compound data subspace, as is shown in figure 7.2d, the occurrence of suboptimal results can be minimized at the expense of additional processing time as multiple index trees need to be examined. Therefore the number of index trees in the forest can be used to configure a trade-off between processing speed and the accuracy of the results.

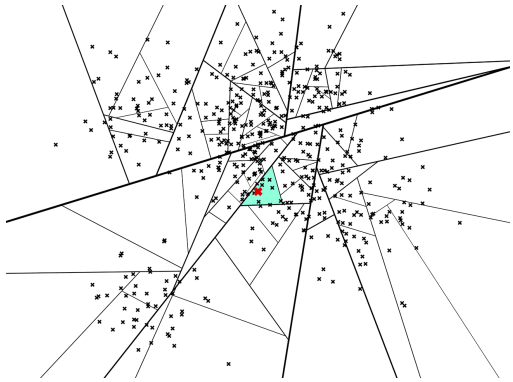
7.2.2 Spectral library searching

During spectral library searching an experimental, unknown, query spectrum is compared to the known spectra in the spectral library. For each candidate spectrum in the spectral library (with the candidates typically determined by a filter on the precursor charge and the precursor mass window) a match score is computed to indicate its similarity to the query spectrum, forming an SSM. Next, to identify the query spectrum, the SSM with the highest score is selected and the query spectrum is assigned the same peptide sequence as the corresponding library spectrum [149, 150].

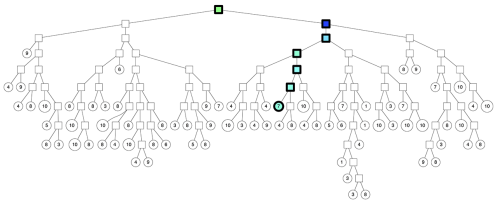
In terms of the terminology introduced in the previous section, to identify an unknown spectrum, i.e. the query, its nearest neighbor is retrieved from the database, i.e. the spectral library (which should not be confused with the 'database' during sequence database searching).



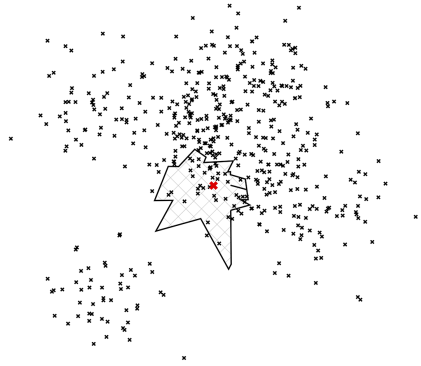
(a) The data is partitioned in subspaces based on random split hyperplanes.



(b) The data subspace to which a query item (highlighted in red) belongs is retrieved to find its nearest neighbors.



(c) The required data subspace can be found in sublinear time using the binary search tree.



(d) A better approximation can be obtained by combining the candidate nearest neighbors from multiple randomly constructed index trees.

Figure 7.2: Representation of ANN searching.

Spectrum preprocessing

First, prior to spectral library searching the spectra are subjected to some preprocessing [223]. Specifically, low-quality spectra are discarded, while the remaining spectra are processed to remove noise peaks and only retain the 50 most intense peaks [152]. Next, the peak intensities are rank transformed [165] or scaled by their square root [152] to de-emphasize overly dominant peaks.

Second, because multidimensional indexing techniques work in vector spaces, the spectra are converted to vectors by dividing a fixed mass range into equispaced mass bins, after which these are normalized to unit length. When choosing the bin width a trade-off between expressivity of the vectors and performance of the ANN index has to be considered. As current high-resolution instruments are able to very accurately measure masses small mass bins can be used to faithfully represent the spectra. In contrast, the ANN index degenerates to a linear search for excessive dimensionalities. Empirically a bin width of 1 Da was determined optimal, which corresponds to previous approaches [75, 152].

Candidate selection

The identification of an unknown query spectrum by spectral library searching (or equivalently by sequence database searching) can be divided into two separate phases: (i) candidate selection, and (ii) candidate scoring. Typically, in the first phase all spectra in the spectral library that have the same precursor charge as the query spectrum and whose precursor mass lies within a specified window around the query's precursor mass are selected as the admissible candidates. In the next phase, for all these candidates a detailed matching score is computed, and the SSM with the highest score is selected as the identification for the query spectrum.

For traditional, 'closed', searches during the candidate selection phase a limited precursor mass window is used, typically at the parts per million (ppm) level or at most a few Dalton. This significantly restricts the number of candidates that have to be scored during the second phase, ensuring that the scoring phase remains computationally feasible. Conversely, during OMS the precursor mass window is considerably opened up and is typically in the order of several hundreds of Dalton [6, 49, 118, 165, 191, 258, 285], which leads to a huge increase in the number of selected candidates that have to be scored. Although the vast majority of the selected candidates will be quite dissimilar from the query spectrum, a matching score will still have to be computed for these candidates, imposing a considerable, yet avoidable, strain on the computational resources.

By using an ANN index the candidate selection phase can be optimized. Instead of filtering the candidates based on the precursor mass window, potentially including very dissimilar spectra as candidates, performing a nearest neighbor query using the ANN index allows to select only the the most similar library spectra as candidates. Subsequently these candidates can still be additionally filtered using a precursor mass window to exclude those candidates whose precursor mass differs excessively. This optimized candidate selection strategy severely reduces the number of candidates for which a computationally expensive matching score has to be computed, significantly shortening the required computational time by orders of magnitude.

Spectral matching

The dot product is commonly used as similarity measure during spectral library searching [58, 88, 89, 152] as it is ideally suited to measure spectral similarity [244]. Using an ANN index spectral matching occurs in two phases. In the first step vectors are matched in the ANN index

to retrieve the candidates. Because multidimensional indexing techniques depend on the triangle inequality to prune the search space they mandate the use of a distance metric as similarity measure. Therefore, instead of the dot product, which in its basic form does not adhere to the triangle inequality, the angular distance defined as the Euclidean distance between two normalized vectors is used as a proxy for the dot product. This angular distance is equivalent to the dot product when comparing unit vectors, without loss of generality for dealing with non-normalized vectors.

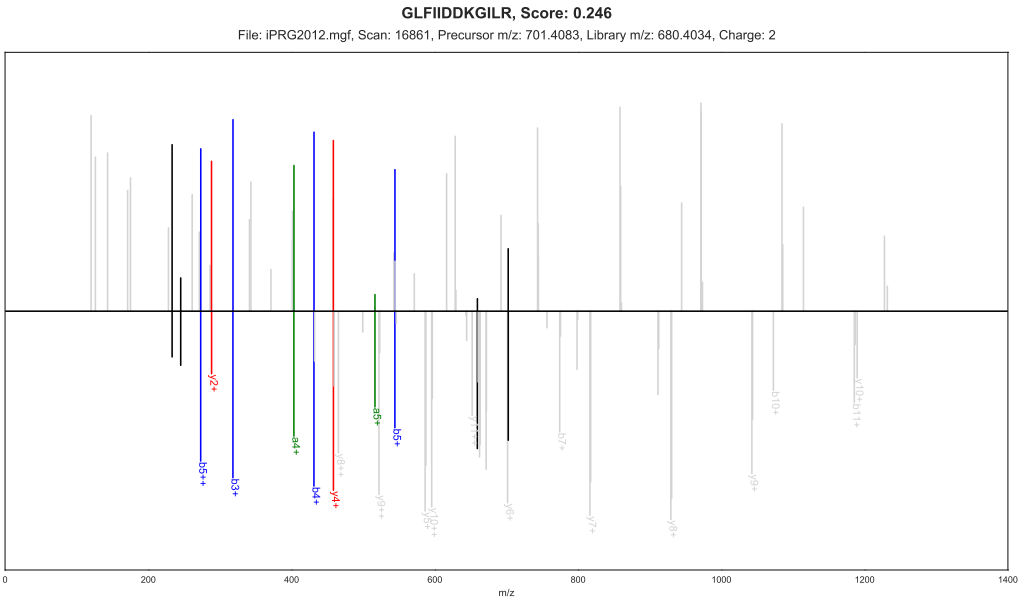
Candidates retrieved from the ANN index are ranked based on the angular distance to the query vector, after which in a second step the similarity between the query and each candidate is refined by making use of the full-resolution spectra instead of their low-dimensional vector representations. We have provided two related similarity measures to score SSMs: the dot product and a variant thereof which takes mass shifts into account to determine peak agreements between two spectra. Whereas the basic dot product is mostly useful to assess whether two spectra represent identical peptides, during OMS unmodified spectra can be compared to their modified variants. Although OMS is fundamentally based on the observation that spectra representing modified and unmodified variants of the same peptide are similar to some extent, knowledge on PTM characteristics can be taken into account to sensitively assess spectral similarity. When comparing two spectra the shifted dot product checks for directly corresponding peaks, equivalent to the basic dot product, while additionally allowing peaks to be shifted according to the precursor mass difference between the two spectra that are being compared. For a non-zero precursor mass difference annotated peaks in the library spectrum are allowed to be shifted by the precursor mass difference at a charge of one up to the charge of the peak, while unannotated peaks are allowed to be shifted at a charge of one up to the precursor charge. All peak agreements including shifted and unshifted peaks are scored based on the corresponding peak intensities, after which the optimal agreements are extracted using a priority queue while only allowing each peak to be matched at most once. As illustrated in figure 7.3 this allows peptides differing by a single PTM to be matched accurately and sensitively.

7.3 Speeding up open modification searching

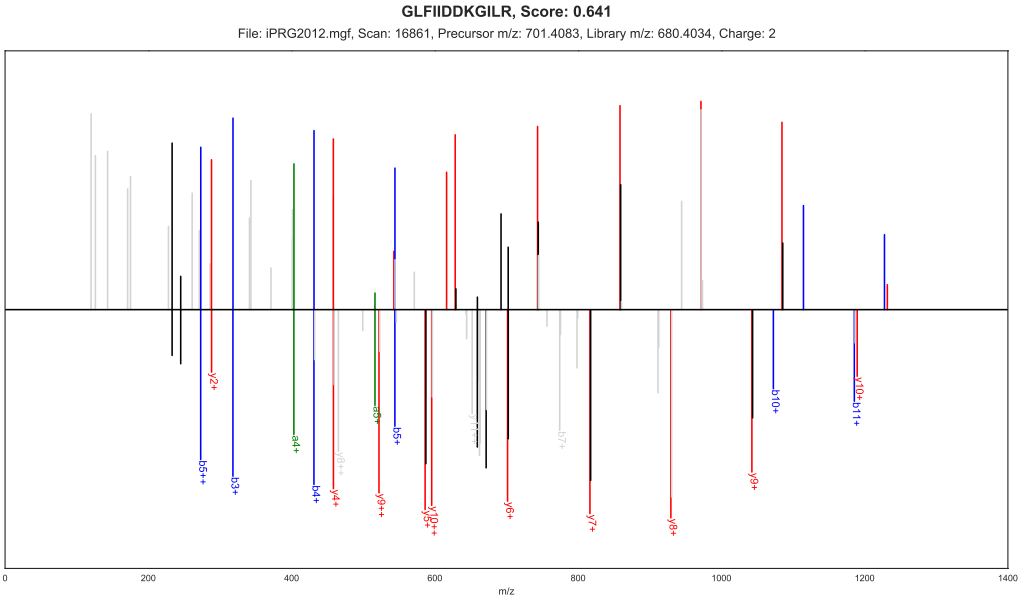
7.3.1 Experimental data

We used publicly available data generated for the purpose of the 2012 study by the Proteome Informatics Research Group (iPRG) of the Association of Biomolecular Resource Facilities (ABRF). The goal of this study was to assess the community's ability to identify modified peptides [47]. To this aim, various participating researchers were asked to identify an unknown dataset, after which their proficiency in handling modified peptides was evaluated. The provided dataset consisted of a mixture of synthetic peptides with biologically occurring modifications combined with a yeast whole cell lysate as background, and the spectra were measured on a TripleTOF instrument. For full details on the sample preparation see the original publication by Chalkley et al. [47]. This is a high-quality dataset which comes recommended as a reference dataset for the evaluation of identification algorithms [93]. All data was downloaded from the MassIVE data repository (MassIVE accession MSV000078492).

To identify the iPRG_2012 dataset the human ion trap (version 2012/05/30) and yeast ion trap (version 2012/04/06) spectral libraries compiled by NIST and obtained from the PeptideAtlas [63] website were used. First, both spectral libraries were concatenated using SpectraST [152], after which additional decoy spectra were added in a $1:1$ ratio [151], resulting in a single large spectral library file containing 799574 spectra.



(a) The default dot product only matches the fragments that do not include the acetylation.



(b) The shifted dot product correctly matches both unmodified and modified fragments.

Figure 7.3: The dot product is used to score the similarity between two spectra. In both figures the top spectrum is the unknown query spectrum and the bottom spectrum is the library spectrum, with matching peaks colored according to the peak annotation. As can be derived from the precursor mass difference the peptide GLFIIDDKGILR has undergone acetylation (mass 42.010565 Da) on the lysine at position 8. The default dot product only takes directly matching peaks into account, while the shifted dot product allows for shifted peaks according to the precursor mass difference, correctly assigning a high score to this match.

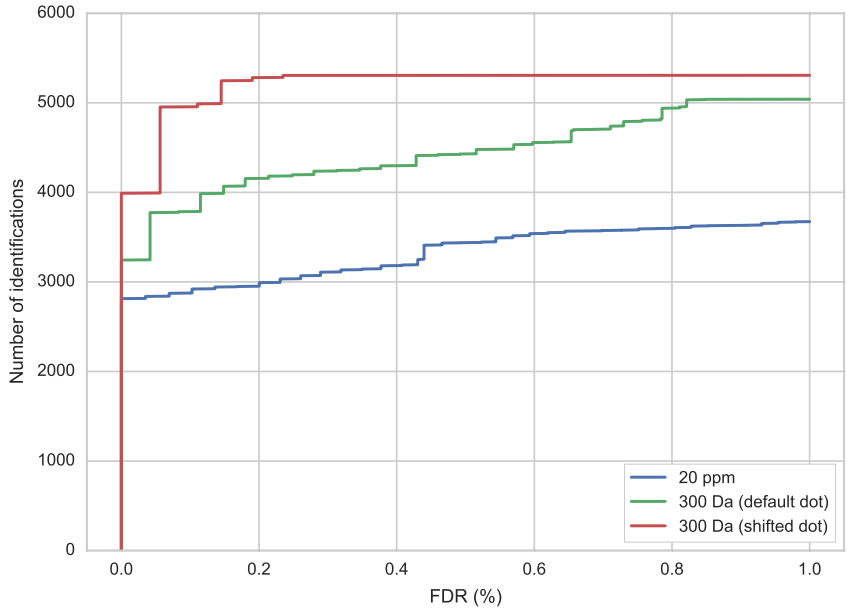
7.3.2 Code availability

The ANN spectral library software is fully coded in Python, making use of scientific libraries such as NumPy [260] and pandas [183]. The ANN indexing is based upon the popular open-source Annoy library [239]. Annoy is a high-performance software library written in C++ with Python bindings originally developed at Spotify to support music recommendations. All code is released as open source and is available at `https://bitbucket.org/proteinspector/ann-solo`.

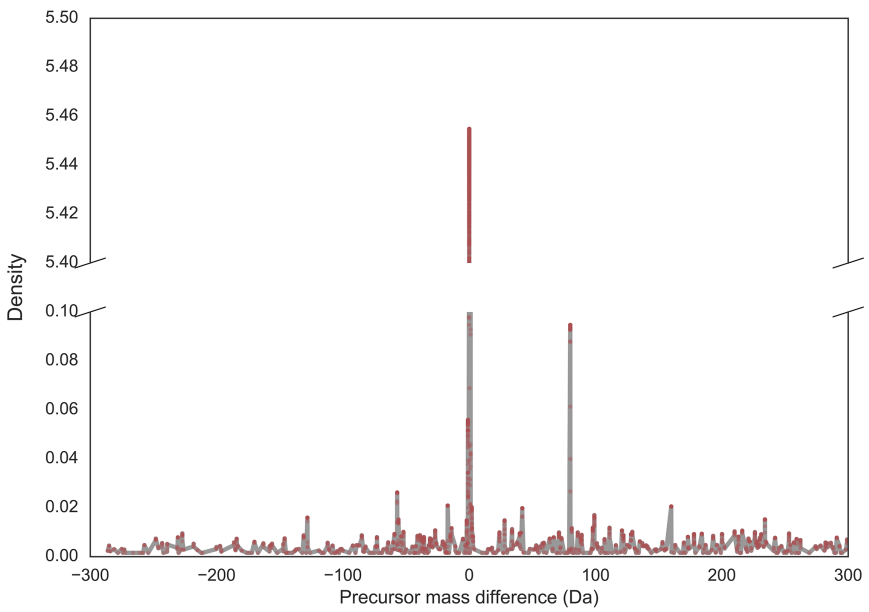
7.3.3 ANN spectral library searching

The big advantage of OMS is that modified peptides can be identified without having to explicitly specify all modifications under consideration. This is done by opening up the precursor mass window in order to match modified spectra to their unmodified variant despite the large mass difference produced by the modification. This makes it possible to successfully identify a considerably higher proportion of spectra compared to the traditional approach. Figure 7.4a shows that for the iPRG_2012 dataset a direct increase in identified spectra of up to 40 % can be achieved by opening up the precursor mass window from 20 ppm to 300 Da to take potential modifications into account. Furthermore, as the shifted dot product is an optimized similarity measure to sensitively match modified spectra, it succeeds in discriminating true SSMs from false SSMs, achieving a further increase in identification rate. This increase is especially pronounced at a low false discovery rate (FDR), indicating that the shifted dot product correctly accounts for systematic peak shifts resulting from modifications. Next, although OMS does not directly indicate or localize PTMs, the type of modifications present can be derived from the difference in precursor mass between the query spectra (modified or not) and the library spectra. As shown in figure 7.4b, the majority of SSMs does not exhibit a precursor mass difference between the query spectrum and the library spectrum, corresponding to unmodified peptides that can be directly matched. Other mass differences indicate the presence of some common modifications, such as phosphorylation (+79.966331 Da), acetylation (+42.010565 Da), or loss of ammonia (−17.026549 Da). Further mass differences correspond to amino acid substitutions between the query spectra and library spectra or a surplus of amino acids at one end of the peptide sequence, confirming that the spectral library does not contain all relevant peptide sequences.

Despite this proven proficiency in identifying modified spectra the OMS strategy remains underused, in large part because of the considerable computational requirements imposed by opening up the precursor mass window. Whereas if the precursor mass window ranges a few ppm only a handful or at most a few dozen candidates need to be considered for each query, by opening up the precursor mass window to the order of hundreds of Dalton a large majority of spectra in the spectral library will need to be considered as candidates for each single query. Consequently, whereas a traditional search can typically be done in a few minutes, OMS can easily require several hours of computation time, vastly surpassing the analytical time required to generate the data. In contrast, by exploiting ANN indexing to reduce the number of candidates that need to be considered OMS can be sped up by several orders of magnitude, as shown in figure 7.5. When making use of ANN indexing to speed up spectral library searching there is a configurable trade-off between precision and speed. By varying how many trees are constructed to index the spectral library and how deep the trees are traversed during the candidate retrieval more accurate results can be obtained at the expense of an increase in computation time. As indicated by figure 7.5 this is subject to diminishing returns: obtaining fully accurate results incurs a disproportionate increase in computation time, even though a speed-up of orders of magnitude is still achieved. These parameters can be varied in function of the use case. For example, a preliminary screening of the modifications that are present in the dataset can be achieved in a



(a) OMS results in a significant increase in the number of identified spectra compared to the traditional search using a limited precursor mass window.



(b) Density plot of the precursor mass differences between the query spectra and the library spectra. The possible presence of a modification and its identity can be derived from the precursor mass difference between the modified and unmodified spectra.

Figure 7.4: OMS identification results for the iPRG_2012 dataset.

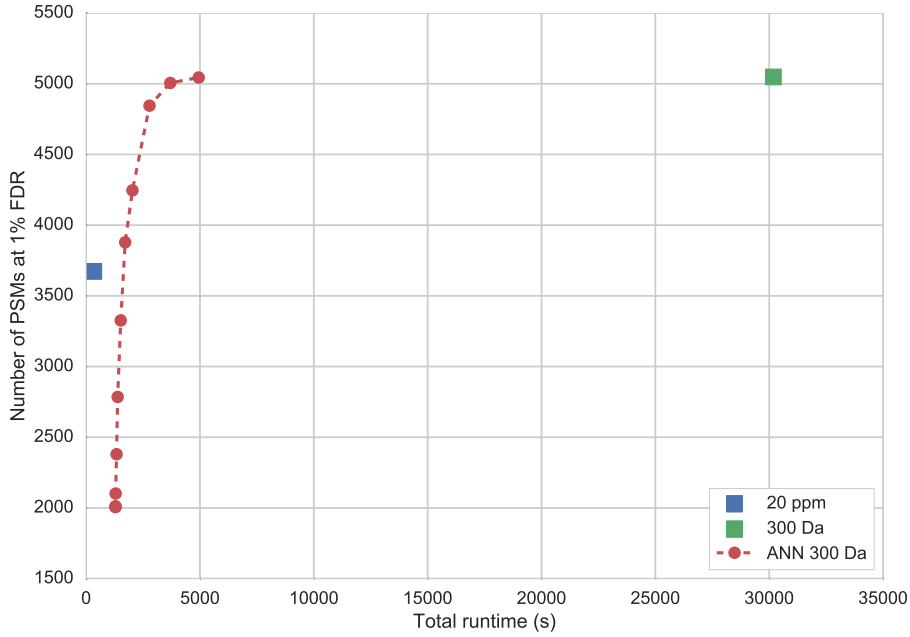


Figure 7.5: Elapsed time for searches with different sizes of precursor mass window. ANN indexing can reduce the time required for OMS by several orders of magnitude.

minimal amount of time, which can be followed by a targeting investigation of modifications of interest.

7.4 Conclusions

We have shown how ANN indexing can be applied to spectral library searching. By filtering the candidate matches for which an SSM needs to be scored during OMS the search procedure can be sped up by several orders of magnitude. Whereas previously OMS has sporadically been lauded for its ability to identify a large number of modified spectra, its adoption remained limited due to its excessive computational requirements. In contrast, by making use of an ANN index the processing time can be kept under control, rendering OMS a viable identification strategy.

This is especially beneficial in light of the ever-increasing size of publicly available high-quality spectral libraries. Current large spectral libraries originate for example from spectral clustering approaches to generate repository-wide spectral libraries [106, 107] or from spectral archives which contain both identified and unidentified spectra [87]. Another interesting approach that aims to address the limited coverage of spectral libraries is to use computational tools to simulate spectral libraries covering a full proteome [286, 287]. As traditional spectral library searching scales linearly in terms of the library size, searching such large spectral libraries would result in a considerable computation time. In contrast, because ANN indexing scales logarithmically in terms of the library size using a larger spectral library would result in only a modest increase in computation time.

Furthermore, besides OMS there are other use cases that have to deal with expensive computations due to a very large search space as well, such as when identifying proteogenomics [195] or metaproteomics [190] datasets. Although ANN indexing is inherently ideally suited to OMS because the candidate retrieval does not depend on the precursor mass, equivalently in these use cases ANN indexing could be used to restrict the search space and significantly reduce the computational requirements.

Chapter 8

Conclusion

To conclude this dissertation we will briefly summarize our work and highlight the major contributions we have presented, and discuss some open problems and interesting avenues of future work.

8.1 Summary of contributions

Mass spectrometry (MS) is a very powerful analytical technique that can be used to identify and quantify the protein content of complex biological samples. However, because MS techniques are necessarily highly sensitive, the results of an MS experiment can be subject to a significant level of variability. Therefore, to inspire confidence in the acquired experimental results it is essential to correctly assess and control this variability, for which a systematic approach to quality control (QC) is needed. In this dissertation we have presented several valuable contributions to computational QC approaches for mass spectrometry-based proteomics. Our work mainly deals with various aspects that are needed to establish a fundamental QC approach supporting the application of mass spectrometry in a biomedical setting.

First, we have shown how secondary QC metrics, such as instrument settings and environmental variables, can provide valuable information that can be related to the quality of an MS experiment. Whereas traditional QC metrics are derived from the spectral data, possibly in combination with the identification results, and are highly dependent on the type of experiment that was performed, instrument settings can be universally monitored irrespective of the experimental set-up. Furthermore, because instrument settings provide very low-level information on the operation of an MS instrument they enable detecting emerging problems as soon as they arise. Because mass spectrometry is so sensitive environmental variables, such as a slight shift in the ambient laboratory temperature for example, can already have a significant impact on the experimental results. This data can therefore provide important information on the performance of an MS instrument. We have shown that commodity hardware can suffice to monitor specific environmental variables and provide novel insights into the quality of the generated data. Systematically monitoring this information, ideally in combination with traditional spectral-derived QC metrics, allows to establish a technological passport specific to an individual MS instrument, which serves to obtain an in-depth understanding of its performance and aids to minimize and understand the observed variability in the experimental results.

Second, we have taken part in the community effort to define the qcML standard file format for QC data. The qcML format provides straightforward yet formal storage of QC information, and by functioning as a common interface it enables interoperability between different QC tools. To provide the necessary technical infrastructure for interacting with qcML data we have developed the jqcML open-source Java application program interface (API), which can be used by other

developers to effortlessly support the qcML format. The development of the qcML format is still ongoing, with the recently established Quality Control working group of the Human Proteome Organization (HUPO) - Proteomics Standards Inititative (PSI) spearheading efforts to improve the format definition, add qcML support to several existing QC tools, and opening up QC procedures to any mass spectrometry-based experimental workflow besides the typical discovery setting.

Third, we have developed an automated approach to discriminate high-quality from low-quality MS experiments based on unsupervised outlier detection applied to a high-dimensional set of QC metrics. Even though multiple tools exist that can compute QC metrics, their output is often highly complex and expert knowledge is required to interpret it. In contrast, our approach is able to accurately assess the quality of an MS experiment without requiring manual intervention. Furthermore, we have implemented an outlier interpretation scheme to highlight the most relevant QC metrics giving rise to the decrease in performance for the detected low-quality experiments. This gives instrument operators actionable knowledge to optimize their experimental set-up, as these interpretations have been shown to correspond to independently published expert knowledge. First of all, by detecting and interpreting low-quality experiments MS practitioners are served with practical information they can use to optimize their experimental set-up. Second, as our approach does not require manual user input it introduces the possibility of automated decision-making. For example, spectral acquisition could be automatically halted if a decrease in quality is detected to avoid undue loss of precious sample content.

The above research contributes to three essential aspects needed to establish a comprehensive QC procedure: being able to record meaningful performance metrics, the ability to store and communicate QC data, and using this information for advanced decision-making. It is only by combining these separate aspects into a unified system that biological mass spectrometry can reach its full potential.

A final contribution relates to analyzing the acquired spectral data after suitable QC procedures have ascertained its validity. In a typical MS experiment only a quarter to half of all spectra can be correctly identified. Recent research has shown that a significant number of spectra remain unidentified because they represent modified peptides whose specific modifications were not considered in the search settings [49]. Instead, open modification searching (OMS) can be used to identify spectra without explicitly having to specify any potential post-translational modifications (PTMs), resulting in a significant increase in the number of identified spectra. OMS remains an underused strategy though, mainly because it suffers from an explosion of the search space, rendering it prohibitively computationally expensive. To solve this issue we have used approximate nearest neighbors (ANN) indexing techniques to limit the search space when performing OMS, resulting in a speed-up of several orders of magnitude. This renders OMS practically feasible, making it possible to identify a considerably higher number of spectra compared to a standard search without having to spend an excessive amount of computational resources.

8.2 Future work

As scientists have rightly realized that suitable quality control procedures are required to advance the field of mass spectrometry-based proteomics, in recent years there has been a proliferation of computational QC tools. However, despite the availability of multiple advanced QC tools their day-to-day usage remains extremely limited, in large part because they lack usability. First, the tools frequently mandate expert knowledge. A common problem is the need for complex user-specified parameters to correctly process the spectral data. Additionally, it is often not straightforward to interpret the generated results, even for expert users. We have tried to address this issue in

part when developing our approach to discriminate low-quality experiments from high-quality experiments [30]. The outlier detection requires only a single user-specified parameter, the size of the local neighborhood to determine whether a specific observation is an outlier or not. This parameter is conceptually easy to understand, and the outlier detection technique is fairly robust to this parameter as shown by the validation results in chapter 6. As such, this method can be automated to some extent. Furthermore, we have implemented an outlier interpretation scheme that limits the data to only the most relevant QC metrics to provide detailed insights into the source of the deterioration in experimental quality.

A second problem contributing to the lack in user-friendliness of the QC tools is that besides expert analytical knowledge to interpret the results, they also often require advanced technical knowledge to configure and troubleshoot. For example, as we have reviewed in chapter 3, installing the SIMPle AuTomatIc Quality COntrol (SimpatiQCo) tool [215] involves a non-trivial procedure consisting of multiple steps. Furthermore, SimpatiQCo requires users to manually insert a new record in its underlying database to set up a new instrument for monitoring. Our contributions are not flawless in this respect either. The Instrument MONitoring DataBase (iMonDB) tool [34] indiscriminately extracts all instrument parameters whose values are present in the experimental output files without using any smart filtering. Depending on the instrument type this can include several dozens to a few hundreds of instrument parameters, of which the majority do not convey any useful information. As the iMonDB tool currently does not include functionality to intelligently serve only the relevant QC metrics users are forced to browse through all parameters to verify if any changes occurred. This obviously severely limits the usefulness of the iMonDB tool, as multiple (inter)national users have communicated.

Unfortunately though a lack of usability is a common issue exhibited by academic software [161]. Although there exist a few computational proteomics groups that have built a reputation of consistently delivering great software tools and which are doing sterling work, in general, research groups often do not have the necessary technical expertise in-house to develop and, importantly, maintain user-friendly software. First, there is a significant gap in level of difficulty between coding a command-line script to perform a specific analysis and developing a robust and full-fledged end-user application or managing a legacy code-base. For the former some 'hacking' skills might suffice, whereas the latter requires profound software engineering skills. A large portion of interdisciplinary bioinformatics researchers lack these skills though, as they have never enjoyed a formal computer science education. This should not detract anything from their accomplishments, as their efforts to learn how to program and to publish their code to the benefit of the community should be applauded, but expectations should be set accordingly. Second, even for research groups that possess the necessary technical skills the current academic climate often does not encourage expending time to develop and support elaborate software tools. Similar to the Pareto principle, during software development (roughly) 80 % of the functionality stems from 20 % of the efforts. With current metrics for measuring academic value putting emphasis on publishing as many papers as possible, it is often deemed not worthwhile to spend the additional development time needed to extend one-off analysis scripts to user-friendly (graphical) tools.

A further issue is that QC tools often require researchers to explicitly expend substantial effort and they function in isolation of both the experimental workflow and the subsequent data analysis, which does not encourage their adoption. In contrast, the tools should be automated and form an integral part of any MS experiment. A tool that is fully integrated within the experimental workflow is able to assist researchers in every step of their experiments. Through comprehensive monitoring of the performance the ideal QC tool would be able to for example optimize wet-laboratory experimental protocols and suggest informed settings for the bioinformatics data analysis. This is not a trivial matter though. As mass spectrometry is such a powerful and versatile technique its use cases are legion. Consequently, it is impossible to define an exclusive

QC directive or develop a single tool that covers all possible applications.

To this end contributions by the HUPO-PSI Quality Control working group [33], as introduced in chapter 5, could turn out to have a high value. Although the qcML format has already been published a few years ago [272], its adoption so far has remained limited. Establishing the qcML format as the definitive HUPO-PSI standard for QC data will encourage third-party developers to support it, which will stimulate the growth of an ecosystem of QC tools around this format. Similar to previous success stories, such as the mzML format [179], this will allow different tools to seamlessly interact with each other. As MS use cases are too diverse to define a single QC directive, modular tools will be able to focus on specific applications. Next, these tools can be combined into a powerful pipeline to offer a rich QC system providing the requisite functionality. With members from the proteomics as well as the metabolomics communities, and involvement from academic, industry, and journal representatives, the HUPO-PSI Quality Control working group is uniquely placed to tackle this important task.

Taking a step back from only quality control for mass spectrometry-based proteomics to machine learning (ML) applied to proteomics data in general, there are multiple opportunities as well. We are convinced that a more data-driven approach can result in significant improvements of the current computational methods. At the moment bioinformatics applications are mostly model-driven, i.e. they function based on explicitly encoded expert knowledge. For example, sequence database search engines identify unknown mass spectra by comparing the experimentally observed spectra to *in silico* simulated spectra generated based on expert rules describing the fragmentation of peptides within a mass spectrometer under certain conditions. In contrast, as famously posited a few years ago under the slogan "the unreasonable effectiveness of data" [109], advanced algorithms can succeed in deriving high-quality models from large datasets instead of having to hand-code complex rules. The field of mass spectrometry-based proteomics seems ready for such an overhaul from rule-driven to data-driven models as well. Data sharing has long been mandated by journals upon publication [225], and public data repositories have seen a steady increase in deposited datasets [65, 266]. Some notable efforts to reanalyze these publicly available datasets to generate novel knowledge have been undertaken in the past few years [264, 265], but the full power of this data remains underused. We surmise that by harnessing this wealth of available public data a deeper understanding of the many processes underlying the generation of MS data can be obtained, which will enable the discovery of novel biological findings.

A concrete application hereof, tying in with the work on measuring instrument parameters presented in chapter 4, is to use ML to optimize spectral acquisition. Currently instrument QC metrics are stored as summary statistics for each run in the iMonDB [34]. Instead measurements pertaining to each individual scan can be collected. These granular metrics would then have to be related to the quality of individual scans, which information can be used to learn patterns corresponding to optimal spectrum acquisition. A comparable approach has been undertaken by Google, who have described how they use supervised learning on multivariate sensor values to minimize the energy consumption of their data centers [62, 104]. Initially this approach could be used to provide insights into the functioning of complex MS machinery to manually optimize the acquisition settings. Ideally though the ML model should directly interface with the instrument to change these settings on the fly.

Unfortunately, in general, an issue hindering the development of algorithms based on MS data is the lack of a ground truth. MS data is subject to a variety of influences, and different analysis methods can yield varying results [31]. In contrast, to validate novel bioinformatics algorithms quantifiable measurements are required instead of the current subjective comparisons [238]. Whereas in general for supervised learning a manually labeled ground truth set is used, determining a conclusive ground truth for MS data is not possible. Advanced expert knowledge is required to manually identify unknown mass spectra, which is time-consuming and cannot be done at

scale. Furthermore, even expert users do not succeed in interpreting a considerable portion of mass spectra. Recent suggestions include using simulated data and curating publicly available gold standard datasets [93], but this is an important problem that will need to be addressed further. Quality control can play an important role here, as the availability of comprehensive QC information describing publicly available datasets can help to determine the data's applicability to be used for training and validating novel computational approaches.

The latest trend in machine learning is without a doubt deep learning (DL) [157]. Based on increases in computational performance and the accumulation of ever larger datasets, deep neural networks (NNs) have revolutionized a variety of ML tasks. So far DL has been most successfully applied to image processing tasks, but there are opportunities to apply this fascinating research to proteomics problems. One of the early breakthroughs at the forefront of the current interest in deep NNs was when a convolutional neural network (ConvNet) achieved a considerable increase in accuracy on the ImageNet challenge [147]. For this challenge participants are required to correctly assign labels to unknown images [228], a task on which the data-driven ConvNet significantly outperformed its traditional rule-driven competitors. One might initially assume that a similar approach can be employed to identify mass spectra, which consist of structured information comparable to images. However, the ImageNet challenge requires users to only predict 1000 different labels [228]. Taking such a sizable number of classes into account is significantly harder than a simple binary classification, however, the number of possible peptides is even several orders of magnitude higher. Furthermore, very different spectra can originate from the same peptide due to PTMs, fragmentation and mass measurement characteristics, etc. Clearly, even current deep NNs are incapable to directly predict such a complex output value. Instead, the systematic structure of a peptide might be exploited using recurrent neural networks (RNNs) [115]. RNNs are used to process sequences of inputs, resulting in state-of-the-art performance on a variety of pattern recognition tasks, such as speech recognition or natural language processing. We hypothesize that these NN architectures can learn how different peaks in a spectrum are related to identify a spectrum at the granularity of individual amino acids, which would result in a sort of implicit, data-driven, *de novo* spectrum identification.

Generative adversarial networks (GANs) are another interesting new DL development. Only recently introduced by Goodfellow et al. [103] in 2014, experts have called it "the most interesting idea in machine learning in the last 10 years." [156] A GAN consists of two separate networks: a generator and a discriminator. The generator produces simulated data, while the discriminator differentiates between true instances and simulated instances produced by the generator. As the two networks compete with each other the generator learns to produce seemingly real data, while the discriminator simultaneously learns to accurately classify data as being real or simulated. State-of-the-art applications of GANs include enhancing pictures by upscaling their resolution [158] or generating photorealistic images based on only a textual description [289]. Applied to mass spectrometry-based proteomics GANs could be trained to discriminate between true and false spectrum identifications. In this case the generator would be trained to simulate false peptide-spectrum matches (PSMs), while the discriminator would be tasked to differentiate these simulated PSMs from high-confidence target PSMs. A straightforward application of such a GAN would be to generate decoy spectral libraries. Current decoy generation procedures typically permute the order of amino acids in target peptide sequences and shuffle the corresponding peaks of the spectra accordingly [151]. However, this approach has some limitations. As the spectrum fragmentation process is not fully understood the peaks are shuffled based on some heuristics. Furthermore, only annotated peaks are deterministically shuffled, whereas unannotated peaks are dealt with in a random fashion. Instead, a GAN could be used to generate highly realistic decoy spectra that are virtually indiscernible from real spectra. Similarly, this approach could be used to generate proteome-wide simulated spectral libraries to increase the libraries' coverage [286, 287]. Even further, instead of making use of trivial fragmentation rules a GAN could be used to

generate *in silico* spectra during sequence database searching. By making use of highly realistic simulated spectra the sequence database searching paradigm and the spectral library searching paradigm would effectively be equalized. In the above cases mainly the GAN's generator is used. However, the discriminator can also be used to great effect, as it has learned to differentiate true positive PSMs from false positive PSMs. Currently true positive identifications are distinguished from false positive identifications based on the target-decoy strategy [74]. However, because the scores of the target identifications overlap with the scores of the decoy identifications to some extent [135] a loss of correct identifications is unavoidable. It has already been shown that ML techniques can be used to provide a better distinction between true positive and false positive identifications based on PSM metadata [129], but a GAN would be able to distinguish true positive from false positive PSMs even more accurately by taking the raw spectral information into account. This would at least result in a higher number of identified spectra, and could possibly obviate the need for the target-decoy strategy by introducing a novel approach to validate spectral identifications.

With current DL research progressing at a breakneck speed highly advanced NNs have made it possible to perform end-to-end learning. This means that the model is able to automatically learn complex internal representations instead of having to explicitly break up a compound task into multiple simpler subtasks. These new approaches can usher in a shift from the current model-driven paradigm in proteomics bioinformatics to a data-driven one. Therefore, we want to conclude by expressing our confidence that by combining state-of-the-art machine learning and computational proteomics knowledge tremendous progress can be achieved in the near future, impacting any application of biological mass spectrometry.

Bibliography

[1] Susan E Abbatiello, D R Mani, Birgit Schilling, Brendan MacLean, et al. "Design, Implementation and Multisite Evaluation of a System Suitability Protocol for the Quantitative Assessment of Instrument Performance in Liquid Chromatography-Multiple Reaction Monitoring-MS (LC-MRM-MS)". In: *Molecular & Cellular Proteomics* 12.9 (Sept. 1, 2013), pp. 2623–2639. DOI: 10.1074/mcp.M112.027078.

[2] Terri A Addona, Susan E Abbatiello, Birgit Schilling, Steven J Skates, et al. "Multi-Site Assessment of the Precision and Reproducibility of Multiple Reaction Monitoring–based Measurements of Proteins in Plasma". In: *Nature Biotechnology* 27.7 (July 2009), pp. 633–641. DOI: 10.1038/nbt.1546.

[3] Ruedi Aebersold and Matthias Mann. "Mass-Spectrometric Exploration of Proteome Structure and Function". In: *Nature* 537.7620 (Sept. 14, 2016), pp. 347–355. DOI: 10.1038/nature19949.

[4] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. "On the Surprising Behavior of Distance Metrics in High Dimensional Space". In: *Proceedings of the 8th International Conference on Database Theory - ICDT '01.* Vol. 1973. Lecture Notes in Computer Science. London, England: Springer Berlin Heidelberg, 2001, pp. 420–434. DOI: 10.1007/3-540-44503-X_27.

[5] Erik Ahrné, Markus Müller, and Frederique Lisacek. "Unrestricted Identification of Modified Proteins Using MS/MS". In: *PROTEOMICS* 10.4 (Feb. 2010), pp. 671–686. DOI: 10.1002/pmic.200900502.

[6] Erik Ahrné, Frederic Nikitin, Frederique Lisacek, and Markus Müller. "QuickMod: A Tool for Open Modification Spectrum Library Searches". In: *Journal of Proteome Research* 10.7 (July 1, 2011), pp. 2913–2921. DOI: 10.1021/pr200152g.

[7] Stephan Aiche, Timo Sachsenberg, Erhan Kenar, Mathias Walzer, et al. "Workflows for Automated Downstream Data Analysis and Visualization in Large-Scale Computational Mass Spectrometry". In: *PROTEOMICS* 15.8 (Apr. 2015), pp. 1443–1447. DOI: 10.1002/pmic.201400391.

[8] Juan-Pablo Albar and Francesc Canals. "Standardization and Quality Control in Proteomics". In: *Journal of Proteomics* 95 (Dec. 16, 2013), pp. 1–2. DOI: 10.1016/j.jprot.2013.11.002.

[9] Brett G Amidan, Daniel J Orton, Brian L LaMarche, Matthew E Monroe, et al. "Signatures for Mass Spectrometry Data Quality". In: *Journal of Proteome Research* 13.4 (Apr. 4, 2014), pp. 2215–2222. DOI: 10.1021/pr401143e.

[10] Genna L Andrews, Ralph A Dean, Adam M Hawkridge, and David C Muddiman. "Improving Proteome Coverage on a LTQ-Orbitrap Using Design of Experiments". In: *Journal of The American Society for Mass Spectrometry* 22.4 (Apr. 2011), pp. 773–783. DOI: 10.1007/s13361-011-0075-2.

[11] Keith A Baggerly, Jeffrey S Morris, Sarah R Edmonson, and Kevin R Coombes. "Signal in Noise: Evaluating Reported Reproducibility of Serum Proteomic Tests for Ovarian Cancer". In: *JNCI Journal of the National Cancer Institute* 97.4 (Feb. 16, 2005), pp. 307–309. DOI: 10.1093/jnci/dji008.

[12] Monya Baker. "1,500 Scientists Lift the Lid on Reproducibility". In: *Nature* 533.7604 (May 25, 2016), pp. 452–454. DOI: 10.1038/533452a.

[13] Marcus Bantscheff, Simone Lemeer, Mikhail M Savitski, and Bernhard Kuster. "Quantitative Mass Spectrometry in Proteomics: Critical Review Update from 2007 to the Present". In: *Analytical and Bioanalytical Chemistry* 404.4 (Sept. 2012), pp. 939–965. DOI: 10.1007/s00216-012-6203-4.

[14] Steven J Bark and Vivian Hook. "Differential Recovery of Peptides from Sample Tubes and the Reproducibility of Quantitative Proteomic Data". In: *Journal of Proteome Research* 6.11 (Nov. 1, 2007), pp. 4511–4516. DOI: 10.1021/pr070294o.

[15] Vicki J Barwick. "Sources of Uncertainty in Gas Chromatography and High-Performance Liquid Chromatography". In: *Journal of Chromatography A* 849.1 (July 16, 1999), pp. 13–33.

[16] Lydia Ashleigh Baumgardner, Avinash Kumar Shanmugam, Henry Lam, Jimmy K Eng, et al. "Fast Parallel Tandem Mass Spectral Library Searching Using GPU Hardware Acceleration". In: *Journal of Proteome Research* 10.6 (June 3, 2011), pp. 2882–2888. DOI: 10.1021/pr200074h.

[17] Daniel W Bearden, Richard D Beger, David Broadhurst, Warwick Dunn, et al. "The New Data Quality Task Group (DQTG): Ensuring High Quality Data Today and in the Future". In: *Metabolomics* 10.4 (Aug. 2014), p. 0. DOI: 10.1007/s11306-014-0679-1.

[18] Ashley Beasley-Green, David Bunk, Paul Rudnick, Lisa Kilpatrick, et al. "A Proteomics Performance Standard to Support Measurement Quality in Proteomics". In: *PROTEOMICS* 12.7 (Apr. 2012), pp. 923–931. DOI: 10.1002/pmic.201100522.

[19] Alexander W Bell, Eric W Deutsch, Catherine E Au, Robert E Kearney, et al. "A HUPO Test Sample Study Reveals Common Problems in Mass Spectrometry–based Proteomics". In: *Nature Methods* 6.6 (June 29, 2009), pp. 423–430. DOI: 10.1038/nmeth.1333.

[20] Keiryn L Bennett, Xia Wang, Cory E Bystrom, Matthew C Chambers, et al. "The 2012/2013 ABRF Proteomic Research Group Study: Assessing Longitudinal Intralaboratory Variability in Routine Peptide Liquid Chromatography Tandem Mass Spectrometry Analyses". In: *Molecular & Cellular Proteomics* 14.12 (Dec. 1, 2015), pp. 3299–3309. DOI: 10.1074/mcp.O115.051888.

[21] Michael S Bereman. "Tools for Monitoring System Suitability in LC MS/MS Centric Proteomic Experiments". In: *PROTEOMICS* 15 (5-6 Mar. 2015), pp. 891–902. DOI: 10.1002/pmic.201400373.

[22] Michael S Bereman, Joshua Beri, Vagisha Sharma, Cory Nathe, et al. "An Automated Pipeline to Monitor System Performance in Liquid Chromatography–tandem Mass Spectrometry Proteomic Experiments". In: *Journal of Proteome Research* 15.12 (Dec. 2, 2016), pp. 4763–4769. DOI: 10.1021/acs.jproteome.6b00744.

[23] Michael S Bereman, Richard Johnson, James Bollinger, Yuval Boss, et al. "Implementation of Statistical Process Control for Proteomic Experiments via LC MS/MS". In: *Journal of The American Society for Mass Spectrometry* 25.4 (Apr. 2014), pp. 581–587. DOI: 10.1007/s13361-013-0824-5.

[24] Joshua Beri, Michael M Rosenblatt, Ethan Strauss, Marjeta Urh, et al. "Reagent for Evaluating Liquid Chromatography–tandem Mass Spectrometry (LC-MS/MS) Performance in Bottom-up Proteomic Experiments". In: *Analytical Chemistry* 87.23 (Dec. 1, 2015), pp. 11635–11640. DOI: 10.1021/acs.analchem.5b04121.

[25] Helen Berman, Kim Henrick, and Haruki Nakamura. "Announcing the Worldwide Protein Data Bank". In: *Nature Structural Biology* 10.12 (Dec. 2003), pp. 980–980. DOI: 10.1038/nsb1203-980.

[26] Michael R Berthold, Nicolas Cebron, Fabian Dill, Thomas R Gabriel, et al. "KNIME - the Konstanz Information Miner: Version 2.0 and Beyond". In: *ACM SIGKDD Explorations Newsletter* 11.1 (June 2009), pp. 26–31. DOI: 10.1145/1656274.1656280.

[27] Kevin S Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. "When Is "Nearest Neighbor" Meaningful?" In: *Proceedings of the 7th International Conference on Database Theory - ICDT '99*. Jerusalem, Israel: Springer-Verlag, 1999, pp. 217–235.

[28] Chris Bielow, Guido Mastrobuoni, and Stefan Kempa. "Proteomics Quality Control: Quality Control Software for MaxQuant Results". In: *Journal of Proteome Research* 15.3 (Mar. 4, 2016), pp. 777–787. DOI: 10.1021/acs.jproteome.5b00780.

[29] Wout Bittremieux, Pieter Kelchtermans, Dirk Valkenborg, Lennart Martens, et al. "jqcML: An Open-Source Java API for Mass Spectrometry Quality Control Data in the qcML Format". In: *Journal of Proteome Research* 13.7 (July 3, 2014), pp. 3484–3487. DOI: 10.1021/pr401274z.

[30] Wout Bittremieux, Pieter Meysman, Lennart Martens, Dirk Valkenborg, et al. "Unsupervised Quality Assessment of Mass Spectrometry Proteomics Experiments by Multivariate Quality Control Metrics". In: *Journal of Proteome Research* 15.4 (Apr. 1, 2016), pp. 1300–1307. DOI: 10.1021/acs.jproteome.6b00028.

[31] Wout Bittremieux, David L Tabb, Francis Impens, Lennart Martens, et al. "Quality Control in Mass-Spectrometry-Based Proteomics". In: *Mass Spectrometry Reviews* (Manuscript submitted).

[32] Wout Bittremieux, Dirk Valkenborg, Lennart Martens, and Kris Laukens. "Computational Quality Control Tools for Mass Spectrometry Proteomics". In: *PROTEOMICS* (Early view Oct. 17, 2016). DOI: 10.1002/pmic.201600159.

[33] Wout Bittremieux, Mathias Walzer, Stefan Tenzer, Weimin Zhu, et al. "The HUPO-PSI Quality Control Working Group: Making Quality Control More Accessible for Biological Mass Spectrometry". In: *Analytical Chemistry* (In revision).

[34] Wout Bittremieux, Hanny Willems, Pieter Kelchtermans, Lennart Martens, et al. "iMonDB: Mass Spectrometry Quality Control through Instrument Monitoring". In: *Journal of Proteome Research* 14.5 (May 1, 2015), pp. 2360–2366. DOI: 10.1021/acs.jproteome.5b00127.

[35] Anna Bodzon-Kulakowska, Anna Bierczynska-Krzysik, Tomasz Dylag, Anna Drabik, et al. "Methods for Samples Preparation in Proteomic Research". In: *Journal of Chromatography B* 849 (1-2 Apr. 15, 2007), pp. 1–31. DOI: 10.1016/j.jchromb.2006.10.040.

[36] Boris Bogdanow, Henrik Zauber, and Matthias Selbach. "Systematic Errors in Peptide and Protein Identification and Quantification by Modified Peptides". In: *Molecular & Cellular Proteomics* 15.8 (Aug. 1, 2016), pp. 2791–2801. DOI: 10.1074/mcp.M115.055103.

[37] Emily S Boja and Henry M Fales. "Overalkylation of a Protein Digest with Iodoacetamide". In: *Analytical Chemistry* 73.15 (Aug. 1, 2001), pp. 3576–3582. DOI: 10.1021/ac0103423.

[38] Adele Bourmaud, Sebastien Gallien, and Bruno Domon. "A Quality Control of Proteomic Experiments Based on Multiple Isotopologous Internal Standards". In: *EuPA Open Proteomics* 8 (Sept. 2015), pp. 16–21. DOI: 10.1016/j.euprot.2015.07.010.

[39] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. "LOF: Identifying Density-Based Local Outliers". In: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data - SIGMOD '00*. Dallas, Texas, USA: ACM, May 2000, pp. 93–104. DOI: 10.1145/335191.335388.

[40] Corey D Broeckling, Adam L Heuberger, Jonathan A Prince, E Ingelsson, et al. "Assigning Precursor–product Ion Relationships in Indiscriminant MS/MS Data from Non-Targeted Metabolite Profiling Studies". In: *Metabolomics* 9.1 (Feb. 2013), pp. 33–43. DOI: 10.1007/s11306-012-0426-4.

[41] Daniel Broudy, Trevor Killeen, Meena Choi, Nicholas Shulman, et al. "A Framework for Installable External Tools in Skyline". In: *Bioinformatics* 30.17 (Sept. 1, 2014), pp. 2521–2523. DOI: 10.1093/bioinformatics/btu148.

[42] Jakob Bunkenborg, Guadalupe Espadas, and Henrik Molina. "Cutting Edge Proteomics: Benchmarking of Six Commercial Trypsins". In: *Journal of Proteome Research* 12.8 (Aug. 2, 2013), pp. 3631–3641. DOI: 10.1021/pr4001465.

[43] Julia Maria Burkhart, Thomas Premsler, and Albert Sickmann. "Quality Control of Nano-LC-MS Systems Using Stable Isotope-Coded Peptides". In: *PROTEOMICS* 11.6 (Mar. 2011), pp. 1049–1057. DOI: 10.1002/pmic.201000604.

[44] Julia Maria Burkhart, Cornelia Schumbrutzki, Stefanie Wortelkamp, Albert Sickmann, et al. "Systematic and Quantitative Comparison of Digest Efficiency and Specificity Reveals the Impact of Trypsin Quality on MS-Based Proteomics". In: *Journal of Proteomics* 75.4 (Feb. 2, 2012), pp. 1454–1462. DOI: 10.1016/j.jprot.2011.11.016.

[45] David A Cairns. "Statistical Issues in Quality Control of Proteomic Analyses: Good Experimental Design and Planning". In: *PROTEOMICS* 11.6 (Mar. 2011), pp. 1037–1048. DOI: 10.1002/pmic.201000579.

[46] Alex Campos, Ramón Díaz, Salvador Martínez-Bartolomé, Jose Sierra, et al. "Multicenter Experiment for Quality Control of Peptide-Centric LC–MS/MS Analysis — A Longitudinal Performance Assessment with nLC Coupled to Orbitrap MS Analyzers". In: *Journal of Proteomics* 127, Part B (Sept. 8, 2015), pp. 264–274. DOI: 10.1016/j.jprot.2015.05.012.

[47] Robert J Chalkley, Nuno Bandeira, Matthew C Chambers, Karl R Clauser, et al. "Proteome Informatics Research Group (iPRG)_2012: A Study on Detecting Modified Peptides in a Complex Mixture". In: *Molecular & Cellular Proteomics* 13.1 (Jan. 1, 2014), pp. 360–371. DOI: 10.1074/mcp.M113.032813.

[48] Matthew C Chambers, Brendan Maclean, Robert Burke, Dario Amodei, et al. "A Cross-Platform Toolkit for Mass Spectrometry and Proteomics". In: *Nature Biotechnology* 30.10 (Oct. 10, 2012), pp. 918–920. DOI: 10.1038/nbt.2377.

[49] Joel M Chick, Deepak Kolippakkam, David P Nusinow, Bo Zhai, et al. "A Mass-Tolerant Database Search Identifies a Large Proportion of Unassigned Spectra in Shotgun Proteomics as Modified Peptides". In: *Nature Biotechnology* 33.7 (June 15, 2015), pp. 743–749. DOI: 10.1038/nbt.3267.

[50] Lynn A Cole and John G Dorsey. "Temperature Dependence of Retention in Reversed-Phase Liquid Chromatography. 1. Stationary-Phase Considerations". In: *Analytical Chemistry* 64.13 (July 1, 1992), pp. 1317–1323. DOI: 10.1021/ac00037a004.

[51] Lynn A Cole, John G Dorsey, and Ken A Dill. "Temperature Dependence of Retention in Reversed-Phase Liquid Chromatography. 2. Mobile-Phase Considerations". In: *Analytical Chemistry* 64.13 (July 1, 1992), pp. 1324–1327. DOI: 10.1021/ac00037a005.

[52] *colims - CompOmics LIMS system*. URL: https://code.google.com/p/colims/ (visited on 11/03/2014).

[53] *common Repository of Adventitious Proteins*. URL: http://www.thegpm.org/crap/ (visited on 06/30/2016).

[54] Richard G Côté, Florian Reisinger, and Lennart Martens. "jmzML, an Open-Source Java API for mzML, the PSI Standard for MS Data". In: *PROTEOMICS* 10.7 (Apr. 7, 2010), pp. 1332–1335. DOI: 10.1002/pmic.200900719.

[55] Jürgen Cox, Marco Y Hein, Christian A Luber, Igor Paron, et al. "Accurate Proteome-Wide Label-Free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ". In: *Molecular & Cellular Proteomics* 13.9 (Sept. 1, 2014), pp. 2513–2526. DOI: 10.1074/mcp.M113.031591.

[56] Jürgen Cox and Matthias Mann. "MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification". In: *Nature Biotechnology* 26.12 (Nov. 30, 2008), pp. 1367–1372. DOI: 10.1038/nbt.1511.

[57] Jürgen Cox, Ivan Matic, Maximiliane Hilger, Nagarjuna Nagaraj, et al. "A Practical Guide to the MaxQuant Computational Platform for SILAC-Based Quantitative Proteomics". In: *Nature Protocols* 4.5 (Apr. 16, 2009), pp. 698–705. DOI: 10.1038/nprot.2009.36.

[58] R Craig, JC Cortens, D Fenyo, and RC Beavis. "Using Annotated Peptide Mass Spectrum Libraries for Protein Identification". In: *Journal of Proteome Research* 5.8 (Aug. 4, 2006), pp. 1843–1849. DOI: 10.1021/pr0602085.

[59] Robertson Craig and Ronald C Beavis. "TANDEM: Matching Proteins with Tandem Mass Spectra". In: *Bioinformatics* 20.9 (Feb. 19, 2004), pp. 1466–1467. DOI: 10.1093/bioinformatics/bth092.

[60] Robertson Craig, John P Cortens, and Ronald C Beavis. "Open Source System for Analyzing, Validating, and Storing Protein Identification Data". In: *Journal of Proteome Research* 3.6 (Dec. 1, 2004), pp. 1234–1242. DOI: 10.1021/pr049882h.

[61] Maria Cristina Ferreira de Oliveira and Haim Levkowitz. "From Visual Data Exploration to Visual Data Mining: A Survey". In: *IEEE Transactions on Visualization and Computer Graphics* 9.3 (July 2003), pp. 378–394. DOI: 10.1109/TVCG.2003.1207445.

[62] DeepMind. *DeepMind AI Reduces Google Data Centre Cooling Bill by 40%*. July 20, 2016. URL: https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/ (visited on 12/21/2016).

[63] Frank Desiere, Eric W Deutsch, Nichole L King, Alexey I Nesvizhskii, et al. "The PeptideAtlas Project". In: *Nucleic Acids Research* 34.90001 (Jan. 1, 2006), pp. D655–D658. DOI: 10.1093/nar/gkj040.

[64] Eric W Deutsch, Juan Pablo Albar, Pierre-Alain Binz, Martin Eisenacher, et al. "Development of Data Representation Standards by the Human Proteome Organization Proteomics Standards Initiative". In: *Journal of the American Medical Informatics Association* 22.3 (Feb. 28, 2015), pp. 495–506. DOI: 10.1093/jamia/ocv001.

[65] Eric W Deutsch, Attila Csordas, Zhi Sun, Andrew Jarnuczak, et al. "The ProteomeXchange Consortium in 2017: Supporting the Cultural Change in Proteomics Public Data Deposition". In: *Nucleic Acids Research* Advance access (Oct. 18, 2016), gkw936. DOI: 10.1093/nar/gkw936.

[66] Julia Dittrich, Susen Becker, Max Hecht, and Uta Ceglarek. "Sample Preparation Strategies for Targeted Proteomics via Proteotypic Peptides in Human Blood Using Liquid Chromatography Tandem Mass Spectrometry". In: *PROTEOMICS - Clinical Applications* 9 (1-2 Feb. 2015), pp. 5–16. DOI: 10.1002/prca.201400121.

[67] Sebastiaan Dolman, Sebastiaan Eeltink, Axel Vaast, and Matthias Pelzing. "Investigation of Carryover of Peptides in Nano-Liquid Chromatography/Mass Spectrometry Using Packed and Monolithic Capillary Columns". In: *Journal of Chromatography B* 912 (Jan. 1, 2013), pp. 56–63. DOI: 10.1016/j.jchromb.2012.11.016.

[68] Warwick B Dunn, Ian D Wilson, Andrew W Nicholls, and David Broadhurst. "The Importance of Experimental Design and QC Samples in Large-Scale and MS-Driven Untargeted Metabolomic Studies of Humans". In: *Bioanalysis* 4.18 (Sept. 2012), pp. 2249–2264. DOI: 10.4155/bio.12.204.

[69] *EclipseLink JPA*. URL: http://www.eclipse.org/eclipselink/jpa.php (visited on 12/12/2013).

[70] *EclipseLink MOXy*. URL: http://www.eclipse.org/eclipselink/moxy.php (visited on 12/12/2013).

[71] Nathan J. Edwards, Mauricio Oberti, Ratna R. Thangudu, Shuang Cai, et al. "The CPTAC Data Portal: A Resource for Cancer Proteomics Research". In: *Journal of Proteome Research* 14.6 (June 5, 2015), pp. 2707–2713. DOI: 10.1021/pr501254j.

[72] Jarrett D Egertson, Brendan MacLean, Richard Johnson, Yue Xuan, et al. "Multiplexed Peptide Analysis Using Data-Independent Acquisition and Skyline". In: *Nature Protocols* 10.6 (May 21, 2015), pp. 887–903. DOI: 10.1038/nprot.2015.055.

[73] Martin Eisenacher, Anke Schnabel, and Christian Stephan. "Quality Meets Quantity - Quality Control, Data Standards and Repositories". In: *PROTEOMICS* 11.6 (Mar. 2011), pp. 1031–1036. DOI: 10.1002/pmic.201000441.

[74] Joshua E Elias and Steven P Gygi. "Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry". In: *Nature Methods* 4.3 (Feb. 27, 2007), pp. 207–214. DOI: 10.1038/nmeth1019.

[75] Jimmy K Eng, Ashley L McCormack, and John R III Yates. "An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database". In: *Journal of the American Society for Mass Spectrometry* 5.11 (Nov. 1994), pp. 976–989. DOI: 10.1016/1044-0305(94)80016-2.

[76] Jimmy K Eng, Brian C Searle, Karl R Clauser, and David L Tabb. "A Face in the Crowd: Recognizing Peptides through Database Search". In: *Molecular & Cellular Proteomics* 10.11 (Nov. 1, 2011), R111.009522–R111.009522. DOI: 10.1074/mcp.R111.009522.

[77] *Enterprise JavaBeans 3.0*. URL: https://jcp.org/en/jsr/detail?id=220 (visited on 12/12/2013).

[78] Claudia Escher, Lukas Reiter, Brendan MacLean, Reto Ossola, et al. "Using iRT, a Normalized Retention Time for More Targeted Measurement of Peptides". In: *PROTEOMICS* 12.8 (Apr. 2012), pp. 1111–1121. DOI: 10.1002/pmic.201100463.

[79] Claire E Eyers, Deborah M Simpson, Stephen C C Wong, Robert J Beynon, et al. "QCAL—a Novel Standard for Assessing Instrument Conditions for Proteome Analysis". In: *Journal of the American Society for Mass Spectrometry* 19.9 (Sept. 2008), pp. 1275–1280. DOI: 10.1016/j.jasms.2008.05.019.

[80] Iakes Ezkurdia, Enrique Calvo, Angela Del Pozo, Jesús Vázquez, et al. "The Potential Clinical Impact of the Release of Two Drafts of the Human Proteome". In: *Expert Review of Proteomics* 12.6 (Nov. 2, 2015), pp. 579–593. DOI: 10.1586/14789450.2015.1103186.

[81] Pan Fang, Mingqi Liu, Yu Xue, Jun Yao, et al. "Controlling Nonspecific Trypsin Cleavages in LC-MS/MS-Based Shotgun Proteomics Using Optimized Experimental Conditions". In: *Analyst* 140.22 (2015), pp. 7613–7621. DOI: 10.1039/C5AN01505G.

[82] David Fenyö, Jan Eriksson, and Ronald Beavis. "Mass Spectrometric Protein Identification Using the Global Proteome Machine". In: *Computational Biology*. Ed. by David Fenyö. Vol. 673. Methods in Molecular Biology. Totowa, NJ: Humana Press, Aug. 17, 2010, pp. 189–202.

[83] Oliver Fiehn, Gert Wohlgemuth, Martin Scholz, Tobias Kind, et al. "Quality Control for Plant Metabolomics: Reporting MSI-Compliant Studies". In: *The Plant Journal* 53.4 (Feb. 2008), pp. 691–704. DOI: 10.1111/j.1365-313X.2007.03387.x.

[84] Erin J Finehout, Jason R Cantor, and Kelvin H Lee. "Kinetic Characterization of Sequencing Grade Modified Trypsin". In: *PROTEOMICS* 5.9 (June 2005), pp. 2319–2321. DOI: 10.1002/pmic.200401268.

[85] Joseph M Foster, Sven Degroeve, Laurent Gatto, Matthieu Visser, et al. "A Posteriori Quality Control for the Curation and Reuse of Public Proteomics Data". In: *PROTEOMICS* 11.11 (June 11, 2011), pp. 2182–2194. DOI: 10.1002/pmic.201000602.

[86] Frédéric Fournier, Charles Joly Beauparlant, René Paradis, and Arnaud Droit. "rTANDEM, an R/Bioconductor Package for MS/MS Protein Identification". In: *Bioinformatics* 30.15 (Aug. 1, 2014), pp. 2233–2234. DOI: 10.1093/bioinformatics/btu178.

[87] Ari M Frank, Matthew E Monroe, Anuj R Shah, Jeremy J Carver, et al. "Spectral Archives: Extending Spectral Libraries to Analyze Both Identified and Unidentified Spectra". In: *Nature Methods* 8.7 (May 15, 2011), pp. 587–591. DOI: 10.1038/nmeth.1609.

[88] Barbara E Frewen, Gennifer E Merrihew, Christine C Wu, William Stafford Noble, et al. "Analysis of Peptide MS/MS Spectra from Large-Scale Proteomics Experiments Using Spectrum Libraries". In: *Analytical Chemistry* 78.16 (Aug. 15, 2006), pp. 5678–5684. DOI: 10.1021/ac060279n.

[89] Barbara Frewen and Michael J MacCoss. "Using BiblioSpec for Creating and Searching Tandem MS Peptide Libraries". In: *Current Protocols in Bioinformatics* 20 (Dec. 2007), pp. 13.7.1–13.7.12. DOI: 10.1002/0471250953.bi1307s20.

[90] Himanshu S Gadgil, Gary D Pipes, Thomas M Dillon, Michael J Treuheit, et al. "Improving Mass Accuracy of High Performance Liquid Chromatography/Electrospray Ionization Time-of-Flight Mass Spectrometry of Intact Antibodies". In: *Journal of the American Society for Mass Spectrometry* 17.6 (June 2006), pp. 867–872. DOI: 10.1016/j.jasms.2006.02.023.

[91] Sebastien Gallien, Adele Bourmaud, and Bruno Domon. "A Simple Protocol to Routinely Assess the Uniformity of Proteomics Analyses". In: *Journal of Proteome Research* 13.5 (May 2, 2014), pp. 2688–2695. DOI: 10.1021/pr4011712.

[92] Laurent Gatto and Andy Christoforou. "Using R and Bioconductor for Proteomics Data Analysis". In: *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1844.1 (Jan. 2014), pp. 42–51. DOI: 10.1016/j.bbapap.2013.04.032.

[93] Laurent Gatto, Kasper D Hansen, Michael R Hoopmann, Henning Hermjakob, et al. "Testing and Validation of Computational Methods for Mass Spectrometry". In: *Journal of Proteome Research* 15.3 (Mar. 4, 2016), pp. 809–814. DOI: 10.1021/acs.jproteome.5b00852.

[94] Lewis Y Geer, Sanford P Markey, Jeffrey A Kowalak, Lukas Wagner, et al. "Open Mass Spectrometry Search Algorithm". In: *Journal of Proteome Research* 3.5 (Oct. 11, 2004), pp. 958–964. DOI: 10.1021/pr0499491.

[95] Scott J Geromanos, Johannes P C Vissers, Jeffrey C Silva, Craig A Dorschel, et al. "The Detection, Correlation, and Comparison of Peptide Precursor and Product Ions from Data Independent LC-MS with Data Dependant LC-MS/MS". In: *PROTEOMICS* 9.6 (Mar. 2009), pp. 1683–1695. DOI: 10.1002/pmic.200800562.

[96] Piero Giansanti, Liana Tsiatsiani, Teck Yew Low, and Albert J R Heck. "Six Alternative Proteases for Mass Spectrometry–based Proteomics beyond Trypsin". In: *Nature Protocols* 11.5 (Apr. 28, 2016), pp. 993–1006. DOI: 10.1038/nprot.2016.057.

[97] Helen G Gika, Georgios A Theodoridis, Mark Earll, and Ian D Wilson. "A QC Approach to the Determination of Day-to-Day Reproducibility and Robustness of LC–MS Methods for Global Metabolite Profiling in Metabonomics/Metabolomics". In: *Bioanalysis* 4.18 (Sept. 2012), pp. 2239–2247. DOI: 10.4155/bio.12.212.

[98] Ludovic C Gillet, Pedro Navarro, Stephen Tate, Hannes Röst, et al. "Targeted Data Extraction of the MS/MS Spectra Generated by Data-Independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis". In: *Molecular & Cellular Proteomics* 11.6 (June 1, 2012), O111.016717–O111.016717. DOI: 10.1074/mcp.O111.016717.

[99] Timo Glatter, Christina Ludwig, Erik Ahrné, Ruedi Aebersold, et al. "Large-Scale Quantitative Assessment of Different in-Solution Protein Digestion Protocols Reveals Superior Cleavage Efficiency of Tandem Lys-C/Trypsin Proteolysis over Trypsin Digestion". In: *Journal of Proteome Research* 11.11 (Nov. 2, 2012), pp. 5145–5156. DOI: 10.1021/pr300273g.

[100] Joanna Godzien, Vanesa Alonso-Herranz, Coral Barbas, and Emily Grace Armitage. "Controlling the Quality of Metabolomics Data: New Strategies to Get the Best out of the QC Sample". In: *Metabolomics* 11.3 (June 2015), pp. 518–528. DOI: 10.1007/s11306-014-0712-4.

[101] Miriam Goebel-Stengel, Andreas Stengel, Yvette Taché, and Joseph R Reeve. "The Importance of Using the Optimal Plasticware and Glassware in Studies Involving Peptides". In: *Analytical Biochemistry* 414.1 (July 1, 2011), pp. 38–46. DOI: 10.1016/j.ab.2011.02.009.

[102] Faviel F Gonzalez-Galarza, Craig Lawless, Simon J Hubbard, Jun Fan, et al. "A Critical Appraisal of Techniques, Software Packages, and Standards for Quantitative Proteomic Analysis". In: *OMICS: A Journal of Integrative Biology* 16.9 (Sept. 10, 2012), pp. 431–442. DOI: 10.1089/omi.2012.0022.

[103] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, et al. "Generative Adversarial Networks". In: *arXiv:1406.2661 [cs, stat]* (June 10, 2014). arXiv: 1406.2661.

[104] Google. *Better data centers through machine learning*. May 28, 2014. URL: https://googleblog.blogspot.com/2014/05/better-data-centers-through-machine.html (visited on 12/21/2016).

[105] Johannes Griss. "Spectral Library Searching in Proteomics". In: *PROTEOMICS* 16.5 (Mar. 2016), pp. 729–740. DOI: 10.1002/pmic.201500296.

[106] Johannes Griss, Joseph M Foster, Henning Hermjakob, and Juan Antonio Vizcaíno. "PRIDE Cluster: Building a Consensus of Proteomics Data". In: *Nature Methods* 10.2 (Jan. 30, 2013), pp. 95–96. DOI: 10.1038/nmeth.2343.

[107] Johannes Griss, Yasset Perez-Riverol, Steve Lewis, David L Tabb, et al. "Recognizing Millions of Consistently Unidentified Spectra across Hundreds of Shotgun Proteomics Datasets". In: *Nature Methods* 13.8 (June 27, 2016), pp. 651–656. DOI: 10.1038/nmeth.3902.

[108] P Hagel, J J T Gerding, W Fieggen, and H Bloemendal. "Cyanate Formation in Solutions of Urea". In: *Biochimica et Biophysica Acta (BBA) - Protein Structure* 243.3 (Sept. 28, 1971), pp. 366–373. DOI: 10.1016/0005-2795(71)90003-1.

[109] Alon Halevy, Peter Norvig, and Fernando Pereira. "The Unreasonable Effectiveness of Data". In: *IEEE Intelligent Systems* 24.2 (Mar. 2009), pp. 8–12. DOI: 10.1109/MIS.2009.36.

[110] Piliang Hao, Yan Ren, Andrew J Alpert, and Siu Kwan Sze. "Detection, Evaluation and Minimization of Nonenzymatic Deamidation in Proteomic Sample Preparation". In: *Molecular & Cellular Proteomics* 10.10 (Oct. 1, 2011), O111.009381–O111.009381. DOI: 10.1074/mcp.O111.009381.

[111] Kenneth Haug, Reza M Salek, Pablo Conesa, Janna Hastings, et al. "MetaboLights–an Open-Access General-Purpose Repository for Metabolomics Studies and Associated Meta-Data". In: *Nucleic Acids Research* 41 (D1 Jan. 2013), pp. D781–D786. DOI: 10.1093/nar/gks1004.

[112] Kenny Helsens, Mi-Youn Brusniak, Eric Deutsch, Robert L Moritz, et al. "jTraML: An Open Source Java API for TraML, the PSI Standard for Sharing SRM Transitions". In: *Journal of Proteome Research* 10.11 (Nov. 4, 2011), pp. 5260–5263. DOI: 10.1021/pr200664h.

[113] *Hibernate*. URL: http://hibernate.org/ (visited on 11/03/2014).

[114] Siri Hildonen, Trine Grønhaug Halvorsen, and Léon Reubsaet. "Why Less Is More When Generating Tryptic Peptides in Bottom-up Proteomics". In: *PROTEOMICS* 14 (17-18 Sept. 2014), pp. 2031–2041. DOI: 10.1002/pmic.201300479.

[115] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (Nov. 15, 1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[116] Kelly Hodge, Sara Ten Have, Luke Hutton, and Angus I Lamond. "Cleaning up the Masses: Exclusion Lists to Reduce Contamination with HPLC-MS/MS". In: *Journal of Proteomics* 88 (Aug. 2, 2013), pp. 92–103. DOI: 10.1016/j.jprot.2013.02.023.

[117] Stephen W Holman, Lynn McLean, and Claire E Eyers. "RePLiCal: A QconCAT Protein for Retention Time Standardization in Proteomics Studies". In: *Journal of Proteome Research* 15.3 (Mar. 4, 2016), pp. 1090–1102. DOI: 10.1021/acs.jproteome.5b00988.

[118] Oliver Horlacher, Frederique Lisacek, and Markus Müller. "Mining Large Scale Tandem Mass Spectrometry Data for Protein Modifications Using Spectral Libraries". In: *Journal of Proteome Research* 15.3 (Mar. 4, 2016), pp. 721–731. DOI: 10.1021/acs.jproteome.5b00877.

[119] Jianhua Hu, Kevin R Coombes, Jeffrey S Morris, and Keith A Baggerly. "The Importance of Experimental Design in Proteomic Mass Spectrometry Experiments: Some Cautionary Tales". In: *Briefings in Functional Genomics and Proteomics* 3.4 (Jan. 1, 2005), pp. 322–331. DOI: 10.1093/bfgp/3.4.322.

[120] Ting Huang, Jingjing Wang, Weichuan Yu, and Zengyou He. "Protein Inference: A Review". In: *Briefings in Bioinformatics* 13.5 (Sept. 1, 2012), pp. 586–614. DOI: 10.1093/bib/bbs004.

[121] Nicola C Hughes, Ernest Y K Wong, Juan Fan, and Navgeet Bajaj. "Determination of Carryover and Contamination for Mass Spectrometry-Based Chromatographic Assays". In: *The AAPS Journal* 9.3 (Sept. 2007), E353–E360. DOI: 10.1208/aapsj0903042.

[122] Piotr Indyk and Rajeev Motwani. "Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality". In: *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing - STOC '98*. Dallas, Texas, USA: ACM Press, 1998, pp. 604–613. DOI: 10.1145/276698.276876.

[123] Alexander R Ivanov, Christopher M Colangelo, Craig P Dufresne, David B Friedman, et al. "Interlaboratory Studies and Initiatives Developing Standards for Proteomics". In: *PROTEOMICS* 13.6 (Mar. 2013), pp. 904–909. DOI: 10.1002/pmic.201200532.

[124] *Java Architecture for XML Binding (JAXB)*. URL: https://jaxb.java.net/ (visited on 12/12/2013).

[125] *Java Persistence API*. URL: https://jcp.org/en/jsr/detail?id=317 (visited on 12/12/2013).

[126] Andrew R Jones, Martin Eisenacher, Gerhard Mayer, Oliver Kohlbacher, et al. "The mzIdentML Data Standard for Mass Spectrometry-Based Proteomics Results". In: *Molecular & Cellular Proteomics* 11.7 (July 1, 2012), pp. M111.014381–M111.014381. DOI: 10.1074/mcp.M111.014381.

[127] Johannes Junker, Chris Bielow, Andreas Bertsch, Marc Sturm, et al. "TOPPAS: A Graphical Workflow Editor for the Analysis of High-Throughput Proteomics Data". In: *Journal of Proteome Research* 11.7 (July 6, 2012), pp. 3914–3920. DOI: 10.1021/pr300187f.

[128] Yuri Kalambet, Yuri Kozmin, Ksenia Mikhailova, Igor Nagaev, et al. "Reconstruction of Chromatographic Peaks Using the Exponentially Modified Gaussian Function". In: *Journal of Chemometrics* 25.7 (July 2011), pp. 352–356. DOI: 10.1002/cem.1343.

[129] Lukas Käll, Jesse D Canterbury, Jason Weston, William Stafford Noble, et al. "Semi-Supervised Learning for Peptide Identification from Shotgun Proteomics Datasets". In: *Nature Methods* 4.11 (Oct. 21, 2007), pp. 923–925. DOI: 10.1038/nmeth1113.

[130] Anastasia Kalli and Sonja Hess. "Effect of Mass Spectrometric Parameters on Peptide and Protein Identification Rates for Shotgun Proteomic Experiments on an LTQ-Orbitrap Mass Analyzer". In: *PROTEOMICS* 12.1 (Jan. 2012), pp. 21–31. DOI: 10.1002/pmic.201100464.

[131] Anastasia Kalli, Geoffrey T Smith, Michael J Sweredoski, and Sonja Hess. "Evaluation and Optimization of Mass Spectrometric Settings during Data-Dependent Acquisition Mode: Focus on LTQ-Orbitrap Mass Analyzers". In: *Journal of Proteome Research* 12.7 (July 5, 2013), pp. 3071–3086. DOI: 10.1021/pr3011588.

[132] Oskar Karlsson, Jonas Bergquist, and Malin Andersson. "Quality Measures of Imaging Mass Spectrometry Aids in Revealing Long-Term Striatal Protein Changes Induced by Neonatal Exposure to the Cyanobacterial Toxin β-N-Methylamino-L-Alanine (BMAA)". In: *Molecular & Cellular Proteomics* 13.1 (Jan. 1, 2014), pp. 93–104. DOI: 10.1074/mcp.M113.031435.

[133] Jonathan A Karty, Marcia M E Ireland, Yves V Brun, and James P Reilly. "Artifacts and Unassigned Masses Encountered in Peptide Mass Mapping". In: *Journal of Chromatography B* 782 (1-2 Dec. 25, 2002), pp. 363–383. DOI: 10.1016/S1570-0232(02)00550-0.

[134] Borivoj Keil. *Specificity of Proteolysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1992.

[135] Andrew Keller, Alexey I Nesvizhskii, Eugene Kolker, and Ruedi Aebersold. "Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search". In: *Analytical Chemistry* 74.20 (Oct. 15, 2002), pp. 5383–5392. DOI: 10.1021/ac025747h.

[136] Bernd O Keller, Jie Sui, Alex B Young, and Randy M Whittal. "Interferences and Contaminants Encountered in Modern Mass Spectrometry". In: *Analytica Chimica Acta* 627.1 (Oct. 3, 2008). Ed. by David M Lubman, Patrick A Limbach, and Liang Li, pp. 71–81. DOI: 10.1016/j.aca.2008.04.043.

[137] Jong-Seo Kim, Matthew E Monroe, David G Camp, Richard D Smith, et al. "In-Source Fragmentation and the Sources of Partially Tryptic Peptides in Shotgun Proteomics". In: *Journal of Proteome Research* 12.2 (Feb. 1, 2013), pp. 910–916. DOI: 10.1021/pr300955f.

[138] Min-Sik Kim, Sneha M Pinto, Derese Getnet, Raja Sekhar Nirujogi, et al. "A Draft Map of the Human Proteome". In: *Nature* 509.7502 (May 28, 2014), pp. 575–581. DOI: 10.1038/nature13302.

[139] Min-Sik Kim, Jun Zhong, and Akhilesh Pandey. "Common Errors in Mass Spectrometry-Based Analysis of Post-Translational Modifications". In: *PROTEOMICS* 16.5 (Mar. 2016), pp. 700–714. DOI: 10.1002/pmic.201500355.

[140] Sangtae Kim and Pavel A Pevzner. "MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics". In: *Nature Communications* 5 (Oct. 31, 2014), p. 5277. DOI: 10.1038/ncomms6277.

[141] Christopher R Kinsinger, James Apffel, Mark Baker, Xiaopeng Bian, et al. "Recommendations for Mass Spectrometry Data Quality Metrics for Open Access Data (Corollary to the Amsterdam Principles)". In: *Molecular & Cellular Proteomics* 10.12 (Feb. 3, 2012), O111.015446–O111.015446. DOI: 10.1074/mcp.O111.015446.

[142] John Klimek, James S Eddes, Laura Hohmann, Jennifer Jackson, et al. "The Standard Protein Mix Database: A Diverse Data Set to Assist in the Production of Improved Peptide and Protein Identification Software Tools". In: *Journal of Proteome Research* 7.1 (Jan. 1, 2008), pp. 96–103. DOI: 10.1021/pr070244j.

[143] Thomas Köcher, Peter Pichler, Remco Swart, and Karl Mechtler. "Quality Control in LC-MS/MS". In: *PROTEOMICS* 11.6 (Mar. 6, 2011), pp. 1026–1030. DOI: 10.1002/pmic.201000578.

[144] Laxmikanth Kollipara and René P Zahedi. "Protein Carbamylation: In Vivo Modification or in Vitro Artefact?" In: *PROTEOMICS* 13.6 (Mar. 2013), pp. 941–944. DOI: 10.1002/pmic.201200452.

[145] Alexandra Kraut, Marlène Marcellin, Annie Adrait, Lauriane Kuhn, et al. "Peptide Storage: Are You Getting the Best Return on Your Investment? Defining Optimal Storage Conditions for Proteomics Samples". In: *Journal of Proteome Research* 8.7 (July 6, 2009), pp. 3778–3785. DOI: 10.1021/pr900095u.

[146] Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. "Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering". In: *ACM Transactions on Knowledge Discovery from Data* 3.1 (Mar. 2009), pp. 1–58. DOI: 10.1145/1497577.1497578.

[147] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F Pereira, C J C Burges, L Bottou, and K Q Weinberger. Curran Associates, Inc., 2012, pp. 1097–1105.

[148] Oleg V Krokhin, Mihaela Antonovici, Werner Ens, John A Wilkins, et al. "Deamidation of -Asn-Gly- Sequences during Sample Preparation for Proteomics: Consequences for MALDI and HPLC-MALDI Analysis". In: *Analytical Chemistry* 78.18 (Sept. 1, 2006), pp. 6645–6650. DOI: 10.1021/ac061017o.

[149] Henry Lam. "Building and Searching Tandem Mass Spectral Libraries for Peptide Identification". In: *Molecular & Cellular Proteomics* 10.12 (Dec. 1, 2011), R111.008565–R111.008565. DOI: 10.1074/mcp.R111.008565.

[150] Henry Lam and Ruedi Aebersold. "Building and Searching Tandem Mass (MS/MS) Spectral Libraries for Peptide Identification in Proteomics". In: *Methods* 54.4 (Aug. 2011), pp. 424–431. DOI: 10.1016/j.ymeth.2011.01.007.

[151] Henry Lam, Eric W Deutsch, and Ruedi Aebersold. "Artificial Decoy Spectral Libraries for False Discovery Rate Estimation in Spectral Library Searching in Proteomics". In: *Journal of Proteome Research* 9.1 (Jan. 4, 2010), pp. 605–610. DOI: 10.1021/pr900947u.

[152] Henry Lam, Eric W Deutsch, James S Eddes, Jimmy K Eng, et al. "Development and Validation of a Spectral Library Searching Method for Peptide Identification from MS/MS". In: *PROTEOMICS* 7.5 (Mar. 5, 2007), pp. 655–667. DOI: 10.1002/pmic.200600625.

[153] Henry Lam, Eric W Deutsch, James S Eddes, Jimmy K Eng, et al. "Building Consensus Spectral Libraries for Peptide Identification in Proteomics". In: *Nature Methods* 5.10 (Sept. 21, 2008), pp. 873–875. DOI: 10.1038/nmeth.1254.

[154] Lydie Lane, Amos Bairoch, Ronald C Beavis, Eric W Deutsch, et al. "Metrics for the Human Proteome Project 2013–2014 and Strategies for Finding Missing Proteins". In: *Journal of Proteome Research* 13.1 (Jan. 3, 2014), pp. 15–20. DOI: 10.1021/pr401144x.

[155] Dorothée Lebert, Mathilde Louwagie, Sandra Goetze, Guillaume Picard, et al. "DIGESTIF: A Universal Quality Standard for the Control of Bottom-up Proteomics Experiments". In: *Journal of Proteome Research* 14.2 (Feb. 6, 2015), pp. 787–803. DOI: 10.1021/pr500834z.

[156] Yann LeCun. *What are some recent and potentially upcoming breakthroughs in deep learning?* July 28, 2016. URL: https://www.quora.com/What-are-some-recent-and-potentially-upcoming-breakthroughs-in-deep-learning (visited on 12/21/2016).

[157] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning". In: *Nature* 521.7553 (May 27, 2015), pp. 436–444. DOI: 10.1038/nature14539.

[158] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, et al. "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network". In: *arXiv:1609.04802 [cs, stat]* (Sept. 15, 2016). arXiv: 1609.04802.

[159] Kimberly A Lee, Chris Farnsworth, Wen Yu, and Leo E Bonilla. "24-Hour Lock Mass Protection". In: *Journal of Proteome Research* 10.2 (Feb. 4, 2011), pp. 880–885. DOI: 10.1021/pr100780b.

[160] Pierre Legrain, Ruedi Aebersold, Alexander Archakov, Amos Bairoch, et al. "The Human Proteome Project: Current State and Future Direction". In: *Molecular & Cellular Proteomics* 10.7 (July 1, 2011), p. M111.009993. DOI: 10.1074/mcp.M111.009993.

[161] Daniel Lemire. *On the quality of academic software.* June 18, 2012. URL: http://lemire.me/blog/2012/06/18/on-the-quality-of-academic-software/ (visited on 12/21/2016).

[162] Kathryn S Lilley, Michael J Deery, and Laurent Gatto. "Challenges for Proteomics Core Facilities". In: *PROTEOMICS* 11.6 (Mar. 2011), pp. 1017–1025. DOI: 10.1002/pmic.201000693.

[163] Miao-Fang Lin, Christie Williams, Michael V Murray, Greg Conn, et al. "Ion Chromatographic Quantification of Cyanate in Urea Solutions: Estimation of the Efficiency of Cyanate Scavengers for Use in Recombinant Protein Manufacturing". In: *Journal of Chromatography B* 803.2 (Apr. 25, 2004), pp. 353–362. DOI: 10.1016/j.jchromb.2004.01.017.

[164] Philip L Loziuk, Jack Wang, Quanzi Li, Ronald R Sederoff, et al. "Understanding the Role of Proteolytic Digestion on Discovery and Targeted Proteomic Measurements Using Liquid Chromatography Tandem Mass Spectrometry and Design of Experiments". In: *Journal of Proteome Research* 12.12 (Dec. 6, 2013), pp. 5820–5829. DOI: 10.1021/pr4008442.

[165] Chun Wai Manson Ma and Henry Lam. "Hunting for Unexpected Post-Translational Modifications by Spectral Library Searching with Tier-Wise Scoring". In: *Journal of Proteome Research* 13.5 (May 2, 2014), pp. 2262–2271. DOI: 10.1021/pr401006g.

[166] Ze-Qiang Ma, Surendra Dasari, Matthew C Chambers, Michael D Litton, et al. "IDPicker 2.0: Improved Protein Assembly with High Discrimination Peptide Identification Filtering". In: *Journal of Proteome Research* 8.8 (Aug. 7, 2009), pp. 3872–3881. DOI: 10.1021/pr900360j.

[167] Ze-Qiang Ma, Kenneth O Polzin, Surendra Dasari, Matthew C Chambers, et al. "QuaMeter: Multivendor Performance Metrics for LC-MS/MS Proteomics Instrumentation". In: *Analytical Chemistry* 84.14 (July 17, 2012), pp. 5845–5850. DOI: 10.1021/ac300629p.

[168] Brendan MacLean, Daniela M Tomazela, Nicholas Shulman, Matthew Chambers, et al. "Skyline: An Open Source Document Editor for Creating and Analyzing Targeted Proteomics Experiments". In: *Bioinformatics* 26.7 (Apr. 1, 2010), pp. 966–968. DOI: 10.1093/bioinformatics/btq054.

[169] Evelyne Maes, Pieter Kelchtermans, Wout Bittremieux, Kurt De Grave, et al. "Designing Biomedical Proteomics Experiments: State-of-the-Art and Future Perspectives". In: *Expert Review of Proteomics* 13.5 (Apr. 25, 2016), pp. 495–511. DOI: 10.1586/14789450.2016.1172967.

[170] Sameh Magdeldin, James J Moresco, Tadashi Yamamoto, and John R Yates. "Off-Line Multidimensional Liquid Chromatography and Auto Sampling Result in Sample Loss in LC/LC–MS/MS". In: *Journal of Proteome Research* 13.8 (Aug. 1, 2014), pp. 3826–3836. DOI: 10.1021/pr500530e.

[171] Douglas W Mahoney, Terry M Therneau, Carrie J Heppelmann, LeeAnn Higgins, et al. "Relative Quantification: Characterization of Bias, Variability and Fold Changes in Mass Spectrometry Data from iTRAQ-Labeled Peptides". In: *Journal of Proteome Research* 10.9 (Sept. 2, 2011), pp. 4325–4333. DOI: 10.1021/pr2001308.

[172] Matthias Mann. "Comparative Analysis to Guide Quality Improvements in Proteomics". In: *Nature Methods* 6.10 (Oct. 2009), pp. 717–719. DOI: 10.1038/nmeth1009-717.

[173] Matthias Mann and Ole N Jensen. "Proteomic Analysis of Post-Translational Modifications". In: *Nature Biotechnology* 21.3 (Mar. 2003), pp. 255–261. DOI: 10.1038/nbt0303-255.

[174] J R Marier and Dyson Rose. "Determination of Cyanate, and a Study of Its Accumulation in Aqueous Solutions of Urea". In: *Analytical Biochemistry* 7.3 (Mar. 1964), pp. 304–314. DOI: 10.1016/0003-2697(64)90135-6.

[175] Gyorgy Marko-Varga, Gilbert S Omenn, Young-Ki Paik, and William S Hancock. "A First Step toward Completion of a Genome-Wide Characterization of the Human Proteome". In: *Journal of Proteome Research* 12.1 (Jan. 4, 2013), pp. 1–5. DOI: 10.1021/pr301183a.

[176] Lennart Martens. "A Report on the ESF Workshop on Quality Control in Proteomics". In: *Molecular BioSystems* 6.6 (2010), pp. 935–938. DOI: 10.1039/c003912h.

[177] Lennart Martens. "Bringing Proteomics into the Clinic: The Need for the Field to Finally Take Itself Seriously". In: *PROTEOMICS - Clinical Applications* 7 (5-6 June 2013), pp. 388–391. DOI: 10.1002/prca.201300020.

[178] Lennart Martens. "Public Proteomics Data: How the Field Has Evolved from Sceptical Inquiry to the Promise of in Silico Proteomics". In: *EuPA Open Proteomics* 11 (June 2016), pp. 42–44. DOI: 10.1016/j.euprot.2016.02.005.

[179] Lennart Martens, Matthew Chambers, Marc Sturm, Darren Kessner, et al. "mzML—a Community Standard for Mass Spectrometry Data". In: *Molecular & Cellular Proteomics* 10.1 (Jan. 1, 2011), R110.000133–R110.000133. DOI: 10.1074/mcp.R110.000133.

[180] Gerhard Mayer, Andrew R Jones, Pierre-Alain Binz, Eric W Deutsch, et al. "Controlled Vocabularies and Ontologies in Proteomics: Overview, Principles and Practice". In: *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1844.1 (Jan. 2014), pp. 98–107. DOI: 10.1016/j.bbapap.2013.02.017.

[181] Gerhard Mayer, Luisa Montecchi-Palazzi, David Ovelleiro, Andrew R Jones, et al. "The HUPO Proteomics Standards Initiative- Mass Spectrometry Controlled Vocabulary". In: *Database* 2013 (Feb. 19, 2013), bat009–bat009. DOI: 10.1093/database/bat009.

[182] W Hayes McDonald, David L Tabb, Rovshan G Sadygov, Michael J MacCoss, et al. "MS1, MS2, and SQT—three Unified, Compact, and Easily Parsed File Formats for the Storage of Shotgun Proteomic Spectra and Identifications". In: *Rapid Communications in Mass Spectrometry* 18.18 (Sept. 30, 2004), pp. 2162–2168. DOI: 10.1002/rcm.1603.

[183] Wes McKinney. "Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. Austin, Texas, USA, 2010, pp. 51–56.

[184] Doris Meder, Mònica Morales, Rainer Pepperkok, Ralph Schlapbach, et al. "Institutional Core Facilities: Prerequisite for Breakthroughs in the Life Sciences". In: *EMBO reports* 17.8 (Aug. 1, 2016), pp. 1088–1093. DOI: 10.15252/embr.201642857.

[185] Katalin F. Medzihradszky and Robert J Chalkley. "Lessons in de Novo Peptide Sequencing by Tandem Mass Spectrometry". In: *Mass Spectrometry Reviews* 34.1 (Jan. 2015), pp. 43–63. DOI: 10.1002/mas.21406.

[186] Barbora Micenková, Xuan-Hong Dang, Ira Assent, and Raymond T Ng. "Explaining Outliers by Subspace Separability". In: *Proceedings of the Thirteenth IEEE International Conference on Data Mining - ICDM '13*. Dallas, Texas, USA: IEEE, Dec. 7, 2013, pp. 518–527. DOI: 10.1109/ICDM.2013.132.

[187] Goran Mitulović, Christoph Stingl, Ines Steinmacher, Otto Hudecz, et al. "Preventing Carryover of Peptides and Proteins in Nano LC-MS Separations". In: *Analytical Chemistry* 81.14 (July 15, 2009), pp. 5955–5960. DOI: 10.1021/ac900696m.

[188] Luisa Montecchi-Palazzi, Samuel Kerrien, Florian Reisinger, Bruno Aranda, et al. "The PSI Semantic Validator: A Framework to Check MIAPE Compliance of Proteomics Data". In: *PROTEOMICS* 9.22 (Nov. 22, 2009), pp. 5112–5119. DOI: 10.1002/pmic.200900189.

[189] Luminita Moruz and Lukas Käll. "Peptide Retention Time Prediction". In: *Mass Spectrometry Reviews* (Early view Jan. 22, 2016). DOI: 10.1002/mas.21488.

[190] Thilo Muth, Bernhard Y. Renard, and Lennart Martens. "Metaproteomic Data Analysis at a Glance: Advances in Computational Microbial Community Proteomics". In: *Expert Review of Proteomics* 13.8 (Aug. 2, 2016), pp. 757–769. DOI: 10.1080/14789450.2016.1209418.

[191] Seungjin Na, Nuno Bandeira, and Eunok Paek. "Fast Multi-Blind Modification Search through Tandem Mass Spectrometry". In: *Molecular & Cellular Proteomics* 11.4 (Apr. 1, 2012), pp. M111.010199–M111.010199. DOI: 10.1074/mcp.M111.010199.

[192] Seungjin Na and Eunok Paek. "Software Eyes for Protein Post-Translational Modifications". In: *Mass Spectrometry Reviews* 34.2 (Apr. 2015), pp. 133–147. DOI: 10.1002/mas.21425.

[193] Stefan Naulaerts, Pieter Meysman, Wout Bittremieux, Trung Nghia Vu, et al. "A Primer to Frequent Itemset Mining for Bioinformatics". In: *Briefings in Bioinformatics* 16.2 (Mar. 2015), pp. 216–231. DOI: 10.1093/bib/bbt074.

[194] Pedro Navarro, Jörg Kuharev, Ludovic C Gillet, Oliver M Bernhardt, et al. "A Multicenter Study Benchmarks Software Tools for Label-Free Proteome Quantification". In: *Nature Biotechnology* 34 (Oct. 3, 2016), pp. 1130–1136. DOI: 10.1038/nbt.3685.

[195] Alexey I Nesvizhskii. "Proteogenomics: Concepts, Applications and Computational Strategies". In: *Nature Methods* 11.11 (Oct. 30, 2014), pp. 1114–1125. DOI: 10.1038/nmeth.3144.

[196] Alexey I Nesvizhskii, Olga Vitek, and Ruedi Aebersold. "Analysis and Validation of Proteomic Data Generated by Tandem Mass Spectrometry". In: *Nature Methods* 4.10 (Sept. 27, 2007), pp. 787–797. DOI: 10.1038/nmeth1088.

[197] William Stafford Noble and Michael J MacCoss. "Computational and Statistical Analysis of Protein Mass Spectrometry Data". In: *PLoS Computational Biology* 8.1 (Jan. 26, 2012). Ed. by Philip E Bourne, e1002296. DOI: 10.1371/journal.pcbi.1002296.

[198] Jesper V Olsen, Lyris M F de Godoy, Guoqing Li, Boris Macek, et al. "Parts per Million Mass Accuracy on an Orbitrap Mass Spectrometer via Lock Mass Injection into a C-Trap". In: *Molecular & Cellular Proteomics* 4.12 (Oct. 24, 2005), pp. 2010–2021. DOI: 10.1074/mcp.T500030-MCP200.

[199] Gilbert S Omenn, Lydie Lane, Emma K Lundberg, Ronald C Beavis, et al. "Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification". In: *Journal of Proteome Research* 14.9 (Sept. 4, 2015), pp. 3452–3460. DOI: 10.1021/acs.jproteome.5b00499.

[200] Gilbert S Omenn, Lydie Lane, Emma K Lundberg, Ronald C Beavis, et al. "Metrics for the Human Proteome Project 2016: Progress on Identifying and Characterizing the Human Proteome, Including Post-Translational Modifications". In: *Journal of Proteome Research* 15.11 (Nov. 4, 2016), pp. 3951–3960. DOI: 10.1021/acs.jproteome.6b00511.

[201] Ben Orsburn. *Case study – Beautiful looking data but NO peptide IDs!!!!* Mar. 31, 2016. URL: https://proteomicsnews.blogspot.com/2016/03/case-study-beautiful-looking-data-but.html (visited on 11/09/2016).

[202] Ben Orsburn. *Unsupervised quality control!* Mar. 20, 2016. URL: https://proteomicsnews.blogspot.com/2016/03/unsupervised-quality-control.html (visited on 07/24/2016).

[203] Eystein Oveland, Thilo Muth, Erdmann Rapp, Lennart Martens, et al. "Viewing the Proteome: How to Visualize Proteomics Data?" In: *PROTEOMICS* 15.8 (Apr. 2015), pp. 1341–1355. DOI: 10.1002/pmic.201400412.

[204] Andrew Palmer, Ekaterina Ovchinnikova, Mikael Thuné, Régis Lavigne, et al. "Using Collective Expert Judgements to Evaluate Quality Measures of Mass Spectrometry Images". In: *Bioinformatics* 31.12 (June 10, 2015), pp. i375–i384. DOI: 10.1093/bioinformatics/btv266.

[205] Ljiljana Pasa-Tolić, Christophe Masselon, Richard C Barry, Yufeng Shen, et al. "Proteomic Analyses Using an Accurate Mass and Time Tag Strategy". In: *BioTechniques* 37.4 (Oct. 2004), 621–624, 626–633, 636 passim.

[206] Amanda G Paulovich, Dean Billheimer, Amy-Joan L Ham, Lorenzo Vega-Montoto, et al. "Interlaboratory Study Characterizing a Yeast Performance Standard for Benchmarking LC-MS Platform Performance". In: *Molecular & Cellular Proteomics* 9.2 (Feb. 1, 2010), pp. 242–254. DOI: 10.1074/mcp.M900222-MCP200.

[207] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, et al. "Scikit-Learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (Oct 2011), pp. 2825–2830.

[208] Andrew J Percy, Andrew G Chambers, Derek S Smith, and Christoph H Borchers. "Standardized Protocols for Quality Control of MRM-Based Plasma Proteomic Workflows". In: *Journal of Proteome Research* 12.1 (Jan. 4, 2013), pp. 222–233. DOI: 10.1021/pr300893w.

[209] Andrew J Percy, Andrew G Chambers, Juncong Yang, Angela M Jackson, et al. "Method and Platform Standardization in MRM-Based Quantitative Plasma Proteomics". In: *Journal of Proteomics* 95 (Dec. 16, 2013), pp. 66–76. DOI: 10.1016/j.jprot.2013.07.026.

[210] Andrew J Percy, Carol E Parker, and Christoph H Borchers. "Pre-Analytical and Analytical Variability in Absolute Quantitative MRM-Based Plasma Proteomic Studies". In: *Bioanalysis* 5.22 (Nov. 2013), pp. 2837–2856. DOI: 10.4155/bio.13.245.

[211] Yasset Perez-Riverol, Emanuele Alpi, Rui Wang, Henning Hermjakob, et al. "Making Proteomics Data Accessible and Reusable: Current State of Proteomics Databases and Repositories". In: *PROTEOMICS* 15 (5-6 Mar. 2015), pp. 930–950. DOI: 10.1002/pmic.201400302.

[212] David N Perkins, Darryl J C Pappin, David M Creasy, and John S Cottrell. "Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data". In: *Electrophoresis* 20.18 (Dec. 1, 1999), pp. 3551–3567. DOI: 10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2.

[213] Brett Phinney. *This is why you do not let artificially smelly people stand next to the Q-Exactive.* Oct. 7, 2014. URL: http://proteomics.ucdavis.edu/2014/10/07/this-is-why-you-do-not-let-artificially-smelly-people-stand-by-the-q-exactive/ (visited on 11/08/2016).

[214] Brett Phinney. *Prelim data: The Promega Lys-C/Trypsin mix might be the culprit for our high missed cleavages. Courtesy of Anthony Herren in the core.* Oct. 25, 2016. URL: https://twitter.com/UCDProteomics/status/791037953300574208 (visited on 11/07/2016).

[215] Peter Pichler, Michael Mazanek, Frederico Dusberger, Lisa Weilnböck, et al. "SIMPATIQCO: A Server-Based Software Suite Which Facilitates Monitoring the Time Course of LC-MS Performance Metrics on Orbitrap Instruments". In: *Journal of Proteome Research* 11.11 (Nov. 2, 2012), pp. 5540–5547. DOI: 10.1021/pr300163u.

[216] Paola Picotti and Ruedi Aebersold. "Selected Reaction Monitoring–based Proteomics: Workflows, Potential, Pitfalls and Future Directions". In: *Nature Methods* 9.6 (May 30, 2012), pp. 555–566. DOI: 10.1038/nmeth.2015.

[217] Paul D Piehowski, Vladislav A Petyuk, Daniel J Orton, Fang Xie, et al. "Sources of Technical Variability in Quantitative LC–MS Proteomics: Human Brain Tissue Sample Analysis". In: *Journal of Proteome Research* 12.5 (May 3, 2013), pp. 2128–2137. DOI: 10.1021/pr301146m.

[218] Jennifer L Proc, Michael A Kuzyk, Darryl B Hardie, Juncong Yang, et al. "A Quantitative Study of the Effects of Chaotropic Agents, Surfactants, and Solvents on the Digestion Efficiency of Human Plasma Proteins by Trypsin". In: *Journal of Proteome Research* 9.10 (Oct. 1, 2010), pp. 5422–5437. DOI: 10.1021/pr100656u.

[219] Da Qi, Ritesh Krishna, and Andrew R Jones. "The jmzQuantML Programming Interface and Validator for the mzQuantML Data Standard". In: *PROTEOMICS* 14.6 (Mar. 2014), pp. 685–688. DOI: 10.1002/pmic.201300281.

[220] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2016.

[221] Janice Reimer, Vic Spicer, and Oleg V Krokhin. "Application of Modern Reversed-Phase Peptide Retention Prediction Algorithms to the Houghten and DeGraw Dataset: Peptide Helicity and Its Effect on Prediction Accuracy". In: *Journal of Chromatography A* 1256 (Sept. 21, 2012), pp. 160–168. DOI: 10.1016/j.chroma.2012.07.092.

[222] Florian Reisinger, Ritesh Krishna, Fawaz Ghali, Daniel Ríos, et al. "jmzIdentML API: A Java Interface to the mzIdentML Standard for Peptide and Protein Identification Data". In: *PROTEOMICS* 12.6 (Mar. 2012), pp. 790–794. DOI: 10.1002/pmic.201100577.

[223] Bernhard Y Renard, Marc Kirchner, Flavio Monigatti, Alexander R Ivanov, et al. "When Less Can Yield More - Computational Preprocessing of MS/MS Spectra for Peptide Identification". In: *PROTEOMICS* 9.21 (Nov. 2009), pp. 4978–4984. DOI: 10.1002/pmic.200900326.

[224] Bertrand Rochat, Emmanuel Kottelat, and Justin McMullen. "The Future Key Role of LC–high-Resolution-MS Analyses in Clinical Laboratories: A Focus on Quantification". In: *Bioanalysis* 4.24 (Dec. 2012), pp. 2939–2958. DOI: 10.4155/bio.12.243.

[225] Henry Rodriguez, Mike Snyder, Mathias Uhlén, Phil Andrews, et al. "Recommendations from the 2008 International Summit on Proteomics Data Release and Sharing Policy: The Amsterdam Principles". In: *Journal of Proteome Research* 8.7 (July 6, 2009), pp. 3689–3692. DOI: 10.1021/pr900023z.

[226] Hannes L Röst, Timo Sachsenberg, Stephan Aiche, Chris Bielow, et al. "OpenMS: A Flexible Open-Source Software Platform for Mass Spectrometry Data Analysis". In: *Nature Methods* 13.9 (Aug. 30, 2016), pp. 741–748. DOI: 10.1038/nmeth.3959.

[227] Paul A Rudnick, Karl R Clauser, Lisa E Kilpatrick, Dmitrii V Tchekhovskoi, et al. "Performance Metrics for Liquid Chromatography-Tandem Mass Spectrometry Systems in Proteomics Analyses". In: *Molecular & Cellular Proteomics* 9.2 (Feb. 1, 2010), pp. 225–241. DOI: 10.1074/mcp.M900223-MCP200.

[228] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision* 115.3 (Dec. 2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[229] Reza M Salek, Masanori Arita, Saravanan Dayalan, Timothy Ebbels, et al. "Embedding Standards in Metabolomics: The Metabolomics Society Data Standards Task Group". In: *Metabolomics* 11.4 (Aug. 2015), pp. 782–783. DOI: 10.1007/s11306-015-0821-8.

[230] Susanna-Assunta Sansone, Daniel Schober, Helen J Atherton, Oliver Fiehn, et al. "Metabolomics Standards Initiative: Ontology Working Group Work in Progress". In: *Metabolomics* 3.3 (Sept. 2007), pp. 249–256. DOI: 10.1007/s11306-007-0069-z.

[231] Richard A Scheltema and Matthias Mann. "SprayQc: A Real-Time LC-MS/MS Quality Monitoring System to Maximize Uptime Using off the Shelf Components". In: *Journal of Proteome Research* 11.6 (June 1, 2012), pp. 3458–3466. DOI: 10.1021/pr201219e.

[232] Oliver Serang and Lukas Käll. "Solution to Statistical Challenges in Proteomics Is More Statistics, Not Less". In: *Journal of Proteome Research* 14.10 (Oct. 2, 2015), pp. 4099–4103. DOI: 10.1021/acs.jproteome.5b00568.

[233] Wenguang Shao and Henry Lam. "Tandem Mass Spectral Libraries of Peptides and Their Roles in Proteomics Research". In: *Mass Spectrometry Reviews* (Early view July 12, 2016). DOI: 10.1002/mas.21512.

[234] Vagisha Sharma, Josh Eckels, Greg K Taylor, Nicholas J Shulman, et al. "Panorama: A Targeted Proteomics Knowledge Base". In: *Journal of Proteome Research* 13.9 (Sept. 5, 2014), pp. 4205–4210. DOI: 10.1021/pr5006636.

[235] Robbert J C Slebos, Xia Wang, Xaojing Wang, Bing Zhang, et al. "Proteomic Analysis of Colon and Rectal Carcinoma Using Standard and Customized Databases". In: *Scientific Data* 2 (June 23, 2015), p. 150022. DOI: 10.1038/sdata.2015.22.

[236] *SMAQC*. URL: https://github.com/PNNL-Comp-Mass-Spec/SMAQC (visited on 06/30/2016).

[237] Nico P M Smit, Irene van den Broek, Fred P H T M Romijn, Martin Haex, et al. "Quality Requirements for Quantitative Clinical Chemistry Proteomics". In: *Translational Proteomics* 2 (Mar. 2014), pp. 1–13. DOI: 10.1016/j.trprot.2013.10.001.

[238] Robert Smith, Dan Ventura, and John T Prince. "Novel Algorithms and the Benefits of Comparative Validation". In: *Bioinformatics* 29.12 (June 15, 2013), pp. 1583–1585. DOI: 10.1093/bioinformatics/btt176.

[239] *spotify/annoy: Approximate Nearest Neighbors in C++/Python optimized for memory usage and loading/saving to disk.*

[240] George R Stark. "Reactions of Cyanate with Functional Groups of Proteins. III. Reactions with Amino and Carboxyl Groups". In: *Biochemistry* 4.6 (June 1965), pp. 1030–1036. DOI: 10.1021/bi00882a008.

[241] George R Stark, William H Stein, and Stanford Moore. "Reactions of the Cyanate Present in Aqueous Urea with Amino Acids and Proteins". In: *Journal of Biological Chemistry* 235.11 (Nov. 1960), pp. 3177–3181.

[242] *Statistical Process Control in Proteomics*. URL: http://www.qcmylcms.com/ (visited on 06/30/2016).

[243] Hanno Steen and Matthias Mann. "The Abc's (and Xyz's) of Peptide Sequencing". In: *Nature Reviews Molecular Cell Biology* 5.9 (Sept. 2004), pp. 699–711. DOI: 10.1038/nrm1468.

[244] Stephen E Stein and Donald R Scott. "Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification". In: *Journal of the American Society for Mass Spectrometry* 5.9 (Sept. 1994), pp. 859–866. DOI: 10.1016/1044-0305(94)87009-8.

[245] Karel Stejskal, David Potěšil, and Zbyněk Zdráhal. "Suppression of Peptide Sample Losses in Autosampler Vials". In: *Journal of Proteome Research* 12.6 (June 7, 2013), pp. 3057–3062. DOI: 10.1021/pr400183v.

[246] Marc Sturm, Andreas Bertsch, Clemens Gröpl, Andreas Hildebrandt, et al. "OpenMS – An Open-Source Software Framework for Mass Spectrometry". In: *BMC Bioinformatics* 9.1 (Mar. 26, 2008), p. 163. DOI: 10.1186/1471-2105-9-163.

[247] Shisheng Sun, Jian-Ying Zhou, Weiming Yang, and Hui Zhang. "Inhibition of Protein Carbamylation in Urea Solution Using Ammonium-Containing Buffers". In: *Analytical Biochemistry* 446 (Feb. 1, 2014), pp. 76–81. DOI: 10.1016/j.ab.2013.10.024.

[248] Michael J Sweredoski, Geoffrey T Smith, Anastasia Kalli, Robert L J Graham, et al. "LogViewer: A Software Tool to Visualize Quality Control Parameters to Optimize Proteomics Experiments Using Orbitrap and LTQ-FT Mass Spectrometers". In: *Journal of Biomolecular Techniques* 22.4 (Dec. 2011), pp. 122–126. PMID: 22131886.

[249] David L Tabb. "Quality Assessment for Clinical Proteomics". In: *Clinical Biochemistry* 46.6 (Apr. 2013), pp. 411–420. DOI: 10.1016/j.clinbiochem.2012.12.003.

[250] David L Tabb, Christopher G Fernando, and Matthew C Chambers. "MyriMatch: Highly Accurate Tandem Mass Spectral Peptide Identification by Multivariate Hypergeometric Analysis". In: *Journal of Proteome Research* 6.2 (Feb. 2, 2007), pp. 654–661. DOI: 10.1021/pr0604054.

[251] David L Tabb, Lorenzo Vega-Montoto, Paul A Rudnick, Asokan Mulayath Variyath, et al. "Repeatability and Reproducibility in Proteomic Identifications by Liquid Chromatography-Tandem Mass Spectrometry". In: *Journal of Proteome Research* 9.2 (Feb. 5, 2010), pp. 761–776. DOI: 10.1021/pr9006365.

[252] David L Tabb, Xia Wang, Steven A Carr, Karl R Clauser, et al. "Reproducibility of Differential Proteomic Technologies in CPTAC Fractionated Xenografts". In: *Journal of Proteome Research* 15.3 (Mar. 4, 2016), pp. 691–706. DOI: 10.1021/acs.jproteome.5b00859.

[253] Chris F Taylor, Norman W Paton, Kathryn S Lilley, Pierre-Alain Binz, et al. "The Minimum Information about a Proteomics Experiment (MIAPE)". In: *Nature Biotechnology* 25.8 (Aug. 8, 2007), pp. 887–893. DOI: 10.1038/nbt1329.

[254] Ryan M Taylor, Jamison Dance, Russ J Taylor, and John T Prince. "Metriculator: Quality Assessment for Mass Spectrometry-Based Proteomics". In: *Bioinformatics* 29.22 (Nov. 15, 2013), pp. 2948–2949. DOI: 10.1093/bioinformatics/btt510.

[255] The UniProt Consortium. "UniProt: A Hub for Protein Information". In: *Nucleic Acids Research* 43 (D1 Jan. 28, 2015), pp. D204–D212. DOI: 10.1093/nar/gku989.

[256] Timothy K Toby, Luca Fornelli, and Neil L Kelleher. "Progress in Top-down Proteomics and the Analysis of Proteoforms". In: *Annual Review of Analytical Chemistry* 9.1 (June 12, 2016), pp. 499–519. DOI: 10.1146/annurev-anchem-071015-041550.

[257] Liana Tsiatsiani and Albert J R Heck. "Proteomics beyond Trypsin". In: *FEBS Journal* 282.14 (July 2015), pp. 2612–2626. DOI: 10.1111/febs.13287.

[258] Dekel Tsur, Stephen Tanner, Ebrahim Zandi, Vineet Bafna, et al. "Identification of Post-Translational Modifications by Blind Search of Mass Spectra". In: *Nature Biotechnology* 23.12 (Nov. 27, 2005), pp. 1562–1567. DOI: 10.1038/nbt1168.

[259] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data Using t-SNE". In: *Journal of Machine Learning Research* 9 (Nov 2008), pp. 2579–2605.

[260] Stéfan van der Walt, S Chris Colbert, and Gaël Varoquaux. "The NumPy Array: A Structure for Efficient Numerical Computation". In: *Computing in Science & Engineering* 13.2 (Mar. 2011), pp. 22–30. DOI: 10.1109/MCSE.2011.37.

[261] Paul M van Midwoud, Laurent Rieux, Rainer Bischoff, Elisabeth Verpoorte, et al. "Improvement of Recovery and Repeatability in Liquid Chromatography-Mass Spectrometry Analysis of Peptides". In: *Journal of Proteome Research* 6.2 (Feb. 1, 2007), pp. 781–791. DOI: 10.1021/pr0604099.

[262] Elien Vandermarliere, Michael Mueller, and Lennart Martens. "Getting Intimate with Trypsin, the Leading Protease in Proteomics". In: *Mass Spectrometry Reviews* 32.6 (Nov. 2013), pp. 453–465. DOI: 10.1002/mas.21376.

[263] Marc Vaudel, Julia M Burkhart, Albert Sickmann, Lennart Martens, et al. "Peptide Identification Quality Control". In: *PROTEOMICS* 11.10 (May 2011), pp. 2105–2114. DOI: 10.1002/pmic.201000704.

[264] Marc Vaudel, Kenneth Verheggen, Attila Csordas, Helge Raeder, et al. "Exploring the Potential of Public Proteomics Data". In: *PROTEOMICS* 16.2 (Jan. 2016), pp. 214–225. DOI: 10.1002/pmic.201500295.

[265] Kenneth Verheggen and Lennart Martens. "Ten Years of Public Proteomics Data: How Things Have Evolved, and Where the next Ten Years Should Lead Us". In: *EuPA Open Proteomics* 8 (Sept. 2015), pp. 28–35. DOI: 10.1016/j.euprot.2015.07.014.

[266] Juan A Vizcaíno, Eric W Deutsch, Rui Wang, Attila Csordas, et al. "ProteomeXchange Provides Globally Coordinated Proteomics Data Submission and Dissemination". In: *Nature Biotechnology* 32.3 (Mar. 2014), pp. 223–226. DOI: 10.1038/nbt.2839.

[267] Juan Antonio Vizcaíno, Attila Csordas, Noemi del-Toro, José A Dianes, et al. "2016 Update of the PRIDE Database and Its Related Tools". In: *Nucleic Acids Research* 44 (D1 Jan. 4, 2016), pp. D447–D456. DOI: 10.1093/nar/gkv1145.

[268] Juan Antonio Vizcaíno, Lennart Martens, Henning Hermjakob, Randall K Julian, et al. "The PSI Formal Document Process and Its Implementation on the PSI Website". In: *PROTEOMICS* 7.14 (July 14, 2007), pp. 2355–2357. DOI: 10.1002/pmic.200700064.

[269] Trung Nghia Vu, Wout Bittremieux, Dirk Valkenborg, Bart Goethals, et al. "Efficient Reduction of Candidate Matches in Peptide Spectrum Library Searching Using the Top k Most Intense Peaks". In: *Journal of Proteome Research* 13.9 (Sept. 5, 2014), pp. 4175–4183. DOI: 10.1021/pr401269z.

[270] Scott J Walmsley, Paul A Rudnick, Yuxue Liang, Qian Dong, et al. "Comprehensive Analysis of Protein Digestion Using Six Trypsins Reveals the Origin of Trypsin as a Significant Source of Variability in Proteomics". In: *Journal of Proteome Research* 12.12 (Dec. 6, 2013), pp. 5666–5680. DOI: 10.1021/pr400611h.

[271] Christopher T Walsh, Sylvie Garneau-Tsodikova, and Gregory J Gatto. "Protein Posttranslational Modifications: The Chemistry of Proteome Diversifications". In: *Angewandte Chemie International Edition* 44.45 (Nov. 18, 2005), pp. 7342–7372. DOI: 10.1002/anie.200501023.

[272] Mathias Walzer, Lucia Espona Pernas, Sara Nasso, Wout Bittremieux, et al. "qcML: An Exchange Format for Quality Control Metrics from Mass Spectrometry Experiments". In: *Molecular & Cellular Proteomics* 13.8 (Aug. 1, 2014), pp. 1905–1913. DOI: 10.1074/mcp.M113.035907.

[273] Xia Wang, Matthew C Chambers, Lorenzo J Vega-Montoto, David M Bunk, et al. "QC Metrics from CPTAC Raw LC-MS/MS Data Interpreted through Multivariate Statistics". In: *Analytical Chemistry* 86.5 (Mar. 4, 2014), pp. 2497–2509. DOI: 10.1021/ac4034455.

[274] Ralf J M Weber, Eva Li, Jonathan Bruty, Shan He, et al. "MaConDa: A Publicly Accessible Mass Spectrometry Contaminants Database". In: *Bioinformatics* 28.21 (Nov. 1, 2012), pp. 2856–2857. DOI: 10.1093/bioinformatics/bts527.

[275] Bo Wen and Laurent Gatto. *proteoQC: An R package for proteomics data quality control*. R package version 1.6.0. 2016.

[276] Mathias Wilhelm, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, et al. "Mass-Spectrometry-Based Draft of the Human Proteome". In: *Nature* 509.7502 (May 28, 2014), pp. 582–587. DOI: 10.1038/nature13319.

[277] John S Williams, Stephanie H Donahue, Hong Gao, and Christopher L Brummel. "Universal LC–MS Method for Minimized Carryover in a Discovery Bioanalytical Setting". In: *Bioanalysis* 4.9 (May 2012), pp. 1025–1037. DOI: 10.4155/bio.12.76.

[278] Catherine C L Wong, Daniel Cociorva, Christine A Miller, Alexander Schmidt, et al. "Proteomics of Pyrococcus Furiosus (Pfu): Identification of Extracted Proteins by Three Independent Methods". In: *Journal of Proteome Research* 12.2 (Feb. 1, 2013), pp. 763–770. DOI: 10.1021/pr300840j.

[279] Alisa G Woods, Izabela Sokolowska, and Costel C Darie. "Identification of Consistent Alkylation of Cysteine-Less Peptides in a Proteomics Experiment". In: *Biochemical and Biophysical Research Communications* 419.2 (Mar. 9, 2012), pp. 305–308. DOI: 10.1016/j.bbrc.2012.02.016.

[280] Ian Wright and Jennifer E Van Eyk. "A Roadmap to Successful Clinical Proteomics". In: *Clinical Chemistry* (Nov. 2016). DOI: 10.1373/clinchem.2016.254664.

[281] Ping Xu, Duc M Duong, and Junmin Peng. "Systematical Optimization of Reverse-Phase Chromatography for Shotgun Proteomics". In: *Journal of Proteome Research* 8.8 (Aug. 7, 2009), pp. 3944–3950. DOI: 10.1021/pr900251d.

[282] Qing-Wei Xu, Johannes Griss, Rui Wang, Andrew R Jones, et al. "jmzTab: A Java Interface to the mzTab Data Standard". In: *PROTEOMICS* 14.11 (June 2014), pp. 1328–1332. DOI: 10.1002/pmic.201300560.

[283] *XXIndex - a library to index XML files for random access*. URL: http://code.google.com/p/pride-toolsuite/wiki/XXIndex (visited on 12/12/2013).

[284] John R III Yates, Sung Kyu Robin Park, Claire M Delahunty, Tao Xu, et al. "Toward Objective Evaluation of Proteomic Algorithms". In: *Nature Methods* 9.5 (Apr. 27, 2012), pp. 455–456. DOI: 10.1038/nmeth.1983.

[285] Ding Ye, Yan Fu, Rui-Xiang Sun, Hai-Peng Wang, et al. "Open MS/MS Spectral Library Search to Identify Unanticipated Post-Translational Modifications and Increase Spectral Identification Rate". In: *Bioinformatics* 26.12 (June 15, 2010), pp. i399–i406. DOI: 10.1093/bioinformatics/btq185.

[286] Chia-Yu Yen, Stephane Houel, Natalie G Ahn, and William M Old. "Spectrum-to-Spectrum Searching Using a Proteome-Wide Spectral Library". In: *Molecular & Cellular Proteomics* 10.7 (July 1, 2011), pp. M111.007666–M111.007666. DOI: 10.1074/mcp.M111.007666.

[287] Chia-Yu Yen, Karen Meyer-Arendt, Brian Eichelberger, Shaojun Sun, et al. "A Simulated MS/MS Library for Spectrum-to-Spectrum Searching in Large Scale Identification of Proteins". In: *Molecular & Cellular Proteomics* 8.4 (Apr. 1, 2009), pp. 857–869. DOI: 10.1074/mcp.M800384-MCP200.

[288] Bing Zhang, Jing Wang, Xiaojing Wang, Jing Zhu, et al. "Proteogenomic Characterization of Human Colon and Rectal Cancer". In: *Nature* 513.7518 (July 20, 2014), pp. 382–387. DOI: 10.1038/nature13438.

[289] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, et al. "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks". In: *arXiv:1612.03242 [cs, stat]* (Dec. 9, 2016). arXiv: 1612.03242.

[290] Xin Zhang, Yunzi Li, Wenguang Shao, and Henry Lam. "Understanding the Improved Sensitivity of Spectral Library Searching over Sequence Database Searching in Proteomics Data Analysis". In: *PROTEOMICS* 11.6 (Mar. 6, 2011), pp. 1075–1085. DOI: 10.1002/pmic.201000492.

[291] Ying Zhang, Aivett Bilbao, Tobias Bruderer, Jeremy Luban, et al. "The Use of Variable Q1 Isolation Windows Improves Selectivity in LC–SWATH–MS Acquisition". In: *Journal of Proteome Research* 14.10 (Oct. 2, 2015), pp. 4359–4371. DOI: 10.1021/acs.jproteome.5b00543.

[292] Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. "A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data". In: *Statistical Analysis and Data Mining* 5.5 (Oct. 2012), pp. 363–387. DOI: 10.1002/sam.11161.

[293] Roman A Zubarev. "The Challenge of the Proteome Dynamic Range and Its Implications for in-Depth Proteomics". In: *PROTEOMICS* 13.5 (Mar. 2013), pp. 723–726. DOI: 10.1002/pmic.201200451.