

INAUGURAL-DISSERTATION

ZUR
ERLANGUNG DER DOKTORWÜRDE
DER
NATURWISSENSCHAFTLICH-MATHEMATISCHEN GESAMTFAKULTÄT
DER
RUPRECHT-KARLS-UNIVERSITÄT
HEIDELBERG

vorgelegt von
Dipl.-Math. Anna Kreshuk
aus Moskau, Russland

Tag der mündlichen Prüfung: _____

Automated Analysis of Biomedical Data from Low to High Resolution

Advisor: Prof. Dr. Fred A. Hamprecht

Abstract

Recent developments of experimental techniques and instrumentation allow life scientists to acquire enormous volumes of data at unprecedented resolution. While this new data brings much deeper insight into cellular processes, it renders manual analysis infeasible and calls for the development of new, automated analysis procedures. This thesis describes how methods of pattern recognition can be used to automate three popular data analysis protocols:

- **Chapter 3** proposes a method to automatically locate bimodal isotope distribution patterns in Hydrogen Deuterium Exchange Mass Spectrometry experiments. The method is based on L1-regularized linear regression and allows for easy quantitative analysis of co-populations with different exchange behavior. The sensitivity of the method is tested on a set of manually identified peptides, while its applicability to exploratory data analysis is validated by targeted follow-up peptide identification.
- **Chapter 4** develops a technique to automate peptide quantification for mass spectrometry experiments, based on $^{16}O/^{18}O$ labeling of peptides. Two different spectrum segmentation algorithms are proposed: one based on image processing and applicable to low resolution data and one exploiting the sparsity of high resolution data. The quantification accuracy is validated on calibration datasets, produced by mixing a set of proteins in pre-defined ratios.
- **Chapter 5** provides a method for automated detection and segmentation of synapses in electron microscopy images of neural tissue. For images acquired by scanning electron microscopy with nearly isotropic resolution, the algorithm is based on geometric features computed in 3D pixel neighborhoods. For transmission electron microscopy images with poor z-resolution, the algorithm uses additional regularization by performing several rounds of pixel classification with features computed on the probability maps of the previous classification round. The validation is performed by comparing the set of synapses detected by the algorithm against a gold standard detection by human experts. For data with nearly isotropic resolution, the algorithm performance is comparable to that of the human experts.

Zusammenfassung

Neueste Entwicklungen in Experimentiertechnik und Gerätschaften erlauben Biowissenschaftlern sehr große Datenmengen in nie dagewesener Auflösung aufzunehmen. Diese Daten vertiefen unser Verständnis von Zellprozessen, machen aber auch eine manuelle Analyse unmöglich und fordern die Entwicklung neuer, automatischer Analysemethoden. Die vorliegende Arbeit beschreibt Mustererkennungsmethoden zur Automatisierung von drei gängigen Datenanalyseprotokollen:

- **In Kapitel 3** wird eine Methode zur automatischen Lokalisierung bimodaler Isotopenverteilungen in Deuterium-Austausch Massenspektrometrie-Experimenten vorgeschlagen. Sie basiert auf L1-regularisierter linearer Regression und erlaubt eine einfache quantitative Analyse von Ko-Populationen mit unterschiedlichen Austauschverhalten. Die Empfindlichkeit der Methode wurde auf manuell identifizierten Peptiden getestet, während ihre Anwendbarkeit in der explorativen Datenanalyse mit Hilfe gezielter Folgepeptid Identifikation gezeigt wurde.
- **In Kapitel 4** wird eine Technik zur automatischen Peptidquantifizierung für Massenspektrometrie-Experimente entwickelt, die auf $^{16}\text{O}/^{18}\text{O}$ Peptidmarkern beruht. Es werden zwei verschiedene spektrale Segmentierungsalgorithmen vorgeschlagen: Der eine basiert auf Bildverarbeitungstechniken und ist auf niedrig aufgelöste Daten anwendbar; der andere nutzt die Sparsity von hochaufgelösten Daten aus. Die Quantifizierungsgenauigkeit wird auf Kalibrierungsdatensätzen überprüft, die auf Proteinen mit vordefinierten Mischungsverhältnissen aufgenommen wurden.
- **In Kapitel 5** wird eine Methode zur automatischen Erkennung und Segmentierung von Synapsen in elektronenmikroskopischen Aufnahmen von neuronalem Gewebe vorgestellt. Für Bilder, die mit Hilfe von Transmissionselektronenmikroskopie aufgenommen wurden basiert der Algorithmus auf geometrischen Merkmalen, die in einer 3D Nachbarschaft von Pixeln berechnet worden sind. Für Transmissionselektronenmikroskopie-Bilder mit schlechter Z-Auflösung benutzt der Algorithmus eine zusätzliche Regularisierung indem mehrere Male eine Pixelklassifizierung durchgeführt wird, die Merkmale von Wahrscheinlichkeitskarten aus vorangegangenen Klassifizierungen mit einbezieht. Die Methode wird validiert indem die vom Algorithmus gefundenen Synapsen mit einem Gold Standard verglichen werden, der von menschlichen Experten erstellt worden ist. Auf Daten mit annähernd isotroper Auflösung sind die Ergebnisse des Algorithmus mit denen menschlicher Experten vergleichbar.

Acknowledgments

First and foremost I would like to thank Prof. Dr. Fred Hamprecht for supervising the research which led to this thesis. Prof. Hamprecht introduced me to the field of pattern recognition and its exciting applications and created a very friendly and stimulating scientific work environment with ample opportunities for interdisciplinary collaborations. I am very grateful for his support, guidance, passion for science and sense of humor, which made my work so enjoyable and my motivation so high in the last years.

I have also greatly benefited from working alongside other members of the multidimensional image processing (MIP) group. I thank Dr. Ulrich Köthe for the multiple discussions we had, for his readiness to share his extensive expertise and to talk about science, programming and life. I'm grateful to the old mass spectrometry group of Dr. Marc Kirchner, Dr. Bernhard Renard, Dr. Michael Hanselmann, Dr. Xinghua Lou and Bernhard Kausler for opening the field of mass spectrometry to me, for answering the many questions I had and in general for being such a great group of people to work and hang out with. I am especially grateful to Dr. Marc Kirchner for our close collaboration.

I thank Bernhard Kausler, who started in the group at the same time as I did, for challenging what seemed obvious, for always getting to the core of things, for being a great officemate and friend. I thank Thorben Kröger for always making my life easier with his software. Along with the rest of the ilastik team: Dr. Christoph Sommer, Christoph Straehle and Luca Fiaschi, I also thank him for introducing me to the fun of Python programming, for all the surprising ideas, debugging sessions and the good times we had together. I thank Dr. Björn Andres for his advice, wisdom and motivating discourse, and Barbara Werner for her help and support in all things administrative. Besides, I thank all other members of the MIP group: Dr. Frederik Kaster, Rahul Nair, Nathan Huesken, Martin Lindner, Martin Riedl, Buote Xu, Darya Trofimova, Svenja Reich, Thorsten Beier, Oliver Petra and Joachim Schleicher for the excellent group atmosphere and interesting discussions.

During the last three years I had the good fortune to collaborate with several outstanding life scientists. I am grateful to Prof. Dr. Wolf-Dieter Lehmann for providing the data of my first project and for introducing me to the application side of the analysis of mass spectrometry data. I thank Prof. Dr. Hanno Steen and the members of his group, especially Dr. Dominic Winter and Dr. Wiebke Timm, for the Orbitrap data and for their hospitality in Boston. I thank Dr. Matthias Mayer for offering me the bimodal isotope distribution analysis problem, for providing the H/D exchange data and for patiently explaining me the biochemistry of H/D exchange experiments. I am very grateful to Dr. Graham Knott and Dr. Natalya Korogod for our fruitful collaboration, for their enthusiasm for automated processing and for the answers to so many questions on neuroscience and electron microscopy.

On the financial side, I am very grateful for the financial support of the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences (HGS) as well as for the Clearingstelle position provided by the Equal Opportunities office of Heidelberg University during my maternity leave.

Finally, I would like to thank my family, especially my parents, my grandparents and my brother, for always being there for me despite the geographical distance between us and for unanimously supporting me in my decision to do a PhD after a five-year break in studies. And of course, I am most grateful to my husband Marian for his constant help, support and encouragement, for our discussions on science and methodology of research and in general for making me so happy all those years. And to my son Martin, for being such a well-behaved boy and letting me finish.

To my family.

Contents

1	Prologue	5
2	Introduction	7
2.1	Proteomics and Mass Spectrometry	7
2.1.1	Experimental scheme and instrumentation	9
2.1.2	Protein identification	13
2.1.3	Protein quantification	13
2.2	Neural circuit reconstruction and Electron Microscopy	14
3	Quantification of Bimodal Isotope Peak Distributions in H/D Exchange Mass Spectrometry	19
3.1	Introduction	19
3.1.1	H/D Exchange Mass Spectrometry	20
3.1.2	Motivation for Analysis of Bimodal Isotope Peak Distributions	24
3.2	Methods	25
3.2.1	Overview of the HeXicon software	25
3.2.2	Tuning HeXicon to search for bimodal distributions	27
3.3	Experiments	29
3.3.1	Data acquisition	29
3.3.2	Retention time alignment	30
3.4	Results	32
3.4.1	Proof of principle	32
3.4.2	Analysis of a full dataset	35
3.4.3	New insight into HtpG protein dynamics	35
3.5	Discussion	36
3.5.1	Analysis of false positive candidates	36
4	Automated Quantification of $^{16}O/^{18}O$-labeled LC/MS Data	39

4.1	Introduction	39
4.1.1	$^{16}\text{O}/^{18}\text{O}$ labeling	40
4.2	Methods	44
4.2.1	Segmentation	44
4.2.2	Modeling and Quantification	52
4.3	Experiments and Results	57
4.3.1	Low resolution	57
4.3.2	High resolution	58
4.4	Discussion	58
5	Automated Detection and Segmentation of Synapses in Serial Electron Microscopy Images	63
5.1	Introduction	63
5.1.1	Chemical synapse and information transmission in the central nervous system	63
5.1.2	Synapse detection	65
5.1.3	Related work	67
5.2	Methods	68
5.2.1	Random Forest classifier	68
5.2.2	Feature selection	71
5.2.3	Probability map thresholding	73
5.2.4	Software	75
5.3	Experiments	76
5.3.1	Data acquisition and generation of the gold standard	76
5.3.2	Error criteria	78
5.4	Results	78
5.4.1	Human experts	78
5.4.2	Automated detection	79
5.5	Discussion	80
5.6	Outlook: synapse detection in non-isotropic images	85
5.6.1	Introduction	85
5.6.2	Feature dimensionality	85
5.6.3	Context and auto-context	87
5.6.4	Anisotropic features	90
5.6.5	Experiments	90
5.6.6	Preliminary results and Discussion	92
6	Final Discussion and Outlook	99
6.1	Quantification of Bimodal Isotope Peak Distributions in H/D Exchange Mass Spectrometry	99
6.2	Automated Quantification of $^{16}\text{O}/^{18}\text{O}$ -labeled LC/MS Data	100

6.3 Automated Detection and Segmentation of Synapses in Serial Electron Microscopy Images	101
Frequently used Abbreviations	105
Lists of Tables	107
Lists of Figures	109
List of Publications	111
Bibliography	113

Chapter 1

Prologue

In the recent years we have witnessed a rapid increase in the volume of data produced by biomedical experiments. This trend has been most noticeable in the field of genomics, where technical improvement of the sequencing technology has been following Moore's law [123] and now allows to produce as much sequence information in minutes as in the first five years of the human genome project. The development of new high-throughput instrumentation is however not limited to genomics and is now paramount in many areas of biomedical research. From the point of view of computer science, these advancements introduce new challenges in data processing technology, requiring more and more automation not only in the data acquisition process, but also in the experiment data analysis.

The field of proteomics studies the totality of the proteins of an organism, including their post-translational modifications and interactions, as well as their role in the development of disease. The rapid advances in the instrumentation of proteomics in the last 20 years have expanded the research goals of a single experiment from obtaining the sequence of a few proteins to qualitatively and quantitatively assessing the complete proteome of a simple organism [31]. Currently, most of the proteomics experiment data is obtained by Mass Spectrometry. Even low resolution mass spectrometers can now generate amounts of data which are very difficult to analyze manually. The cutting-edge high resolution mass spectrometers render manual analysis infeasible and require automated means of protein quantification and comparison. Besides simplification or complete replacement of manual analysis techniques targeting selected proteins, automated analysis methods allow for fast exploratory scanning of huge data volumes for interesting patterns of changes in protein expression levels. This thesis presents analysis algorithms for two types of proteomics experiments:

- Chapter 3 introduces a method to automatically find sub-populations of different exchange behavior in a protein sample for Hydrogen Deuterium Exchange experi-

ments. It is based on the HeXicon algorithm for deuteration distribution estimation [90], which was modified to search specifically for bimodal distribution patterns and on the general NITPICK peak picking procedure [124], which was tuned to handle the case of mixtures of very correlated peptides.

- Chapter 4 develops an automated quantification approach for $^{16}\text{O}/^{18}\text{O}$ stable isotope labeling experiments. It introduces two different methods for the segmentation of the data: a watershed-based algorithm for low resolution mass spectra - a modified version of the segmentation from [90] - and a novel sparse technique for high resolution mass spectra.

Neural circuit imaging and reconstruction can serve as a further example of a biological domain with unprecedented instrumentation advancements in the recent years. Super-resolution methods, bypassing the diffraction limits, have been introduced in light microscopy [130, 10, 102]. Automation of nervous tissue slice handling now allows for imaging of very large tissue stacks by Transmission Electron Microscopy [57], while the development of new tissue block cutting and milling techniques enabled the use of Scanning Electron Microscopy and brought the native isotropic resolution of the image stacks to $3\times 3\times 3$ nanometers [34, 80]. New data analysis methods have been introduced aiming to partially or even fully automate the processing of the acquired images. A method of this type, offering automated detection and segmentation of synaptic contacts in serial electron microscopy images, is presented in this thesis:

- Chapter 5 introduces a protocol for automated detection and segmentation of asymmetric synapses in isotropic image stacks, produced by focused ion beam/scanning electron microscope. The procedure is based on interactive machine learning and only requires a few labeled synapses for training. The statistical learning is performed on geometrical features of 3D neighborhoods of each voxel and can fully exploit the high z-resolution of the data. On a quantitative validation dataset the error rate of the algorithm was found to be comparable to that of the human experts.
- The Outlook section of Chapter 5 explores several strategies for the extension of the automated synapse detection procedure to image stacks with poor z-resolution produced by serial section Transmission Electron Microscopy.

Chapter 2

Introduction

2.1 Proteomics and Mass Spectrometry

Proteins play a leading role in most biological processes and serve as the main functional unit of the cell. While the genome of an organism encodes all the information on its potential development, it is its proteome - the ensemble of all its proteins - that determines the processes which are actually happening and their rate. Following the rapid development of genomics and the success of the human genome project, proteomics has emerged as the study of the proteome in its widest sense, including not only the abundances or expression levels of all the organism's proteins, but also proteins' function, protein-protein interactions and the study of higher-order protein complexes [152].

Proteins are built from sequences of amino acids, held together by peptide bonds. A shorter amino acid sequence or a fragment of a protein can also be called a peptide. The free amine group of the amino acid on one end of the peptide chain forms the protein N-terminus, while the free carboxyl group on the other end forms the C-terminus. The proteins get translated from messenger RNA from their N-terminus to the C-terminus and, by convention, protein sequences are also listed from the N-terminus to the C-terminus. Fig.2.1 shows a short peptide of just three amino acids.

In the early days of proteomics 2D gel electrophoresis played the role of the leading experimental technique. It allowed to separate proteins in the sample by two characteristics: charge and relative molecular mass. Sequences of the separated proteins could then be found by applying a chemical technique known as Edman degradation [40, 39] to their N-terminus or to the digested protein fragments [118]. Fragmenting a peptide by Edman degradation could take hours or even days. Besides, the reaction requires a large amount of the protein and poses high requirements to its purity [144].

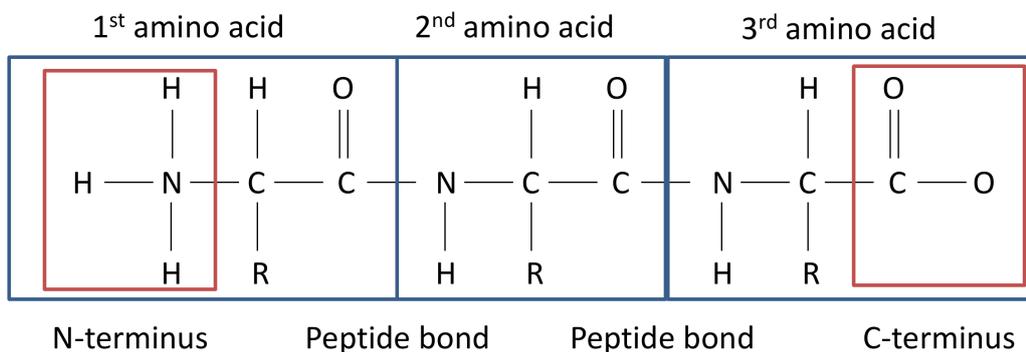


Figure 2.1: A short peptide example.

Although the popularity of mass spectrometry as an analytical technique in chemistry and physics goes back to early 20th century, lack of efficient ionization techniques for large molecules - such as proteins and peptides - has delayed its universal adaptation in biology. This barrier was finally removed in the late 1980s by the pioneering works of John Fenn [46] and Koichi Tanaka [147] on Electrospray Ionization (ESI) and Matrix-Assisted Laser Desorption/Ionization (MALDI). Fenn and Tanaka received the Nobel Prize in Chemistry in 2002 for "their development of soft desorption ionisation methods for mass spectrometric analyses of biological macromolecules" [111]. Combined with ESI or MALDI, mass spectrometry can fragment proteins in seconds, while only requiring femtomole amounts of sample. Furthermore, it does not require homogeneously purified proteins and is not hindered by their post-translational modifications. These advantages have led to rapid adoption of mass spectrometry in the field of proteomics as the primary experimental technology [144].

Another important development in the late 1990s was the coupling of gel-independent fractionation methods with mass spectrometry of proteins. Although liquid and gas chromatographic columns were historically combined with mass spectrometers, it was only in 1995, after the introduction of ESI, that Appella et al [7] demonstrated the superior ability of LC/MS to process very complex protein mixtures. It is also possible to perform in-gel separation first and then cut the gel into slices and subject each slice to LC/MS, thus achieving an even more significant complexity reduction [133].

Finally, following the success of genome sequencing, there appeared analysis tools that could match the mass spectrum of the peptides to their amino acid sequence [120, 44], thus completing the experiment pipeline from the complex mixture separation to identification of proteins in the mixture. The following section gives more detail on the

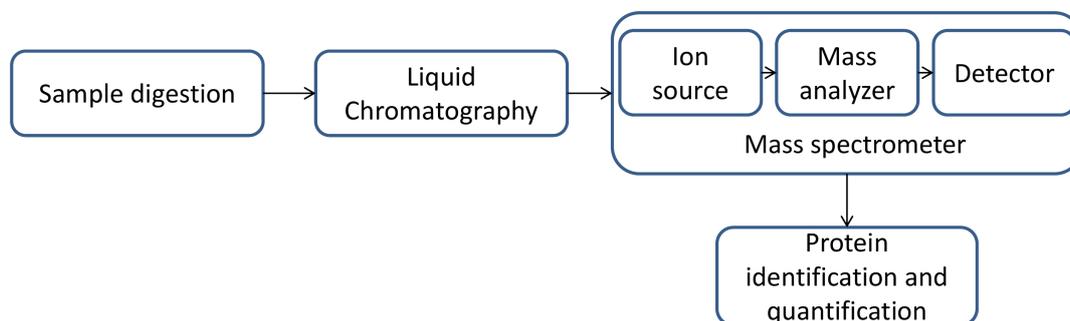


Figure 2.2: A schematic view of an LC/MS proteomics experiment.

parts of the pipeline, which play an important role in analysis of the data used in this thesis.

2.1.1 Experimental scheme and instrumentation

Fig.2.2 shows an overview of an LC/MS proteomics experimental workflow. In the first phase of the workflow, the protein sample is digested into peptides by a protease. One of the most important reasons to use peptides rather than intact proteins is that the mass spectrometer is most efficient in obtaining sequence information for fairly short peptides (up to 20 amino acids) [144]. Top-down proteomics experiments forgo this phase and introduce whole proteins into the mass spectrometer. This approach can be beneficial for detection of post-translational modifications [75], however, all experiments considered in this thesis belong to the bottom-up type and include a digestion phase. Specific or non-specific proteases can be used depending on the required experimental conditions, but site-specific ones are usually preferred as they do not generate overlapping peptides. Trypsin is a popular site-specific protease, which always cleaves the protein after lysine or arginine amino acids. Cleavage sites of pepsin, a non-specific protease, are not so easy to predict, although it also has a preference for certain amino acids. Pepsin is frequently used in Hydrogen-Deuterium Exchange experiments, described in Chapter 3. Trypsin digestion is a common choice for $^{16}\text{O}/^{18}\text{O}$ labeling experiments of Chapter 4.

After digestion the peptide mixture is introduced into a liquid chromatography column. The column “re-orders” the peptides by their hydrophobicity and lets them out into the ion source in small droplets. Each droplet only contains a sub-sample of all the peptides, thus reducing the complexity of the mixture entering the mass spectrometer. A mass spectrum is further acquired separately for each droplet. Consequently, the overall spectrum produced in the experiment has two dimensions: chromatographic

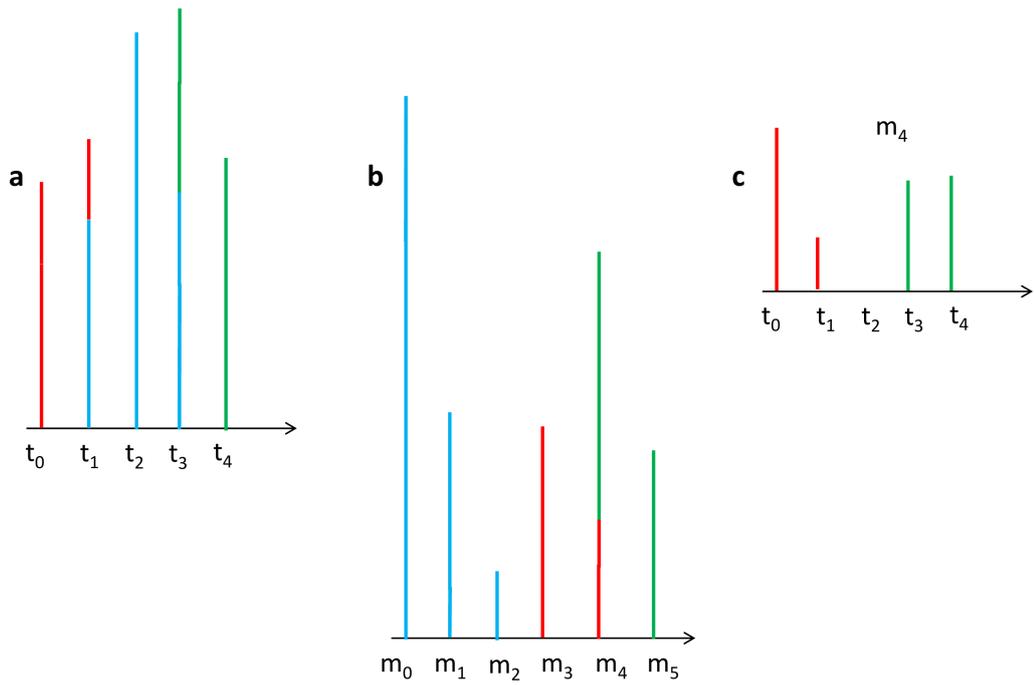
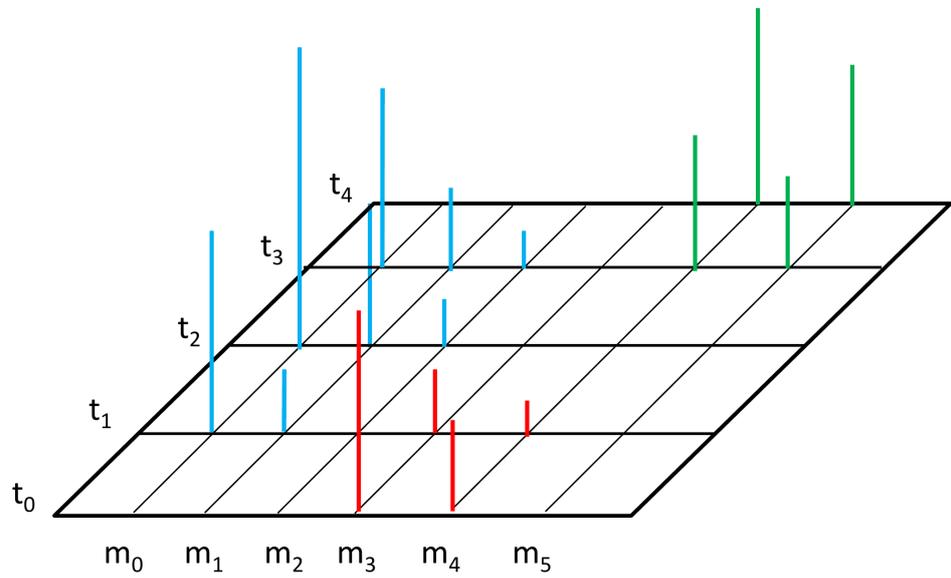


Figure 2.3: Top: a 2D mass spectrum, containing 3 peptides (cyan, red and green). Bottom: a) Total Ion Chromatogram of the 2D spectrum; b) Integrated mass spectrum; c) XIC for m/z value m_4 .

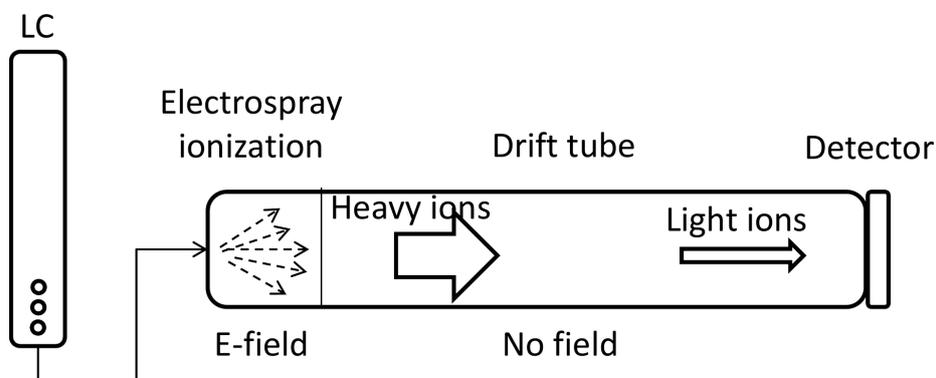


Figure 2.4: A schematic view of a TOF mass spectrometer.

retention time and mass-over-charge ratio. Each peptide is present in several consecutive droplets. The following characteristic plots can be used to study the elution behavior of the sample: a *Total Ion Chromatogram* (TIC) and *Extracted Ion Chromatograms* (XICs) for the individual peptides. *Total Ion Chromatogram* (TIC) is produced by summing up the intensity of the overall spectrum across the mass-over-charge dimension. An *Extracted Ion Chromatogram* (XIC) shows how the intensity of a given mass-over-charge value (with certain tolerance) changes over time. An example of a 2D mass spectrum with a TIC and a XIC is given in Fig. 2.3.

At the exit from the liquid chromatography column, the droplet is vaporized and ionized by a strong electric potential (ESI). Popular ionization techniques include ESI and MALDI mentioned in the previous section, as well as surface enhanced laser desorption/ionization (SELDI) [64] and secondary ion mass spectrometry (SIMS) [89]. As all the experiments considered in this thesis were performed with ESI ionization, we will not describe other methods in detail. Peptides and proteins are usually ionized in positive ionization mode and the resulting charge of a peptide depends on the number of additional protons it attracts, which, in its turn, depends on the length of the peptide and its amino acid composition. The same peptide can be observed in multiple charge states.

The next phase of the workflow brings the ionized peptides into the mass analyzer. The main function of the mass analyzer is to separate the peptides by their mass-over-charge (m/z) ratio. The exact separation technique depends on the type of analyzer,

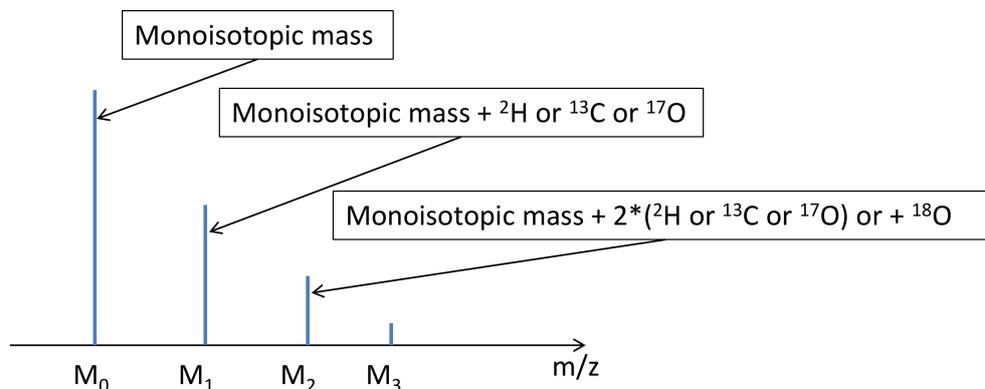


Figure 2.5: An example of an isotope envelope. The charge can be computed from the ratio $M_1 - M_0 = 1/\text{charge}$.

where Quadrupole, Time-of-Flight and Orbitrap analyzers can be mentioned as the most popular ones. Tandem mass spectrometers have two (or even more) mass analyzers, not necessarily of the same type. For example, the data analyzed in Chapter 3 was generated by a Quadrupole-TOF instrument. For illustration, a scheme of the TOF mass analyzer is shown in Fig. 2.4. The separation of ions is based on the time they take to drift through a field-free tube after acceleration by the electric field, which is related to m/z of the peptides as $t^2 = C\sqrt{(m/z)}$ for constant C . An electrostatic mirror (reflectron) can be placed at the end of the tube to force the ions to turn around and drift another tube length, thus improving the m/z resolution of the analyzer [42].

The final part of the mass spectrometer is the detector which transforms the ion signals into a mass spectrum. Most modern mass spectrometers use a microchannel plate detector [37]. We will not focus on its principles in this chapter, but will consider the problem of estimating the charge of detected peptides instead. As noted before, the analyzer separates the peptides based on their mass-over-charge ratio, not on raw mass. Consequently, charge estimation is a pre-requisite for the correct mass estimation. The method relies on the high resolution of the modern mass spectrometers, which record separate signals for the instances of the same peptide with different numbers of stable isotopes. Since 1% of naturally occurring carbon is ^{13}C and not ^{12}C , each C atom of the peptide can turn out to be ^{13}C with 1% probability. Consequently, a sample of this peptide will contain several instances with approximately 1 unit mass difference. In the mass spectrum, these instances will produce peaks at the distance of $1/\text{charge}$, which are referred to as “peptide isotope envelope”. The charge can then be calculated from

the distance between the peaks of the envelope, as shown in Fig. 2.5).

2.1.2 Protein identification

The first phase of LC/MS proteomics data analysis after the acquisition of mass spectra usually performs peptide and protein identification. Identification is based on tandem or MS/MS mass spectra, which are acquired in alternation with regular spectra (also known as MS1 spectra) during the experiment. To be more precise, when a full MS1 spectrum is acquired, the most abundant ions of this spectrum are isolated and fragmented even further, and the mass of the fragments is measured in the second mass analyzer. Their mass spectra are referred to as “MS/MS” spectra, and the original ion that was fragmented is also known as “precursor ion”. Fragmentation achieved by Laser-induced dissociation (LID), Collision-induced dissociation (CID), Electron-capture dissociation (ECD) or Electron transfer dissociation (ETD) [42] usually breaks the backbone of the peptide. Only one half of single-charged peptides retains the charge and appears in the mass spectrum. Alternatively, both halves of double and more charged peptides can be present in the mass spectrum. Since the fragmentation process can break the peptide between any two amino acids, its MS/MS spectrum consists of peaks produced by fragments with one amino acid difference. “*De novo*” sequencing then tries to reconstruct the peptide by finding its amino acids one by one from the mass differences between the peaks of the MS/MS spectrum. However, this approach is not considered very reliable, as the spectra usually contain a certain amount of noise which can corrupt the distance calculation [144]. If the genome of the organism has already been sequenced, peptide identification can be performed by a more robust technique of database matching. This method is based on the observation that only very few of all potential amino acid sequences actually occur in nature. The genome sequence of the organism defines which proteins could potentially be present in the spectrum and thus limits the search space for possible peptide sequences. The observed spectra are matched with the theoretical spectra in the databases and identifications are returned along with the “score” of the matching. If several peptides of a protein were identified with a high score, the protein identification is considered certain. The most popular database search tools include Mascot [120], Sequest [44] and Protein Prophet [109].

2.1.3 Protein quantification

For many types of data analysis, the necessary identification phase must be complemented by an estimate of protein quantity and its changes under experimental conditions. All experiments studied in this thesis refer to this type of analysis and rely on stable isotope labeling to perform relative quantification and compare the abundance of proteins between different samples. Chapter 3 offers an automated technique for studying specific patterns in the changes of the peptide abundance. Chapter 4 proposes

a method to automate quantification for $^{16}\text{O}/^{18}\text{O}$ -labeling - a specific type of stable isotope labeling - and describes other quantification methods in detail.

2.2 Neural circuit reconstruction and Electron Microscopy

The study of the brain at the microscopic level was pioneered by Santiago Ramon y Cajal(1852-1934), who also created the first drawings of neural structure comprised of individual cells and their connections [138]. Ever since subsequent studies aimed to reconstruct the wiring of the neural circuits to further comprehend their function. However, it was not until 1986 that White et al [161] presented the first full reconstruction of a nervous system of an organism - a small roundworm *Caenorhabditis elegans* with 302 neurons. The reconstruction effort required 10 years of manual processing of electron microscopy images.

Over the years of research, many imaging methods have been used to study the neuron structure, starting from Golgi staining and apochromatic objectives used by Cajal [142] to modern isotropic resolution electron microscopes. Amazing advances have recently been made in fluorescent staining and light microscopy, which has now overcome the diffraction limits and reached sub-100 nm resolution with visible light wavelength [53, 130, 10]. The resolution of volumetric imaging was greatly improved by the introduction of array tomography by Micheva and Smith in [102]. While z-resolution of confocal microscopes is limited at approximately 700 nm, array tomography acts on arrays of ultra-thin tissue sections and is thus limited only by the minimal achievable slice thickness, which can be as low as 50 nm. The method can be multiplexed by applying a series of different fluorescent staining techniques and imaging them separately. Moreover, if heavy metal staining is applied in the end, the slices can also be imaged by an electron microscope. Another recent contribution of Micheva et al [101] demonstrates how analysis of proteins present in synapses can be performed with array tomography.

However, electron microscopy remains the only technique capable of simultaneously following all the neural processes in a dense volume of tissue [21]. The oldest and most popular electron microscopy technique is Transmission Electron Microscopy [79], which acquires the images by transmitting an electron beam through a thin slice of tissue and then focusing it on a detector. Staining of tissue with heavy metals makes the neuron membranes and ultracellular structures electron opaque and leaves intracellular and extracellular space transparent, consequently all neural processes of the tissue volume are stained and can be studied at the same time. Volumetric imaging can be performed by serial section Transmission Electron Microscopy (ssTEM) [157] which first cuts the tissue into ultra-thin slices, then images the slices separately and finally combines and registers the images into a stack. A schematic view of the ssTEM pipeline is shown

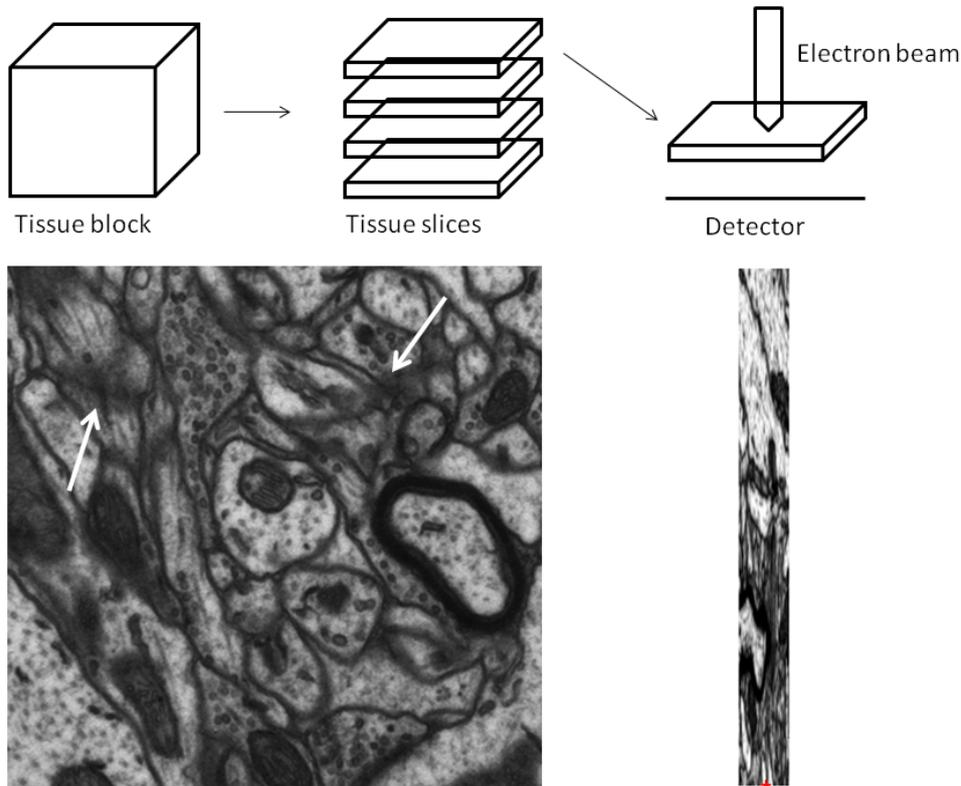


Figure 2.6: Top: a schematic view of ssTEM experimental pipeline. Bottom left: an example of an ssTEM image. Note the smeared membranes, pointed out by the arrows. Bottom right: a side view of an ssTEM stack, showing residual misalignment.

in Fig. 2.6. ssTEM allows to image very large volumes of tissue (this is the technique which was used in [161]) at very high planar resolution, reaching 3×3 nm. However, the z resolution of the image stacks is much worse, at the level of 40-60 nm. As the electrons are recorded after they pass through the tissue slice, areas of very non-uniform electron opacity appear smeared. Such areas can represent, for example, neuron membranes or other ultracellular structures extending in directions at a low angle to the slicing plane. This disadvantage makes it more difficult to follow very thin neural processes through the tissue during subsequent data analysis. Another important disadvantage of the ssTEM technique is the necessity to handle ultra-thin tissue slices. Although the handling has now been automated to a large degree [57], folds, tears and other artefacts still occur in the slices and the imaging conditions are not completely homogeneous, which introduces differences in brightness and contrast, further complicating the image processing.

Besides, images for different slices have to be aligned by a post-processing registration procedure, which is in itself a nontrivial problem [131, 2]. An example of an ssTEM image and the side view of the stack are shown in Fig. 2.6.

The poor z-resolution of ssTEM image stacks can be improved by serial section transmission electron tomography [98, 11], which images each slice hundreds of times, each time tilted at a different angle. Veeraraghavan et al in [154] propose a novel reconstruction method, which allows to limit the number of necessary tilts to five. Although the z-resolution of the stacks does improve and reaches 5 nm, since traditional TEM imaging is used internally, the disadvantages stemming from the necessity of ultra-thin slice handling are still present.

These shortcomings were addressed by Denk and Horstmann [34], introducing Serial Block-Face Scanning Electron Microscopy (SBFSEM). Unlike traditional ssTEM, it does not pre-cut the volume into slices, but uses a custom-designed microtome - diamond knife - to alternatively cut off an ultra-thin tissue slice directly in the sample chamber of the microscope and image the remaining block of tissue by Scanning Electron Microscopy. The image is constructed by measuring backscattered electrons, which allows to distinguish the parts of the tissue stained with heavy metals as in traditional electron microscopy staining. SBFSEM microscopes can acquire image stacks with isotropic resolution reaching 25 nm.

The resolution was further improved by Knott et al in [80], introducing Focused Ion Beam Scanning Electron Microscopy (FIB/SEM). Like SBFSEM, it is based on scanning electron microscopy and it captures electrons backscattered from under the surface of the tissue block. However, unlike SBFSEM it does not use a microtome to cut the slices off, but mills the tissue with an ion beam, parallel to its surface. A schematic view of a FIB/SEM pipeline is shown in Fig. 2.7. FIB/SEM microscopes now demonstrate unprecedented isotropic resolution of 3 nm, which allows for detailed study of synapse morphometry. Also, like SBFSEM microscopy, it does not suffer from slice misalignment. The only disadvantage of this technique is the limitation of the imaging area, however new solutions to this problem are now being proposed and stacks of tens of thousands of pixels in each dimension can now be acquired.

Full reconstruction of the wiring diagram of a volume of neural tissue requires tracing of all the neurons in the tissue and finding the synapses which connect them [27, 28, 68]. While most of the recent literature is devoted to automating the former task, our work, presented in Chapter 5, introduces a method for automated detection and segmentation of synapses in FIB/SEM data and considers different approaches to extending the method to ssTEM data.

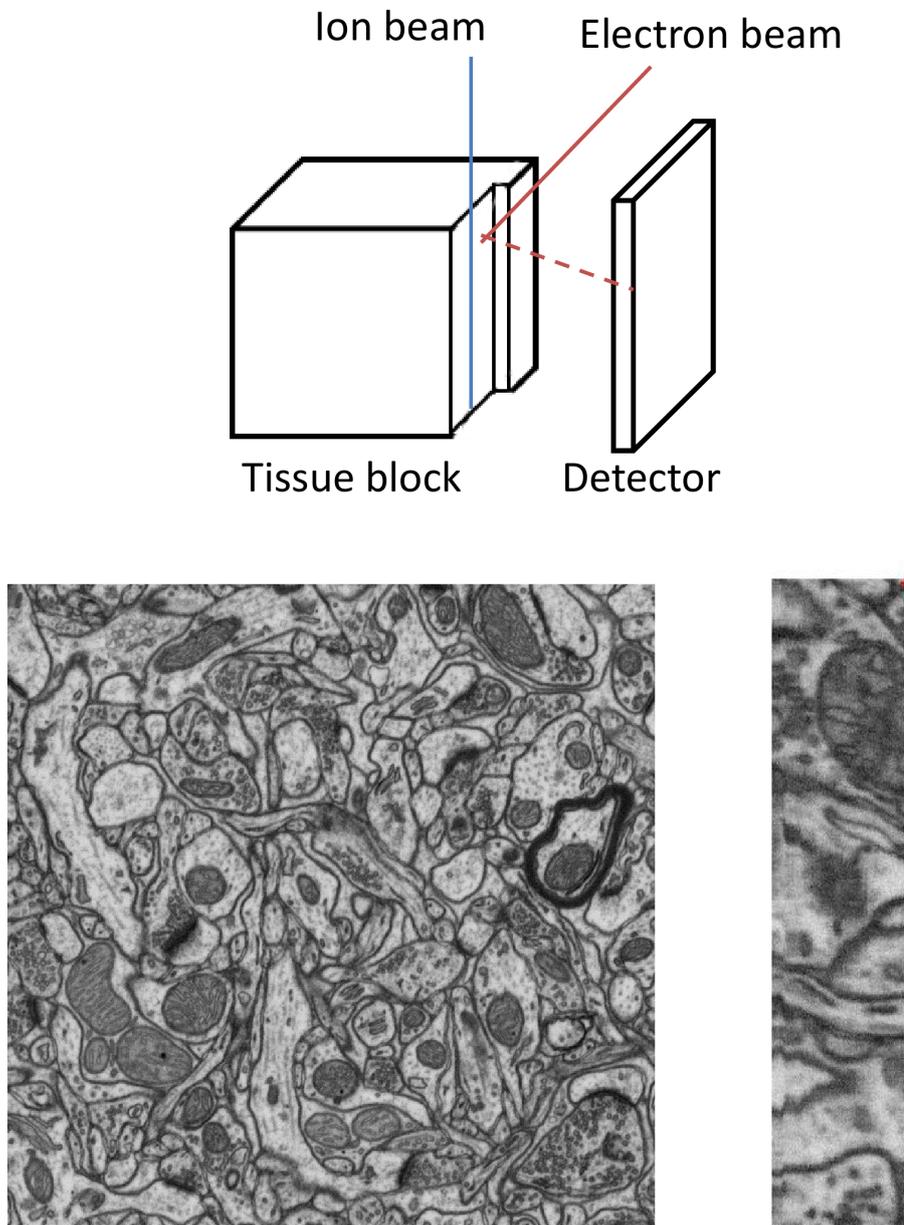


Figure 2.7: Top: a schematic view of a FIB/SEM microscope. Bottom left: an example of an FIB/SEM image. Bottom right: a side view of a FIB/SEM stack, showing its excellent z-resolution.

Chapter 3

Quantification of Bimodal Isotope Peak Distributions in H/D Exchange Mass Spectrometry

3.1 Introduction

Similarly to the genes being defined by sequences of nucleic acids, proteins are defined by sequences of amino acids forming polypeptide chains as shown in Fig. 3.1. However, unlike the genes, which are fixed on the double helix of the DNA, proteins fold into very complex 3D conformations. Four levels of protein structure are usually defined [3]:

- primary structure, specified by the sequence of amino acids, which are joined by peptide bonds into a polypeptide chain
- secondary structure of α -helices and β -sheets, determined by hydrogen bonds between amide and carbonyl groups. The same mechanism of hydrogen bonding between base pairs holds together the double helix of the DNA.
- tertiary structure, determined mostly by the interactions of the side chains of the amino acids. The hydrophilic side chains tend to stay on the surface of the protein, while hydrophobic side chains prefer to fold inside the protein to be protected from water.
- quaternary structure, joining together several polypeptide chains. Not all the proteins consist of more than one polypeptide chain.

Correct folding of the protein is necessary for its biological function and incorrect folding or incapability of the protein to stay in the correctly folded state is the cause of

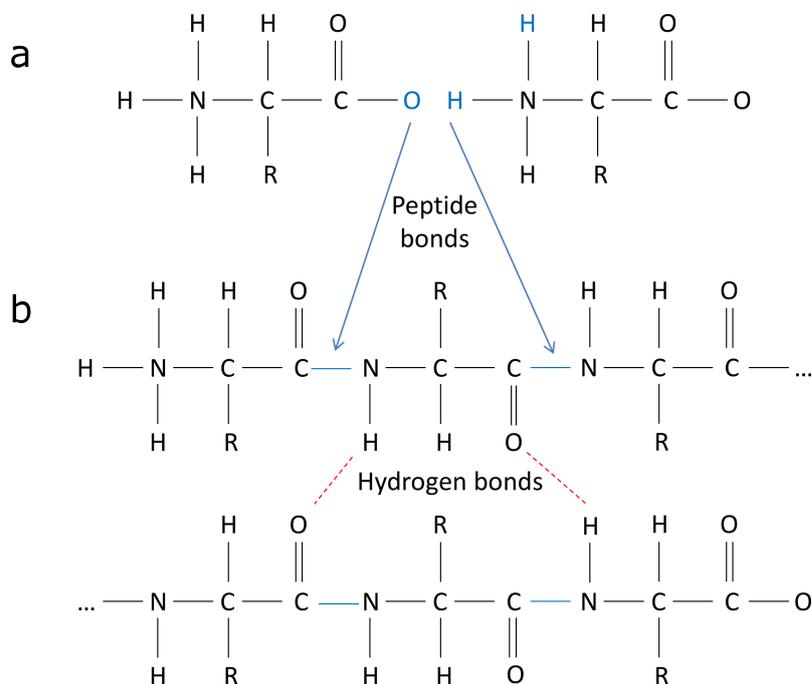


Figure 3.1: a) Formation of a peptide bond, holding together the primary protein structure. b) Hydrogen bonds of the secondary protein structure, shown on the example of a parallel β -sheet.

many pathological conditions, including Alzheimer's disease and prion-related illnesses. However, proteins may undergo a reversible conformational change to switch between states or perform their biological function. Identifying and measuring the rate of such conformational changes is the subject of this chapter.

3.1.1 H/D Exchange Mass Spectrometry

Hydrogen/Deuterium exchange refers to the ability of proteins to exchange their backbone amide hydrogen atoms for deuterium atoms when immersed into heavy water (D_2O). The exchange probability of a given hydrogen atom depends on the position of this atom in the 3D structure of the protein (solvent accessibility) as well as on the

presence of hydrogen bonds [85]. A protected hydrogen, i.e. the one involved in a hydrogen bond or hidden deep inside the protein structure, can not exchange unless a molecular motion exposes it to the solvent or opens its hydrogen bond. Most proteins are intrinsically dynamic under physiological conditions, continuously opening and closing many hydrogen bonds of their secondary structure elements. This flexibility of proteins is considered to be essential for their biological function such as enzymatic activity or allosteric regulation [60]. The opening motions of proteins, ranging from local changes in protein secondary structure to global unfolding events, are reflected in the rate of exchange of each hydrogen and can thus be studied by Hydrogen/Deuterium exchange experiments.

Three rate constants define the most popular kinetic model of the deuteration process of native proteins in D_2O [48]:



where k_{op} is the rate constant of the opening motion and k_{cl} is the rate constant of the closing motion. k_{rc} , the exchange rate of an exposed unprotected hydrogen, depends - among other aspects - on the experimental conditions and on the position of the hydrogen in the amino acid chain and can be computed independently [85]. This model has two limits, EX1 and EX2. EX1 refers to the situation, when k_{rc} is significantly faster than k_{cl} , so a protected hydrogen always exchanges once it transfers to the open state and the exchange rate is limited by the opening rate k_{op} . EX2 refers to the opposite case, when k_{cl} is much faster than k_{rc} and the exchange rate is determined as $k_{ex} = \frac{k_{op}}{k_{cl}} k_{rc}$.

Hydrogen/Deuterium exchange reaction was first studied by Kaj Ulrik Linderstrom-Lang [65, 45], who measured protein deuteration using density gradient tubes. Currently, the majority of HDX experiments are performed either with Nuclear Magnetic Resonance (NMR) spectroscopy or with Mass Spectrometry (MS) instruments. HDX-NMR is based on the difference in the magnetic properties of hydrogen and deuterium and allows to estimate the exchange rate with single amino acid resolution. It is, however, only applicable to smaller proteins of up to 30kDa [128]. HDX-MS does not suffer from this limitation and allows for experiments on larger proteins and multi-protein complexes. Another disadvantage of HDX-NMR is its high requirement for concentration of the protein sample, compared to nano/pico mole samples needed for HDX-MS. Last but not least, the experiment setup time is much shorter for HDX-MS, which allows for studies of fast exchanging hydrogens [114].

HDX-MS studies the hydrogen/deuterium exchange reaction by analyzing the mass-over-charge measurements of a sample comprised of peptides or proteins of interest,

incubated in D_2O for different time intervals (see Fig. 3.3a). An exchange of one hydrogen for a deuterium produces a mass shift of 1 mass unit in the peptide isotope distribution and of $1/\text{charge}$ in the observed peptide spectrum, thus the deuteration of a peptide can be deduced from the deconvolution of a peptide isotope distribution into contributing forms. To localize fast and slow exchanging regions within a protein, samples are digested by a protease after the exchange reaction under low pH quench conditions. Pepsin, which is generally used for proteolytic digestion due to its activity at low pH, has low sequence specificity and generates many overlapping peptides. For larger proteins the number of generated fragments that are then analyzed by liquid chromatography mass spectrometry (LC/MS) is immense and manual data analysis becomes very time consuming. In general, a minimum number of peptides that cover the protein sequence sufficiently are analyzed and the rate of the exchange reaction is estimated from the comparison of spectra, corresponding to different sample incubation times (in D_2O). Even for this limited number of peptides, manual analysis of HX-MS spectra is very non-trivial, as the isotope envelopes of a peptide in different deuteration states (with a different number of exchanged hydrogens) overlap significantly and form complex peak clusters. To facilitate this step, several software packages have recently been proposed for automatic or semi-automatic processing of HX-MS data, including HX-Express [158], The Deuterator [116], TOF2H [110], Hydra [141], and others.

Many of these methods (among others, [115, 158, 110]) only track the changes in the position of the center of mass of the isotope cluster and thus limit their analysis to the estimation of the average peptide deuteration rate and discard a lot of information contained in the spectra. Others ([117, 141, 62]) can estimate the full deuteration distribution, i.e. the relative abundance of peptide isoforms corresponding to each possible number of incorporated deuterium atoms, but are limited to small-scale well-tuned problems that arise, for example, when the peptide of interest is pre-selected and the spectrum is well separated. This limitation prevents such methods from being used in a setting, where a large-scale screening is performed first to localize potentially interesting peptides, which are then analyzed more closely by MS/MS experiments.

Recently, another software package named HeXicon was proposed [90], which does not suffer from any of the limitations above and computes a robust estimate of the full deuteration distribution. It is not limited to peptides previously identified by MS/MS and provides strong sensitivity with balanced specificity. With HeXicon, analysis of the most common continuous labeling HDX-MS experiments (as shown in Fig. 3.3) can now be performed in fully automated manner. However, there are other types of HDX-MS experiments, to which HeXicon can not be applied directly. An important example of such an experiment is described in the following section.

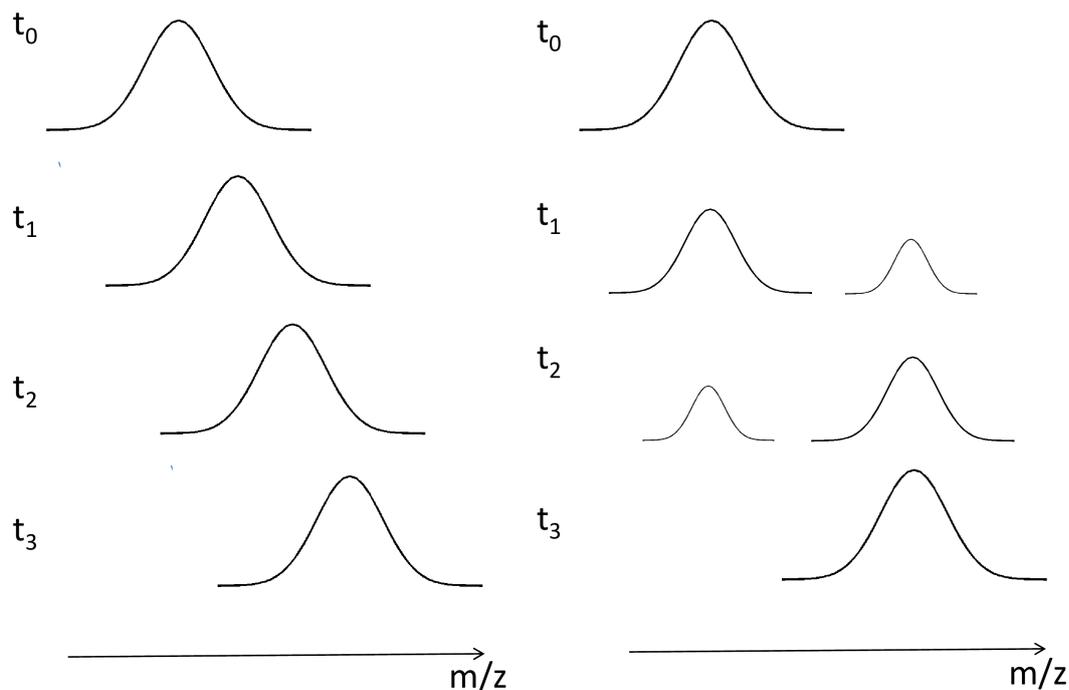


Figure 3.2: Comparison of unimodal and bimodal distributions, bell-shaped peaks depict the centers of mass of the isotope distributions (for the right side of the figure, centers of mass of the slow and fast exchanging populations). Left: The original use case of HeXicon - a unimodal distribution, shifting to higher m/z values as the deuteration time increases. Right: Bimodal deuteration distribution. Note, that both unimodal and bimodal distributions correspond to a continuous labeling experiment, unlike pulse labeling experiment data, used in this study. In our data all samples are immersed in D_2O for the same period of time and bimodality is caused by different ATP incubation times (see also Fig. 3.3). However, the image on the right illustrates the challenges of bimodal deuteration distribution estimation correctly also for our experimental setup.

3.1.2 Motivation for Analysis of Bimodal Isotope Peak Distributions

A major benefit of HDX-MS approaches is their ability to detect coexisting populations of a single protein, which exhibit different exchange behaviors. Such different populations can arise from slow local unfolding due to intrinsic protein instability or from induced conformational changes that are slow compared to the exchange reaction kinetics. In HDX-MS experiments, two coexisting populations with different exchange kinetics are characterized by a bimodal peak distribution within the isotope cluster [159] (Fig. 3.2right). The data of such experiments, while giving important insight into the kinetics of conformational changes, provides an even greater challenge for manual, as well as automated, analysis methods. The current general approach for automated exchange rate determination is based on the detection and tracing of the center of mass of a protein's peptides over all incubation times for which measurements are available. Deuteration level determination based on the center of mass reduces the deuteration distribution for each incubation time to a single point measurement. Consequently, all information contained in the deuteration distribution is lost. In case of a bimodal peak distribution, when a peptide isotope distribution - as shown in Fig. 3.2 - has two possibly overlapping centers of mass which correspond to two different exchange behaviors, this single exchange rate estimate is not informative.

An alternative approach to bimodal distribution analysis has recently been demonstrated in [159]. The authors sought to characterize the EX1 kinetic limit of exchange for proteins which exhibit EX1 kinetics under physiological conditions. This limiting case occurs when in a region of a protein the exchange rate for exposed hydrogens is higher than the rate of closing and consequently, once the region is opened, all its hydrogens exchange before the protein re-folds to protect them again. As only some regions in a protein participate in these cooperative exchange events, peptides resulting from peptic digestion will likely contain both EX2 and EX1 exchanging residues and, depending on the size of the EX1 region, their isotope clusters might overlap substantially. Width of the isotope distribution was proposed by the authors as an estimate of the half-life of the conformational change and the number of hydrogens involved. The width itself was estimated either directly from the shape of empirical isotope envelopes or from the deuteration distributions computed by the DEX software. [62]. As the DEX software is based on Fourier deconvolution of an observed isotope cluster into a native isotope cluster and deuteration-induced changes, it relies on the sequence of a peptide being known to compute the native distribution and can thus only be applied to peptides identified by MS/MS.

With the introduction of HeXicon, we obtained a method to recover isotope distributions in a fully automated manner, also for the peptides, not identified by MS/MS. In this study we introduced several modifications to HeXicon to specifically search a large

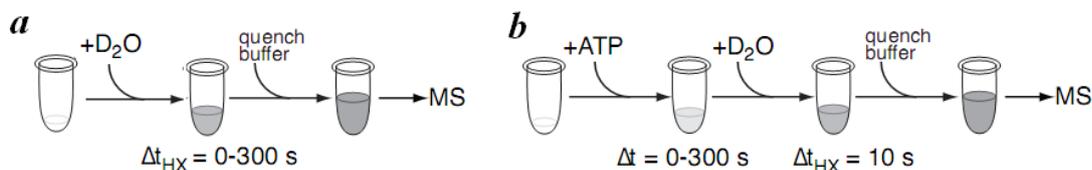


Figure 3.3: a) Continuous labeling experiment - the original use case of HeXicon. b) Pulse labeling combined with ATP incubation as used in this study.

HX-MS dataset for peptides with bimodal distribution patterns and analyzed the kinetics of the conformational changes based on HeXicon results. The method was tested on the LC/MS measurements of pulse-labeling HX-MS experiment with *Escherichia coli* HtpG, a protein that belongs to the family of 90 kDa heat shock proteins [51]. This work has been performed in collaboration with the group of Dr. Mayer (ZMBH, Heidelberg).

3.2 Methods

3.2.1 Overview of the HeXicon software

The workflow of the HeXicon procedure starts by determining which peptides could possibly be observed in the spectrum from MS/MS results and *in silico* digestion of the protein. These peptides candidates are then used as potential components in the mixture model, fit by an L1-regularized regression by application of the NITPICK peak picking algorithm [124]. Before peak picking, the two-dimensional LC/MS spectrum is segmented by a watershed-based algorithm. Segmented parts of the spectrum then contain only the signal of one peptide in all its deuterated forms and possibly other peptides which overlap with it. After Nitpick finds the peaks in the spectra corresponding to different D_2O sample incubation times, peaks which belong to the same peptide are matched across the spectra by Euclidean distance in the two-dimensional space of retention time and m/z . The deuteration distribution of these peptides is then found from the quantitative results of Nitpick. Finally, the protein sequence coverage is improved by also considering the peak groups, for which no MS/MS results exist. This greatly improves the sensitivity of the algorithm, but the specificity can then suffer from the false positive detections. To avoid this, a Random Forest classifier [18] is trained on manually labeled peptide examples to produce a quality score based on several peptide features (such as the variance of the deuteration distribution).

NITPICK peak picker

The core of HeXicon is the NITPICK [124] peak picking: a non-greedy algorithm for fitting of mixture models to spectra. Given a set of stoichiometries of peptides that could potentially be found in a given spectrum, NITPICK can deconvolve complex overlapping isotope distributions and find a sparse subset of the peptides that best explain this spectrum. Peptide stoichiometries can be built based on MS/MS or protein sequence information, or, in case neither is available for a given data point, approximated by the fractional average method. Spectra, corresponding to the stoichiometries, are generated from stable isotope distribution tables and later serve as potential components of the mixture model. Internally, NITPICK uses non-negative lasso regression [148], which can be computed in a very efficient manner by the least angle regression algorithm LARS [41]. Following the notation of [124], the optimization problem of lasso regression for the case of mass spectra fitting can be formulated as

$$\hat{c} = \underset{c}{\operatorname{argmin}} \|s - \Phi c\|,$$

$$s.t. \sum_{i=1}^K c_i \leq t, c_i \geq 0$$

or in the equivalent Lagrangian form as

$$\hat{c} = \underset{c}{\operatorname{argmin}} \|s - \Phi c\| + \lambda \sum_{i=1}^K |c_i|,$$

where s is the observed spectrum, Φ is the matrix of model spectra and c_i are the unknown non-negative model concentration values. Non-negative t or λ parameters control the level of regularization.

The LARS algorithm iteratively adds models to the active set, by, at each iteration, finding the model which is currently the most correlated with the residual, adding it to the active set and moving the coefficients of the active set models in the direction, defined by their joint least squares coefficient of the current residual on the active set until another model has as much correlation. To compute lasso solutions, the algorithm is modified to drop a model from the active set when its coefficient hits zero. Unless an early termination criterion is used, the algorithm returns a set of solutions for each value of λ for which the active set changes, up to the full least squares solution, and the most suitable value of λ then has to be selected by, for example, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) or generalized cross validation. To avoid computing additional solutions when the models in the active set explain the data reasonably well, Nitpick implements an early termination criterion based on the correlation of the new variable, entering the active set at each iteration. If the gain

in correlation by including a new model falls below a pre-defined threshold, the LARS iterations are stopped and the BIC is computed for each value of λ as

$$BIC(\lambda) = \frac{1}{\sigma^2} \|s - \Phi_{A(\lambda)} c_{A(\lambda)}^{nlsq}\|^2 + df(\lambda) \log(N),$$

where $A(\lambda)$ refers to the active set of models at a given value of λ , $c_{A(\lambda)}^{nlsq}$ is the non-negative least squares solution and $df(\lambda)$ can be estimated by generalized degrees of freedom. If for some value of λ the BIC reaches a deep enough minimum, the active set corresponding to this λ is returned. If none of the minima were found to be deep enough, the last active set corresponding to the largest value of λ is returned. The required depth of the minimum is a user-defined parameter, which was set to five in the implementation we used.

3.2.2 Tuning HeXicon to search for bimodal distributions

Improvement of the early termination criterion

While NITPICK has been thoroughly tested on synthetic and real-world mass spectrometry proteomics data, deconvolving the bimodal isotope distributions of HX/MS peptides of this study turned out to be especially challenging for it, as the true model mixture contains a lot of very correlated models. For many such peptides the early termination criterion stopped the LARS iterations too soon and thus prevented the algorithm from including all relevant peptide isoforms into the active set. We have therefore modified the behavior of the termination criterion in case of a very small gain in correlation to “pause” the LARS iterations and compute the BIC trace. If the BIC minimum is not deep enough, the algorithm resumes the LARS iterations until the true BIC minimum is found. This change was beneficial for all use cases of NITPICK and is now part of its standard distribution. An example of the subsequent improvement in the deuteration distribution estimation can be seen on Fig. 3.4

Other modifications

HeXicon was originally developed for continuous-labeling time course experiments (individual samples differ by D_2O incubation time) as shown in Fig. 3.3 and accordingly assumes monotonic increase of deuterium incorporation. However, in the studies of induced conformational change, the deuteration can decrease with the incubation in the change-inducing reagent, if the protein transfers to a more protected state. To preserve the assumption of HeXicon also in this case, one could process the samples in a pairwise manner such that samples at different incubation times are only compared with the reference (non-deuterated) sample. The resulting deuteration distributions for all time points could be combined for each peptide identified by MS/MS or found by the coverage extension procedure. An alternative approach is to introduce the samples into

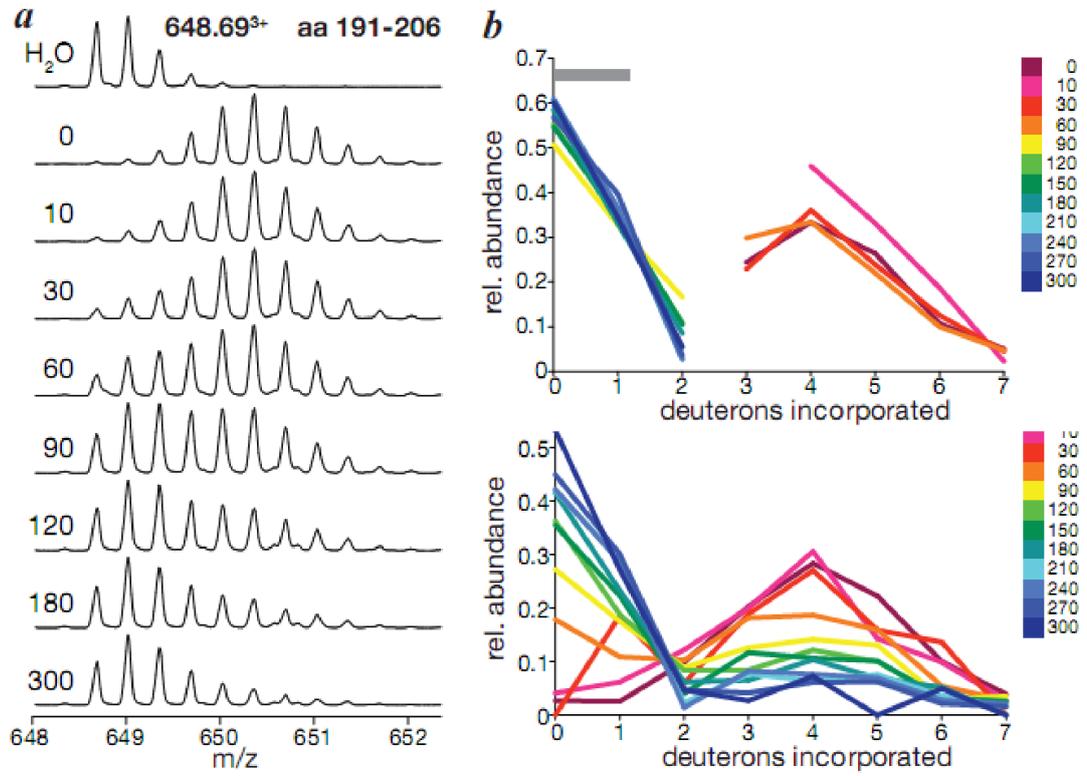


Figure 3.4: a) Spectra of the peptide at residues 191-206 at different ATP incubation times. b) Deuteration distributions estimated by HeXicon. Top: with the original NITPICK implementation. Bottom: with the modified early stopping criterion. Note the improvement in the estimation of the contribution of the lower abundance isoforms.

HeXicon with the deuteration time parameter modified in such a way, that the sequence of time labels is reversed while keeping the time interval constant. In this approach the deuterium incorporation behavior complies with the assumption of HeXicon. The first method is more general as it does not rely on any assumptions about the protein change in the course of the experiment, be it monotonous decrease or increase in deuterium. The second method is more robust for peptides with other high intensity peptides in the vicinity, where peaks found in other samples could help correct the peptide assignment.

Another modification was concerned with the peptide quality score estimation. The final processing step of HeXicon ranks the peptides found in the previous steps with the help of a Random Forest classifier [18]. However, some of the features used for

classification, such as, for example, the variance of the deuteration distribution, are not applicable for the bimodal distribution search. Moreover, since in this limited case we are only interested in the peptides whose distributions follow a certain pattern, a ranking of all found peptides is not needed and can be replaced by a post-processing filtering procedure, tuned to the bimodal pattern of interest. The procedure checks the deuteration distributions of peptides across all samples. It selects the peptides, for which at least for some samples a “2 (local) maxima/2-3 minima” or “1 maximum/2 minima” pattern is observed and several such samples share a maximum position, while at least one other has a different maximum. Sensitivity of the procedure is controlled by the minimal number of samples with the same maximum, after which the distribution is considered bimodal.

3.3 Experiments

The bimodal distribution search was tested on a dataset of LC/MS spectra, produced by pulse-labeling HX-MS experiments with HtpG. Upon addition of Adenosine Triphosphate (ATP) HtpG undergoes a two-step conformational change into a highly protected state. The first step with a half-life of 3 ms occurs too fast for HX-MS experiments as performed in this study (compare [129]), but the second step with a half-life of approximately 120 s is accessible to HX analysis [51]. Manual data analysis was restricted to peptides previously identified by an MS/MS experiment and MASCOT search and yielded 7 peptides with a bimodal distribution. The aim of this study was to screen the entire dataset for additional peptides with a bimodal isotope distribution and to specifically identify these peptides by subsequent MS/MS analysis.

3.3.1 Data acquisition

Fig. 3.3b shows the experimental scheme. ATP was added to HtpG and the mixture was pre-incubated for 0 to 300 s at 30°C before the reaction was diluted into D_2O and incubated for exactly 10 s for pulse-labeling of the protein. Subsequently, the reaction was quenched by the addition of a low-pH-phosphate buffer and analyzed by LC/MS as described previously [127, 126]. As binding of ATP induces the transition of HtpG from the relaxed state with high flexibility and rapid deuteron incorporation to the tensed state with low flexibility and slow deuteron incorporation, isotope peak distributions migrate to lower m/z values with increasing pre-incubation time in the presence of ATP (Fig. 3.7:left and Fig. 3.8:left). All data used in this study was collected on a QSTAR Pulsar mass spectrometer. The analyst .wiff files were converted to mzXML format by mzWiff tool [76] and no further pre-processing such as smoothing was applied.

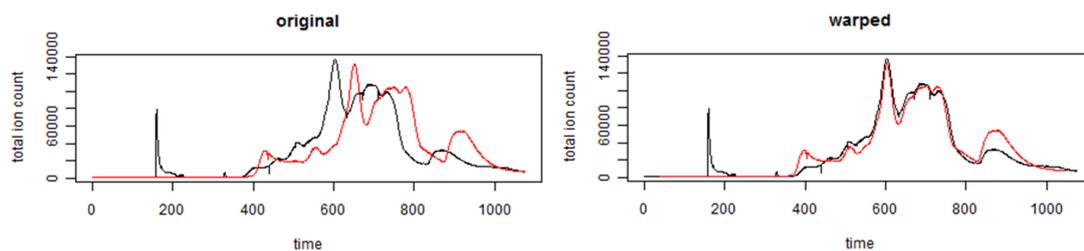


Figure 3.5: Left: original Total Ion Counts (TICs) of 2 spectra. Right: aligned TICs.

3.3.2 Retention time alignment

HeXicon does not have a built-in LC/MS alignment module and often fails to establish the correct inter-sample correspondences when the raw data exhibit severe retention time offsets. This problem occurs frequently in LC/MS based proteomics and numerous algorithms have been proposed in the last years to bring the spectra into alignment. These algorithms can be divided into two classes: the first one aligns ion chromatograms ([122, 153, 29]), and the second one aligns the peak lists after peak picking ([87, 78]). Employing an algorithm of the second type would require modification of the core of the HeXicon algorithm, which we preferred to avoid.

Between the algorithms of the first type, we chose Parametric Time Warping (PTW)[43], a popular algorithm for chromatogram alignment. It belongs to the family of time warping based alignment techniques, which also includes dynamic time warping and correlation optimized warping. An improvement of the original PTW algorithm proposed in [14] provides methods for alignment of Total Ion Chromatograms, global alignment of complete LC/MS spectra (global warping) and global alignment followed by alignment of individual mass traces. Alignment of TICs is the simplest of the three methods, suited for well-behaved data. The other two techniques, which take into account the full LC/MS information, have to be applied if data is acquired in several batches or if the chromatographic columns have to be changed during the experiment. Judging by visual inspection and by the quality of HeXicon results, TIC alignment was sufficient in our case. We aligned the total ion chromatograms of all samples to the reference ATP-free sample.

PTW explicitly models the time warping function by a polynomial, in our case by a parabola. In more detail, if a spectrum $S(t_i)$ is aligned to a reference spectrum $R(t_i)$, the time warping function is defined as such a $w(t)$ that $S(w(t_i))$ matches $R(t_i)$. $w(t)$ is then modeled as $w(t) = \sum_{k=0}^K a_k t^k$. The optimization criterion is formulated as weighted

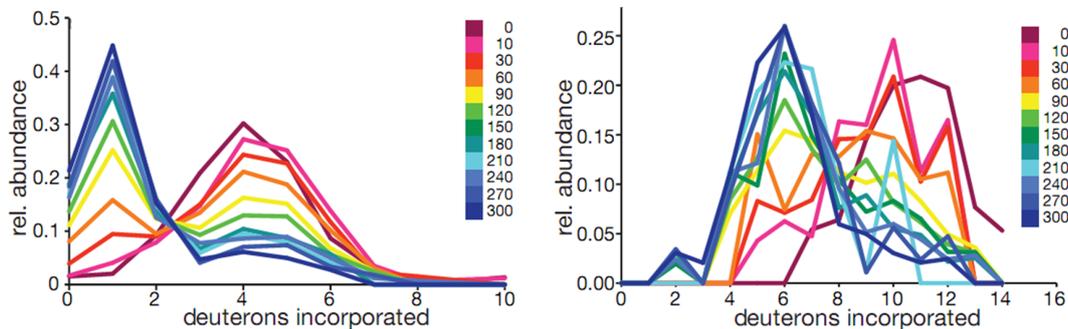


Figure 3.6: Two examples of deuteration distribution with Gaussian peaks.

cross correlation of the spectra [33]:

$$WCC = \frac{R^T W S}{\sqrt{R^T W R} \sqrt{S^T W S}},$$

where W is a banded matrix, representing a triangular weight function. The diagonal elements of W are equal to one and the non-diagonal elements are decreasing linearly as the distance from the diagonal increases. After a certain cut-off distance from the diagonal, the non-diagonal elements are set to zero. The denominator of the fraction makes the WCC value independent of the scale of the data. The WCC criterion, introduced in [14], is an improvement of the RMS criterion used by [43], which is defined as follows:

$$RMS = \sqrt{\sum_i \frac{[R(t_i) - S(w(t_i))]^2}{N}}$$

Unlike RMS , which assigns the maximum penalty if the patterns in two spectra do not overlap regardless of the distance between them, WCC penalizes small shifts between the patterns in the spectra less than large shifts and is thus more precise and easier to optimize.

Publicly available R package `ptw` was used on total ion chromatograms (TICs), extracted from the raw data by the Analyst software suite (see Fig. 3.5 for an example of the alignment results). The retention time values of the raw data in `mzXML` format were updated before being processed by HeXicon. In principle, any algorithm for sample alignment can be used, and in future we hope to be able to provide an alignment module directly in HeXicon [156].

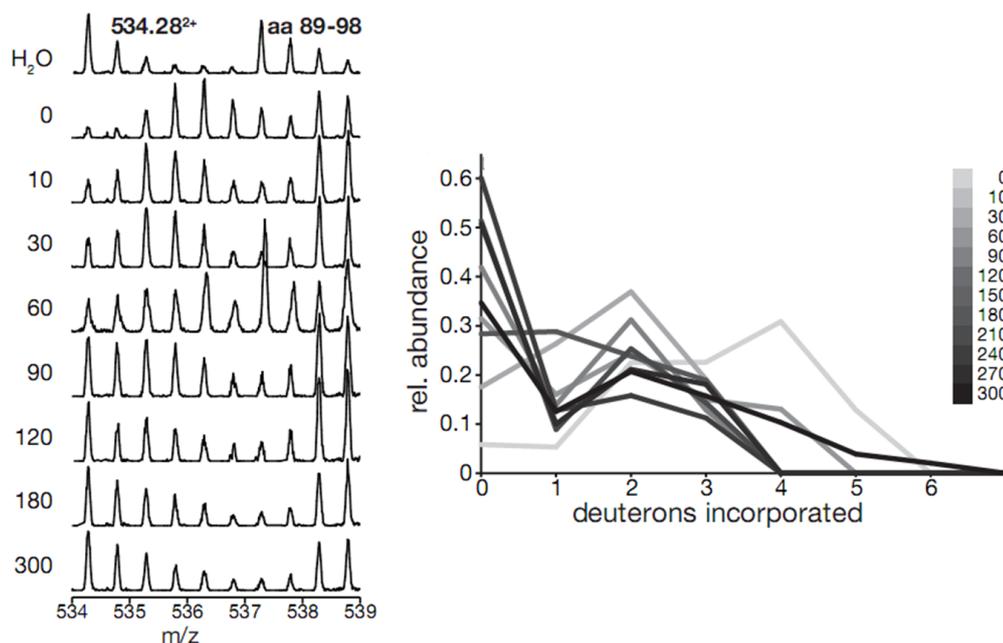


Figure 3.7: An example of a non-Gaussian deuterium distribution caused by two co-eluting peptides.

3.4 Results

3.4.1 Proof of principle

To test the performance of the modified HeXicon algorithm we first restricted the search to m/z values known to represent peptides with bimodal isotope distributions. All seven previously known peptides were identified. Four of these show deuterium distributions that fit well to two Gaussian peaks with the maximum of the earlier time points (red colors) at higher deuterium numbers than the maximum of later time points (blue colors) as expected, see Fig. 3.6. Interestingly, in some cases the distribution of the hydrogen exchange in the absence of ATP (0 s time point) is shifted to higher deuterium numbers as compared to the distribution of the 10 s time point, indicating the fast phase of the conformational change that cannot be analyzed with our HX-MS setup (Fig. 3.7:right). For three peptides the deuterium distribution is not Gaussian shaped but a separation between earlier and later time points is visible (Fig. 3.7). Tuning HeXicon parameters did not lead to more regular shaped deuterium distributions for these peptides, but the separation between fast and slow exchange behaviors was present for all HeXicon parameter settings. Close inspection of the spectra for which HeXicon did not yield clearly separate Gaussian distributions reveals that these spectra are either of low in-

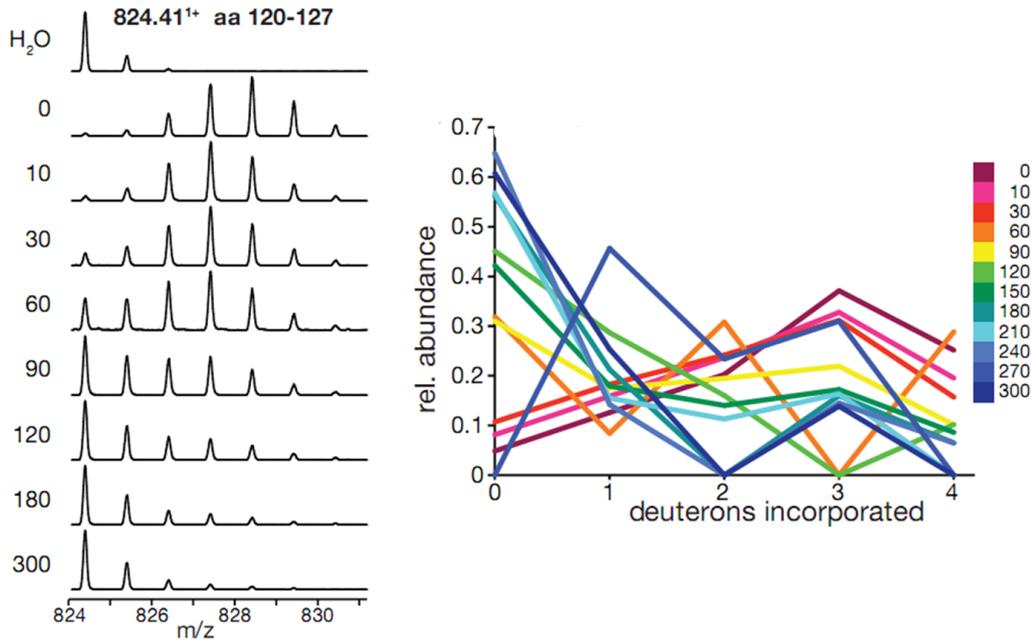


Figure 3.8: An example of a non-Gaussian deuteration distribution caused by low signal-to-noise ratio.

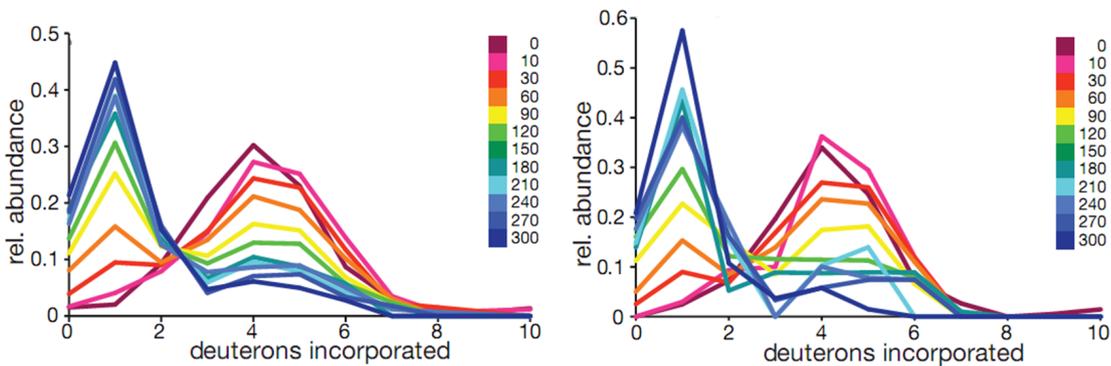


Figure 3.9: Deuteration distribution of the same peptide at different charges

tensity with poor signal to noise ratios (Fig. 3.8) or contain an equally spaced isotope cluster originating from a different peptide which disturbed the determination of the deuteration distribution (Fig. 3.7). These isotope clusters were clearly separated for the undeuterated peptides but merged after deuteration.



Figure 3.10: The amino acid sequence of *E. coli* HtpG is shown with the secondary structure elements as found in the crystal structure of nucleotide-free HtpG (PDB entry code 2IOQ; [139]) above the sequence (dashed lines are not resolved in the crystal structure). The dashed and dotted lines below the sequence indicate the peptic peptides that were analyzed previously [12]. Dotted lines indicate previously known peptides with bimodal distribution. Black solid lines indicate the additional peptides found in this study by HeXicon and verified manually to be correct. All peptides were identified by tandem mass spectrometry.

3.4.2 Analysis of a full dataset

After verifying that the adapted HeXicon was able to detect bimodal isotope distributions we applied the program to the entire dataset with very sensitive post-processing filter settings to avoid missing interesting peptides. The filter produced a plot for a peptide, if in more than two samples the deuteration distributions had the same maximum and at least one more had a different maximum. It only considered distributions following the pattern of “2 (local) maxima/2-3 minima” or “1 maximum/2 minima”. Such a sensitive setting led to a certain amount of false positive results, which were discarded by manual inspection of the produced distribution plots. Examples of the inspected plots can be seen in the figures of this section (Fig. 3.6 and similar). HeXicon identified 27 additional, previously unidentified peptides with deuteration distribution that suggests two populations with distinct exchange kinetics. Nine of these peptides were verified to exhibit ATP-pre-incubation-time-dependent bimodal isotope distributions. For eight of these peptides the identity was determined by a follow-up MS/MS analysis with their monoisotopic masses in the inclusion list (Fig. 3.10). Three of the eight peptides were instances of known species at previously unobserved charge states. As expected, the deuteration distributions determined by HeXicon for the same peptide sequence at different charge states are highly similar (compare the two plots on Fig. 3.9). The remaining five peptides overlap with the previously identified peptides (Fig. 3.10). In addition to our HeXicon analysis, we analyzed the width of the isotope peak distribution as proposed by Weis and coauthors using the HX-Express software tool [5, 9]. For most peptides the half-lives determined by fitting the HeXicon data coincide reasonably well with those estimated by the width of the isotope distributions.

3.4.3 New insight into HtpG protein dynamics

One of the new bimodal peptides found by HeXicon represented a rather short sub-sequence (residues 13-18) of a known longer peptide (residues 1-19) (Fig. 3.11). Identifying this peptide allowed our collaborators from the group of Dr. Matthias Mayer to better understand the mechanism of N-terminal dimerization of HtpG: ”This peptide encompasses a region in HtpG that forms a small α -helix in the homologous yeast Hsp82, which was crystallized in complex with AMPPNP and the cochaperone Sba1. We previously assumed that this α -helix becomes stabilized in the Hsp82 structure upon docking of the two N-terminal domains and by interaction with the cochaperone Sba1. The data for the HtpG peptide identified by HeXicon clearly show that most of the stabilization occurs in the dead time of the experiment, which is considerably before N-terminal docking as determined previously by fluorescence measurements [51]. The stabilization of this α -helix may even be a prerequisite for the docking. This will be explored in a future mutagenesis study.” [84]

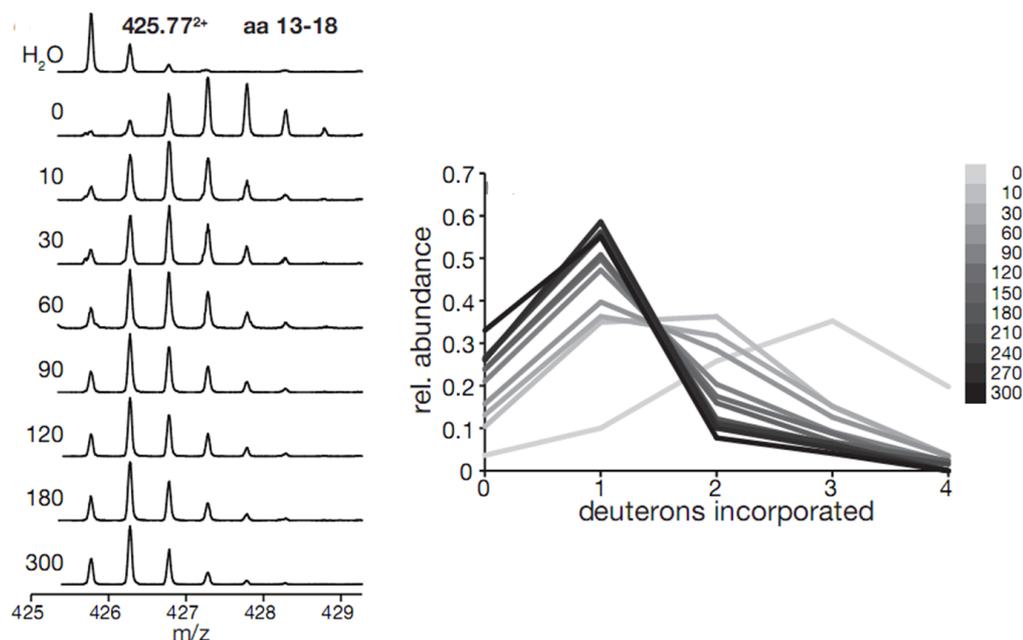


Figure 3.11: Previously unknown peptide with bimodal isotope distribution found by HeXicon

3.5 Discussion

3.5.1 Analysis of false positive candidates

Some of the peptides suggested by HeXicon do not really exhibit a bimodal exchange behavior. The reason for these false positive results is in most cases a second peptide with a close monoisotopic mass co-eluting with the peptide of interest. For example, as shown in Fig. 3.12a, 3.12b, the isotope profiles of peptides at m/z 526.30 and 528.34 overlap for D_2O and ATP incubated samples, misleading HeXicon to incorrectly assign all observed peaks to the same peptide. The seemingly bimodal distribution is then just a sum of the deuteration distributions of two individual peptides. Another example of this behavior is demonstrated in Fig. 3.12c, 3.12d, where a co-eluting peptide of a lower mass interferes with the isotope distribution of the peptide at m/z 960.00, and again causes HeXicon to find a false positive bimodal isotope distribution. The underlying cause for such incorrect assignments is the fairly large m/z error tolerance of 0.1 Da used by HeXicon when establishing correspondences between the peptides in the reference sample and their possible deuterated forms in other samples. Based on our experience with other, non-HX experiments, we anticipate that this will be considerably

less of a problem with experiments that are performed on a higher resolution instrument. As the measured signal peaks become narrower, the isotope distributions of neighboring peptides will no longer overlap, unless it's a perfect overlap of the same peptide in different deuteration states. We could then lower the tolerance for m/z deviations between the supposed signals from the same peptide across different samples (currently, a difference of up to 0.1Da is considered acceptable) and thus improve the accuracy of the peptide matching across the samples. Other examples for false-positive assignments of bimodal exchange behaviors could result from carry-over contaminations. Such samples would exhibit bimodal isotope distribution with seemingly unexchanged species appearing. Since in these cases the bimodal distributions are real, the problem cannot be solved by mass spectrometer or analysis software but only by improving the HPLC conditions, for example by introducing a run without injection in between two analysis runs. HeXicon could help to detect such problems since the amplitude of the appearing unexchanged fraction should not depend on the incubation times or conditions.

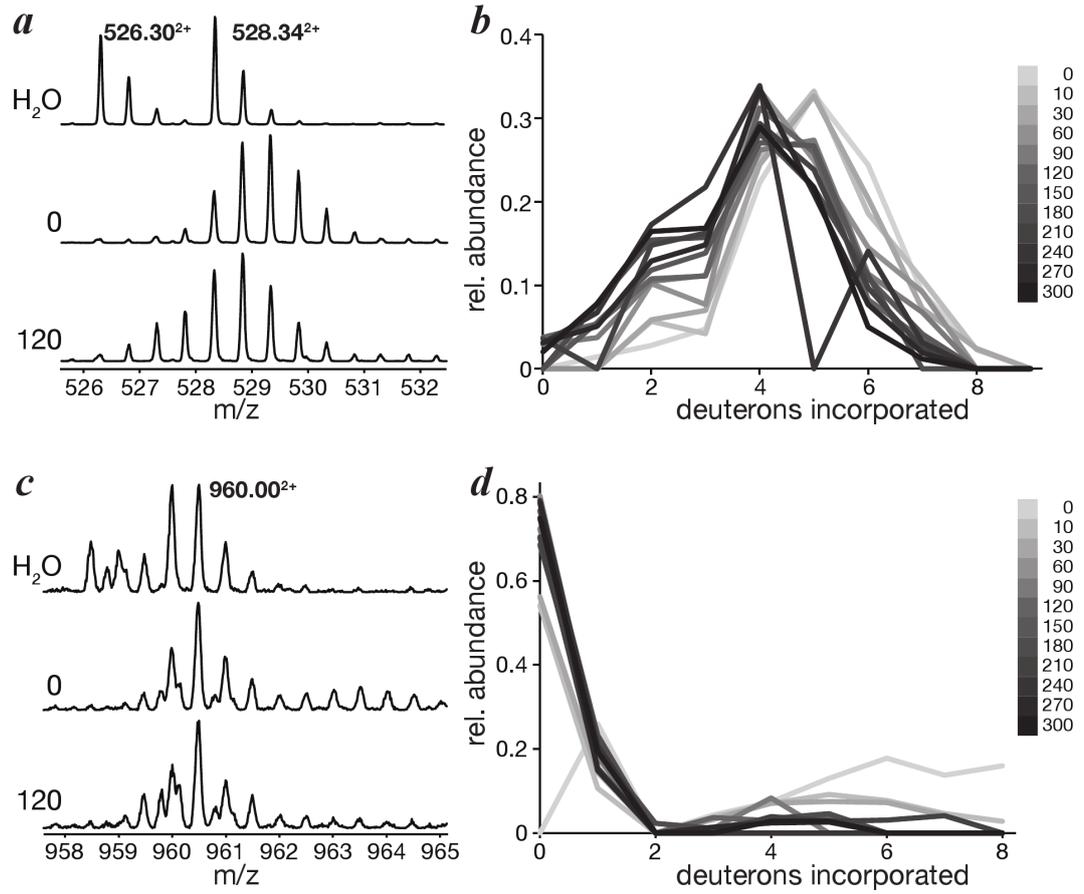


Figure 3.12: Peptides, falsely indicated by HeXicon as bimodal

Chapter 4

Automated Quantification of $^{16}\text{O}/^{18}\text{O}$ -labeled LC/MS Data

4.1 Introduction

Many biological problems, such as, for example, biomarker discovery, require a means to quantitatively compare protein expression levels across different samples. Mass spectrometry is not an intrinsically quantitative method, as the amplitude of a peptide's signal in the mass spectrometer depends on its ionization, and ionization efficiency varies a lot between peptides. Consequently, comparisons have to be relative and each peptide can only be compared with the same peptide in other samples. When absolute quantification is required, it is performed against a labeled internal standard [77].

Peptide relative quantification methods can be divided into methods based on stable isotope labeling and label-free approaches. When stable isotope labeling is used, tags of stable isotopes of different known masses are attached to all the peptides of samples to be compared. Samples are then pooled and processed during a single LC/MS run. Signals from different samples in the resulting spectrum can be distinguished by the mass shift induced by the tag. When label-free quantification is used, an LC/MS spectrum is acquired separately for each sample and the protein signals across samples are compared based either on the ion chromatograms of their peptides or by simply counting the number of spectra in which peptides of this protein were identified (spectral counting). Label-free methods are cheaper and more scalable; they do not require expertise at labeling techniques and allow for comparison of more than 8 samples at the same time. However, they can only handle simple biochemical workflows and are not as accurate as stable isotope labeling quantification, since comparisons are done across different runs with different systematic errors [9].

Element	Symbol	Mass	Natural abundance
Hydrogen	^1H	1.00783	99.99
	^2H	2.01410	0.01
Carbon	^{12}C	12.0000	98.91
	^{13}C	13.0034	1.09
Nitrogen	^{14}N	14.0031	99.6
	^{15}N	15.0001	0.4
Oxygen	^{16}O	15.9949	99.76
	^{17}O	16.9991	0.04
	^{18}O	17.9992	0.2
Sulfur	^{32}S	31.9721	95.02
	^{33}S	32.9715	0.76
	^{34}S	33.9676	4.22

Table 4.1: Mass of atoms and natural abundances of the stable isotopes of hydrogen, carbon, nitrogen, oxygen and sulfur [42]

From the data analysis side, label-free experiments can use any software for LC/MS data quantification, while stable isotope labeling experiments may require custom algorithms for deconvolution of the signals of differently labeled peptides, in case the induced mass shift is not large enough to separate their isotope clusters. Of all the popular stable isotope labeling methods, labeling with ^2H , described in Chapter 3 produces the smallest mass shift of just one unit. $^{16}\text{O}/^{18}\text{O}$ labeling, the subject of this chapter, produces a mass shift of 2-4 mass units, depending on the label incorporation. SILAC (Stable Isotope Labeling with Amino Acids in Cell Culture), another popular labeling technique, shifts by 6 mass units [113], while the very first labeling method, ICAT (Isotope Coded Affinity Tag) makes the labeled peptides different by 8 mass units, in newer versions by 9 [54, 108].

The shift of 2-4 mass units, introduced by a heavy oxygen label, is not enough to separate the isotope distributions of the “light” (unlabeled) and “heavy” (labeled) peptides. In the previous chapter we described a method for quantitative analysis of HDX experiments. In this chapter, we will introduce an algorithm for automated quantification of $^{16}\text{O}/^{18}\text{O}$ labeling experiments in low and high resolution.

4.1.1 $^{16}\text{O}/^{18}\text{O}$ labeling

Isotopic $^{16}\text{O}/^{18}\text{O}$ labeling is a popular differential proteomics technique, which allows for relative quantification of peptides and proteins between two biological samples [163]. It is based on the ability of peptides to exchange the ^{16}O atoms on their C-terminals for ^{18}O atoms when immersed in H_2^{18}O . Label introduction is performed either during pro-

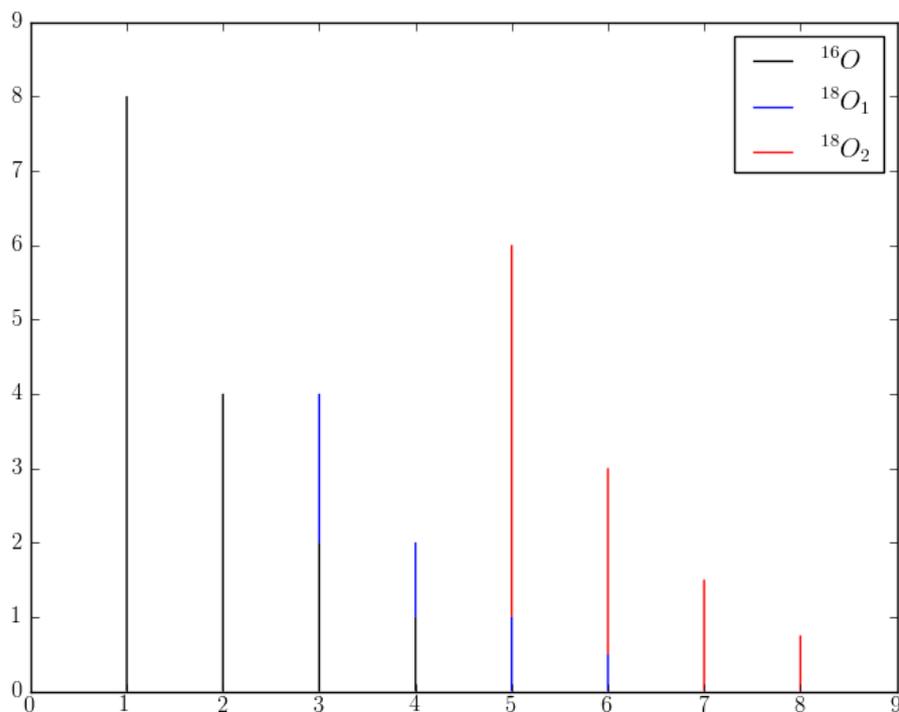


Figure 4.1: A schematic view of a spectrum, generated by a peptide of charge 1, with the unlabeled, single and double labeled forms present.

teolytic digestion or as an additional step after digestion by incubation with the protease [9]. After one or both C-terminal oxygens exchange, the mass spectrum of the labeled peptide is shifted by 2 or 4 mass units. After one of the samples is labeled, it is combined with the unlabeled sample and subjected to LC/MS. Spectrum of the combined sample is formed by the overlaps of the isotope distributions of labeled and unlabeled peptides, as shown in Fig. 4.1. Deconvolution of the overlap has to be performed before quantitative comparison can be made, and, once labeled/unlabeled peptide abundance ratios are known for multiple peptides, conclusions on the differences in protein expression levels can be drawn.

The main advantages of ^{18}O labeling include its universality, simplicity of the labeling procedure and low reagent cost, uniform labeling of all peptides (at least one ^{18}O atom is always incorporated) and low requirements to the sample size [47, 8]. Of these

properties, universality is arguably the most important. Unlike another popular relative quantification method, Stable Isotope Labeling with Amino Acids in Cell Culture (SILAC)[113], ^{18}O labeling is not limited to cells and can be applied to any kind of proteins with all possible post-translational modifications. In conditions where SILAC is applicable (*in vivo* label incorporation), the precision of the two methods has been shown to be comparable [88]. The experimental protocol is constantly being improved and issues like back-exchange or incomplete label incorporation, which caused a lot of problems in the early days of ^{18}O labeling, are now far less severe ([162, 104, 121] to mention just a few). The most important disadvantage of the technique is the small mass shift between labeled and unlabeled peptides, which calls for careful deconvolution of the isotope distribution of each peptide and severely complicates manual analysis. In the recent years, several solutions to this problem have been proposed, ranging from specially designed peak picking algorithms to generic full scale software suites, however, to the best of our knowledge, none of them except [165] gained a wide-spread popularity.

Johnson et al in [69] perform relative quantification based on accurate mass tags (AMTs) of peptides generated by the FT-ICR mass spectrometer. This approach allows them to forgo the exact peptide identification by tandem mass spectrometry and fit the sub-spectrum corresponding to each interesting peptide using averagine models. [103] proposes a method for ^{18}O labeling experiments using MALDI-MS with internal standards. The isotope distributions of ^{18}O -labeled peptides are estimated during a separate MS run and then used in a multivariate regression model along with native peptide isotope distributions. [38] builds on the work of [69] and [103] and uses multivariate regression with averagine models. This approach allows the authors to avoid performing a separate MS run with the labeled sample. However, although the multivariate regression as applied to ^{18}O labeling quantification is very well developed in [38] and includes corrections for back-exchange and impurity of the ^{18}O -water, the method works directly on the sub-spectrum corresponding to labeled/unlabeled peptide pair. This sub-spectrum, free of overlaps with other peptides, is supposed to be acquired by other means. Andersen et al in [4] also start from a peptide spectrum and use multivariate regression to fit the theoretical models for labeled and unlabeled peptide forms. However, to account for possible labeling inefficiency, they don't use averagine models for the labeled peaks, but run a label-swap experiment and estimate it directly.

Halligan et al in [55] propose a complete framework for the analysis of $^{16}\text{O}/^{18}\text{O}$ labeling experiments, performed on ion trap mass spectrometers. Zoom scans of peptides, identified by Sequest[44] are used for quantification. Sadygov et al [132] propose another software suite, MassXplorer, which is particularly well suited to low resolution ion trap instruments, but is not suitable for high resolution ion traps, such as Orbitrap. Both methods are limited to ion trap instruments.

For high-resolution data, [94] develops a quantification procedure, based on grouping the centroided peak lists and model fitting by linear regression via the non-negative least squares method. The regression is performed for each possible alignment of the theoretical distribution to the experimentally observed peaks. This procedure is only applicable to high-resolution data, but of all the methods reviewed here, it is the closest in spirit to our work and also to the work of Cox and Mann [32] on SILAC data quantification. A more detailed comparison can be found in the Discussion section.

msInspect[12] and Xpress[56] provide full scale frameworks for stable isotope labeling quantification. However, they base their quantification only on the ^{16}O and $^{18}\text{O}_2$ peaks, ignoring the possible incomplete label incorporation, which can appear as an $^{18}\text{O}_1$ peak (see also Fig. 4.1. Ye et al [164] demonstrated, by running an experiment with ^{18}O -labeled sample only, that incomplete label incorporation can reach 45%, with an average of 21% for the BSA protein. Consequently, ignoring the $^{18}\text{O}_1$ peak can severely underestimate the abundance of the labeled sample. [164] in turn propose to use trapezoid rule integration over retention time and Poisson distribution to model the labeled and unlabeled peaks in the isotope envelope of the identified peptides. This method also relies on the peptide spectra being extracted in advance and is not applicable to non-identified peptides.

All the methods described above are performing the quantification based on MS1 scans. A different approach is taken by [165, 160]. Since the label in the ^{18}O labeling procedure attaches itself on the C-terminal of peptides, it produces a mass shift not only in the MS1 spectrum, but also in all the y-ions of the MS/MS scan of the labeled peptide. MS/MS-based quantification can be more precise than the methods based on the full scan, as MS/MS spectra have a better signal-to-noise ratio. However, it is obviously only applicable to identified peptides and can not be used in exploratory analysis or in case when the peptides of interest are of low abundance (caused, by example, by a post-translational modification) and of unknown masses.

To summarize, to the best of our knowledge, there are currently no solutions available, which would take a raw 2D spectrum, either in high or in low resolution, and output labeled/unlabeled abundance ratios for both identified and non-identified peptides. For high-resolution data, the method of [94] provides a workflow of this type, however, as its quantification is based on non-negative least squares, it would not be able to handle peptide signal overlaps and noise in low resolution data. We have observed in Chapter3 that peak picking based on regularized regression techniques can successfully handle deconvolution of very complex signal mixtures of HDX-MS data. The goal of the project of this chapter was to create a framework, which could first localize the peptide signals in both low and high resolution data, then compute the abundances of labeled and unlabeled peptides by regularized regression and return the corresponding ratios.

The project has been realized in collaboration with the groups of Prof.Dr. Lehmann (DKFZ, Heidelberg) and Prof.Dr. Steen (Harvard Medical School, Children's Hospital Boston).

4.2 Methods

As the NITPICK peak picker successfully performed deconvolution of very complex mixtures in the data of Chapter 3, we decided to also use it for the quantification of overlapping spectra of $^{16}\text{O}/^{18}\text{O}$ experiments. The main difference between the HDX data and $^{16}\text{O}/^{18}\text{O}$ data from the peak picking point of view is that while in HDX data more peptide isoforms are mixed, the spectra are acquired for several D_2O incubation times, as well as for the reference sample, incubated in H_2O only. Comparison of peak picking results across spectra makes it more robust, as the deuterium incorporation is continuous and thus the decisions of the peak picker about the presence or absence of certain isoforms can be supported or overruled by the peaks found in the neighboring spectra. In contrast, $^{16}\text{O}/^{18}\text{O}$ labeling experiments only acquire one mixed spectrum for both the labeled and unlabeled sample.

An enormous simplification of the peak picking problem can be achieved by first segmenting the LC/MS spectrum into regions, corresponding to just a few overlapping peptide isotope distributions. The peak picker is then applied to the spectrum of the region, summed across its retention time domain. The retention time profile of each component can be reconstructed by an additional linear regression step. Segmentation greatly improves the signal-to-noise ratio of the data and, by limiting the range of the problem, makes the regression much faster and requires far less memory. In the following sections we first describe the segmentation method we applied to low resolution data of this and previous chapters and then another, sparse segmentation method which we developed specifically for the modern data of very high resolution.

4.2.1 Segmentation

Low resolution

For low resolution data, we followed the method of [90]. In this approach the two-dimensional LC/MS spectrum is treated as an image (with appropriate binning) and fast image segmentation algorithms are used to break it up into smaller pieces. The signal is first smoothed by convolution with a Gaussian kernel with variance selected so, that peaks at a distance of less than 1 Dalton don't get separated. Then, a difference of Gaussians filter is used to find blobs in the blurred image. These blobs serve as seeds for the third algorithm step, which performs seeded watershed segmentation [1]. Finally, the signal in the segmented regions is integrated over the retention time dimension and

NITPICK can be applied on the integrated spectrum.

We introduced an additional pre-processing step to deal with the problem of non-equidistant binning in the retention time dimension. In the Q-TOF mass spectrometer, acquisition of MS1 spectra is interrupted for MS/MS spectra acquisition and consequently the retention time values of individual MS1 spectra are not equispaced. Since the watershed segmentation treats the spectrum as an image, i.e. it assumes equidistant binning in both dimensions, the segmentation results can be flawed. In our modification, we first denoised the image by removing all signals with intensity less than 2, as well as all isolated signals, which did not have any neighbors in a 1Da range in m/z or in any of the neighboring scans in retention time. Then we re-binned the spectrum to 1 scan/second rate before the segmentation, using cubic spline interpolation for the missing values. The interpolated spectra were only used for segmentation, Lasso regression was performed on the original data values. The improvement in the segmentation results can be seen in Fig. 4.2.

High resolution data

The recent introduction of Orbitrap mass spectrometers [93, 63] brought a fundamental change to mass spectrometry-based proteomics, both on the biological and on the data analysis side [136]. The dynamic range of the instrument is much better than that of the Q-TOF. Orbitrap can take up to 6 MS/MS scans between regular MS1 scans, which brings a higher protein identification rate. In particular for peak picking, this translates into a much high number of exact theoretical peptide models, which makes the fitting more precise. But the most important improvement of the Orbitrap compared to time-of-flight instruments lies in its resolving power, which reaches 1,000,000 (normally used at 60,000), compared to 5,000-7,000 for Q-TOF. The MS1 spectra, produced at such resolution, have almost no overlaps, except the “perfect” ones which are produced by different isoforms of the same peptide 4.3. As we have observed in Chapter3 and in the low resolution part of this study, the NITPICK peak picker is very well suited for the deconvolution of such perfect overlaps, especially if the LC/MS spectrum is first segmented into parts of lower complexity.

However, the algorithm described above for low resolution spectra is not applicable for high resolution data. While the watershed-based segmentation works sufficiently well in low-resolution, the high-resolution data of the Orbitrap mass spectrometer can no longer be treated as an image and requires a sparse approach. The main argument for the necessity of sparse treatment of the data is the prohibitively large size of the image, which would result from a peak-shape preserving binning, such as the one used in the previous section. Also, this segmentation method does not make full use of such properties of high resolution data as the absence of peak overlaps. These considerations

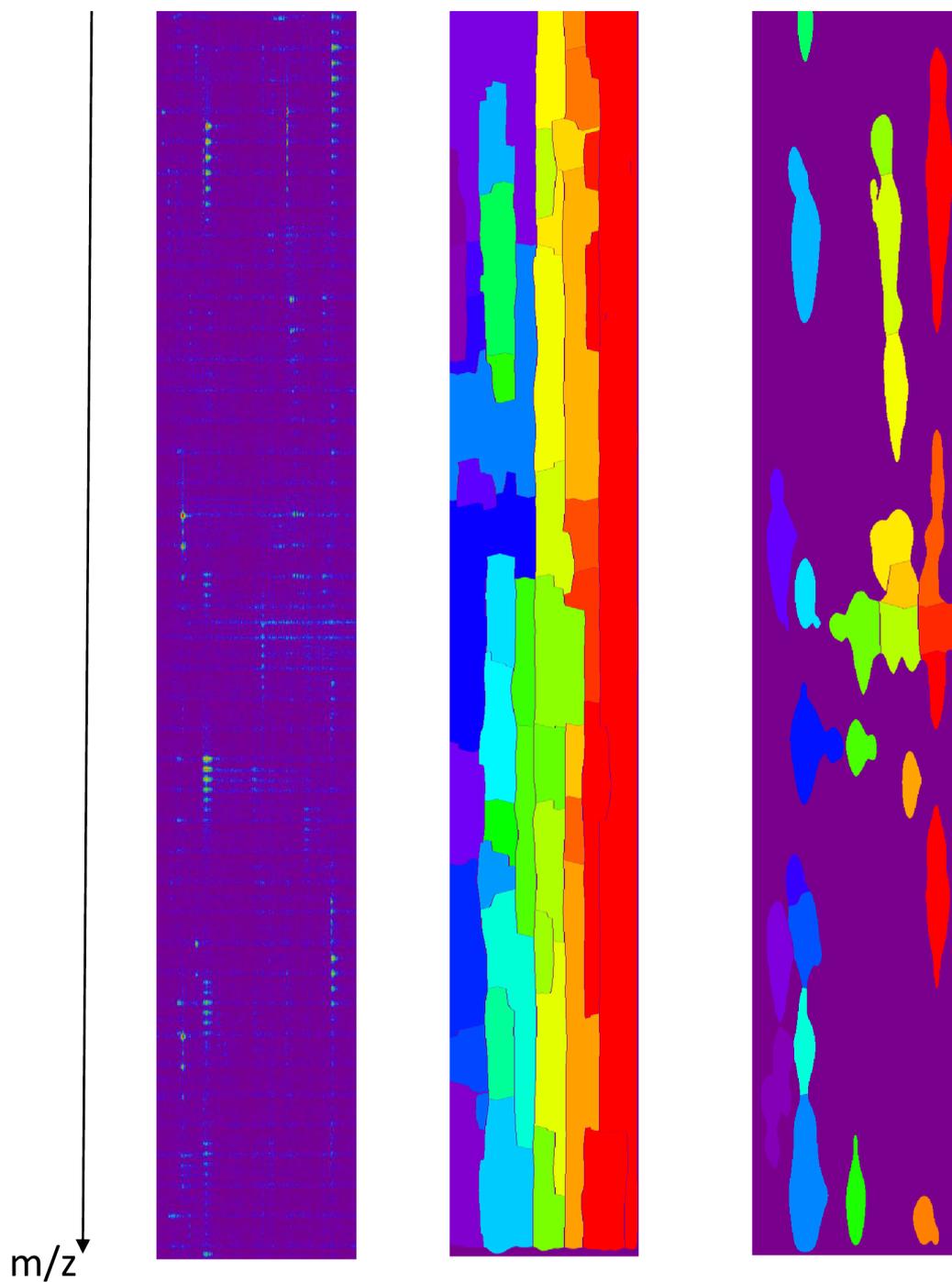


Figure 4.2: Segmentation of the low resolution data. Top: a fragment of the original spectrum. Middle: results of applying [90] method directly. Bottom: Applying [90] method after additional spectrum de-noising, re-binning and interpolation.

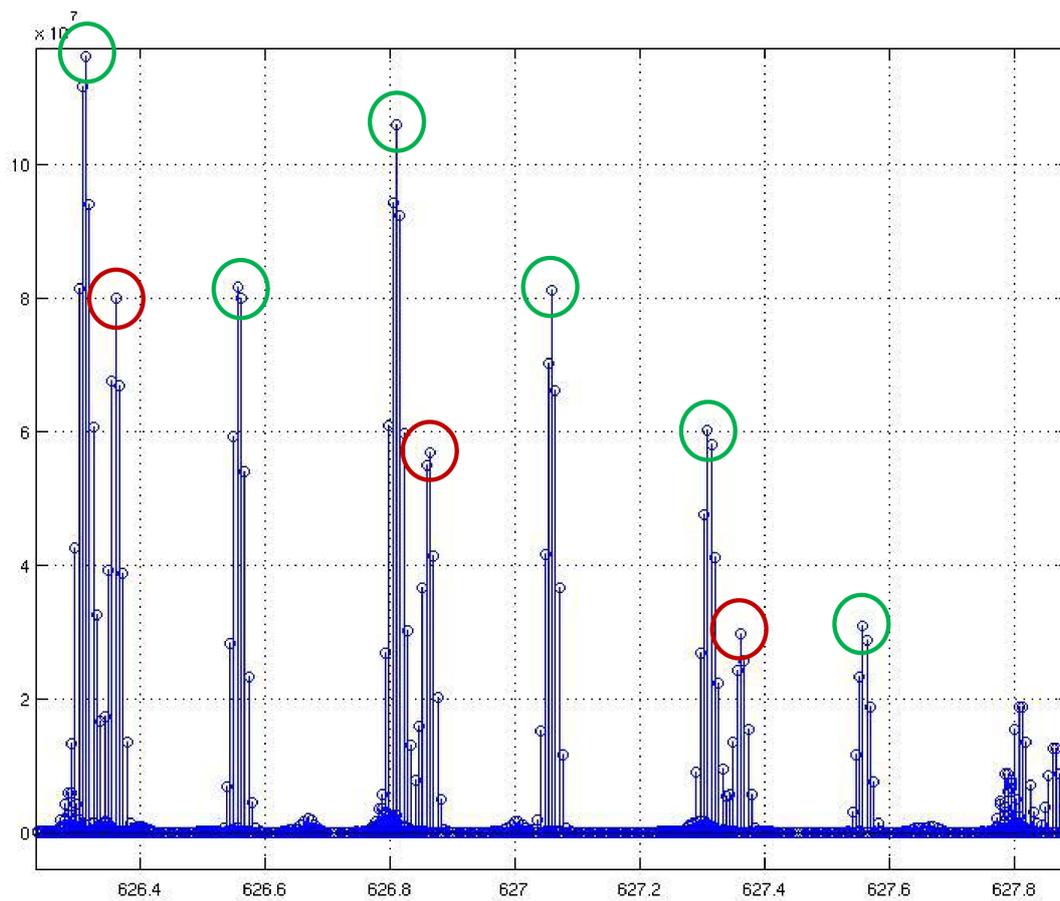


Figure 4.3: An example of the Orbitrap-produced spectrum, showing the advantages of its high resolution. The peaks, marked by green circles, belong to one peptide of charge two, the ones, marked by red circles, do a different peptide of charge one. These two peak groups are very well separated at Orbitrap resolution.

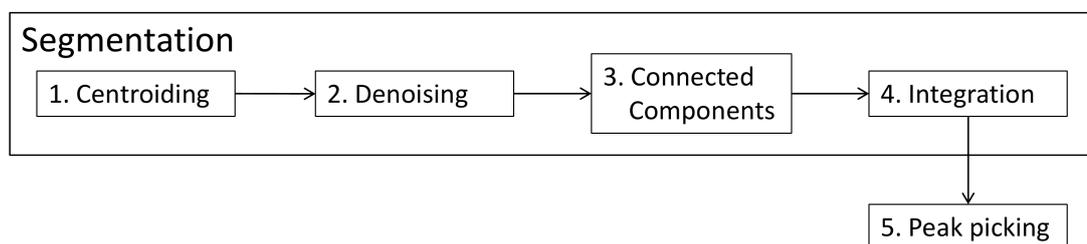


Figure 4.4: Workflow of the sparse segmentation algorithm.

led us to develop a new, sparse algorithm for the segmentation of high-resolution LC/MS data.

The algorithm is based on three assumptions about the shape of the peptide signal at high resolution:

- Peaks corresponding to the same peptide are aligned along the retention time dimension.
- Along m/z dimension, distances between the peaks of the same peptide are of the form $1/n$, $1 \leq n \leq \max(\text{charge})$.
- There is no significant overlap in m/z between the peaks, corresponding to different peptides.

A simplified algorithm scheme is presented in Fig. 4.4. Before describing each step of the algorithm in more detail, we need to introduce the data structure, which serves as the basis for all our spectrum operations.

kd-Trees

A kd-tree (k-dimensional tree) is a space partitioning data structure, which divides the full multi-dimensional space of data into regions with approximately the same number of points [13]. When the tree is fully grown, the end nodes only contain one point each and the regions at the intermediate levels of the tree are of approximately the same size. Assuming that all the data points are present at the time of tree construction, the tree growing algorithm can be formulated as follows:

1. Put all the points into the root node of the tree
2. Find the median of the data coordinates on the first axis

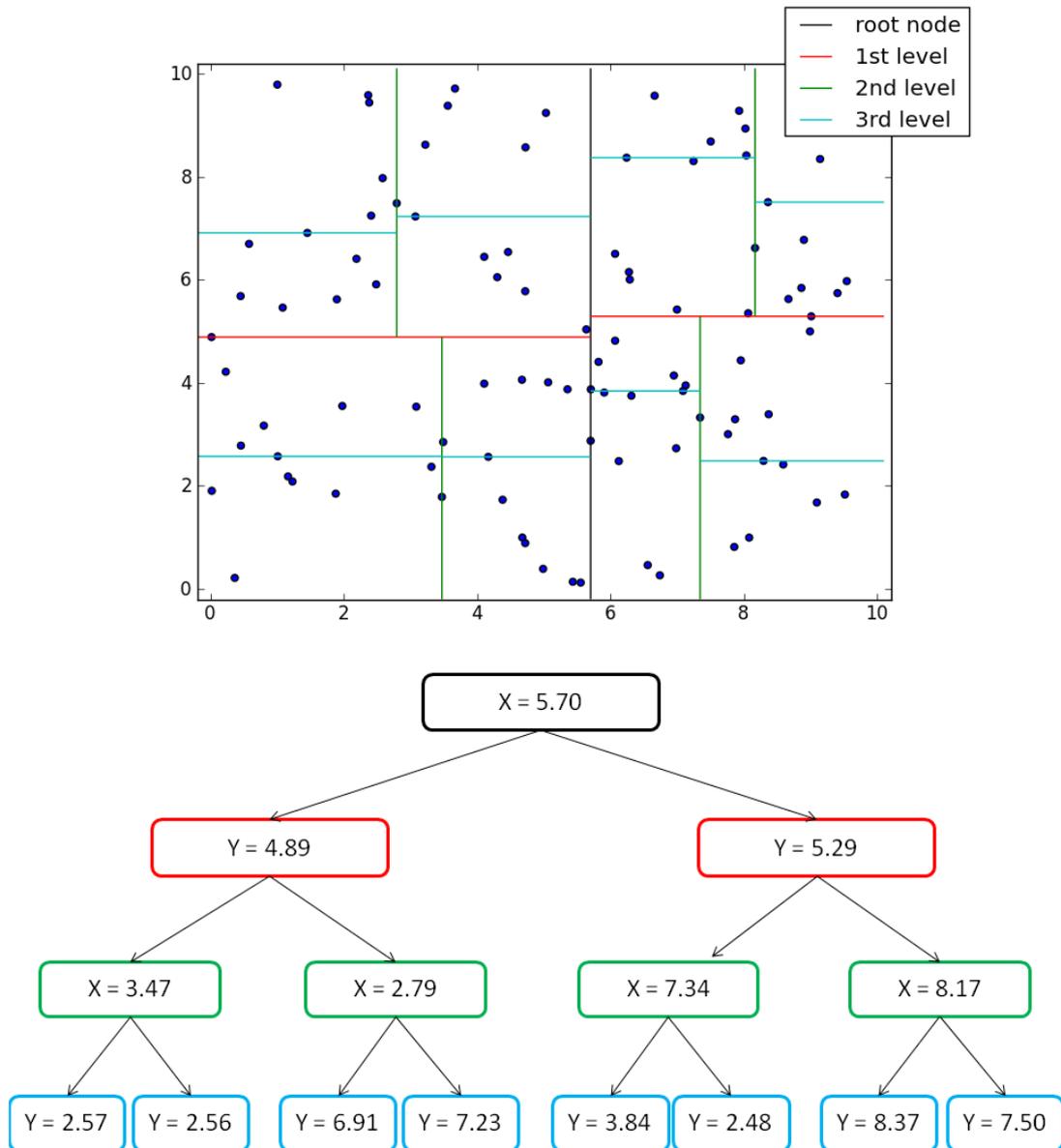


Figure 4.5: An example of how a kd-tree is built on 100 random points. Top: the splitting lines of the root node and the first 3 tree levels. Bottom: the split values, corresponding to the lines.

3. Add a left child to the root node which would hold the points with the chosen coordinate less than the median and a right child with the points with the chosen coordinate greater than the median
4. Repeat steps 1-3 for the children until the node size reaches a pre-defined value or one.

Several approaches exist for selecting which axis to cut on at each step. The simplest is to loop through the axis in order. A popular alternative is to cut on the axis with the largest data spread or even to select the axis randomly. Fig. 4.5 shows the top 4 levels of a kd-tree, built on 100 random points.

kd-trees were developed for fast spatial intersection queries, such as nearest neighbor or k-nearest neighbors. The computational cost of building or balancing a kd-tree is $O(n\log(n))$, with $O(n)$ of memory required for storage. Nearest neighbor queries in a balanced tree can be performed in $O(\log(n))$ time. Insertion or deletion of non-root node also is also $O(\log n)$. Partial match queries with t keys require $O(n(k-t)/k)$ time [13]. Queries to an unbalanced tree are slower, so it is usually recommended to either build the tree when all the points are available or to re-balance it after new elements are added or deleted. kd-trees are badly suited for very high dimensional data, but as LC/MS spectra are only two-dimensional, this limitation is not significant in our case.

We used the kd-tree implementation from the libsrckdtree library [125]. Libsrckdtree is a C++ header-only template library, which follows most rules of C++ STL associative containers. It supports keys of any dimensionality and values of all data types that can be stored in an STL container. This flexibility allowed us to perform all the multiple range queries of our algorithm on kd-trees only and to store our own data structures directly in kd-trees. To further illustrate this point, some of the kd-trees built during the execution of the program are listed below:

- a tree of peak centroids, with m/z and retention time values as key and a structure of the signals of the centroid and their weighted abundance as value.
- a tree of extracted ion chromatograms (XIC), with m/z and retention time values as key and a structure with the signals comprising the XIC, their weighted average mass, retention time, and other characteristics.
- a tree of connected components, with m/z , retention time and connected component number as key and XIC as value
- a tree of detected peaks, with monoisotopic m/z , retention time, connected component and maximum m/z of the peptide (for higher mass peptides, the largest peak is not the monoisotopic peak) as key and a peak structure of abundance, peptide and protein identification, and other peak features as value.

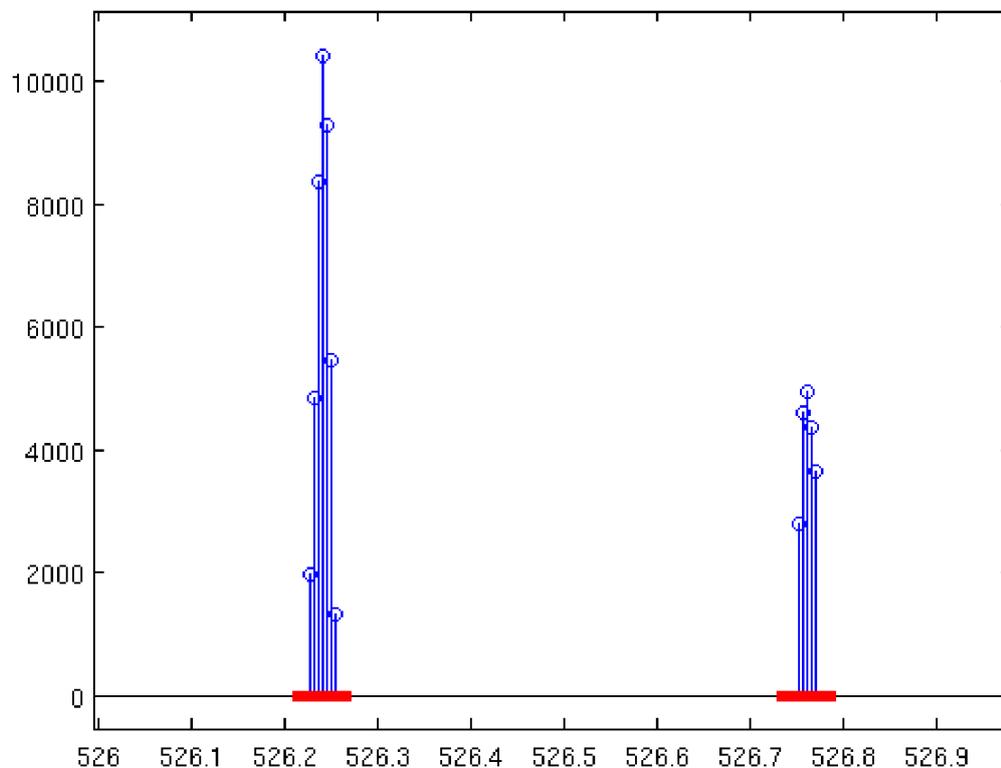


Figure 4.6: Centroiding along m/z dimension. The integration area for each peak is highlighted in red.

Centroiding

As we assume no significant overlap between the peaks of different peptides, bell-shaped peaks can be reduced to stick-like centroids without a substantial loss in information. To compute the appropriate width of a peak at a given m/z , we use the theory of mass spectrometer peak shape functions, developed in [72]. Based on the physics of the measurement process, [72] derives the dependency of the shape of peaks, produced by the Orbitrap, on m/z and computes the corresponding full width at half maximum (FWHM). The peak shape parameters have to be found by calibrating the function on the most abundant 1-dimensional spectrum. First we locate all the local maxima, which will serve as the base for centroids and then add to each local maximum the signals to the left and to the right of it until a saddle point to another peak reached or the distance to the centroid becomes larger than $2 \times \text{FWHM}$. Once all the peaks of the future centroid

are found, we fit a Gaussian to the three central points of the centroid, as described in [32]. Cox and Mann in [32] showed that a 3-point Gaussian fit provides a better abundance estimate than a weighted average and does not improve significantly if more points are added. Fig. 4.6 shows the integration area for several peaks. The resulting centroids are stored as a 2-dimensional kd-tree.

Denoising and XICs

For each centroid in the kd-tree we then construct its extracted ion chromatogram (XIC) by following its m/z value over retention time until a scan with a zero at this m/z value is found. The construction is performed as finding connected components in the implicit centroid graph, where two centroids are connected by an edge, if they have the same m/z value and are found in neighboring retention time scans. Such neighbors for each centroid can easily be found by querying the centroid kd-tree. XICs with very few members correspond to signals which only span very few retention time scans and are filtered out as noise. The results of such denoising can be seen in Fig. 4.7. The extracted retention time profiles are further broken into smaller XICs at the local minima and a kd-tree of XICs is created.

Connecting isotope clusters

The next connected components search is performed in the tree of XICs. The vertices of the implicit graph are connected by an edge if the m/z distance between them is either very small or of the form $1/n, 1 \leq n \leq \max(\text{charge})$ and their retention time profiles are correlated. As we are using regularized regression for quantification, we don't have to build different potential isotope clusters for all possible charge states, but we combine all the XICs at correct m/z distances together and rely on the regression procedure to find the correct charge assignment. Fig. 4.8 shows the isotope clusters, found in a small part of a spectrum. Note, how at this resolution, it's impossible to judge the quality of the segmentation by visual inspection in 2D. The right group of peaks on the bottom left image seem to form an isotope cluster in 2D, but closer inspection in 3D (bottom right image) shows that the algorithm was correct in its grouping decision.

4.2.2 Modeling and Quantification

The extracted segments are integrated over the retention time dimension. For the m/z values of the resulting 1D spectrum, we construct a set of models from the Mascot identification results and, for the m/z values without corresponding Mascot identifications, from fractional average models. The set of exact models is further expanded by adding the single and double labeled forms of all Mascot identified peptides. After the set of models is constructed, NITPICK peak picking is run to select the best fitting

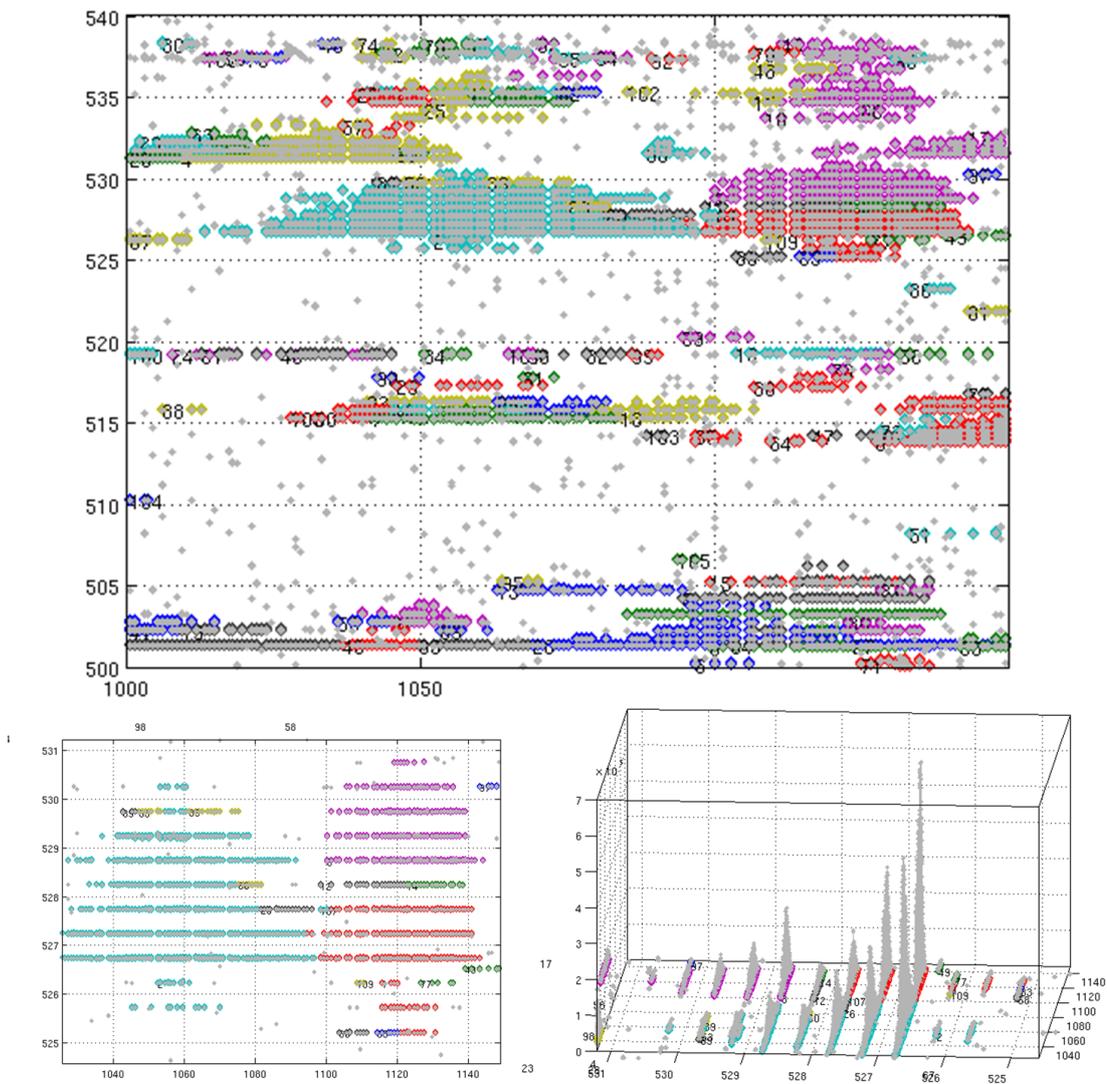


Figure 4.8: Finding connected components in a filtered spectrum. Top: 2D view of a part of a high-resolution spectrum. Bottom: a close up view of 2 peak groups in 2D and 3D. The 3D view proves the correctness of the algorithm decision to break the right peak group into 3 clusters.

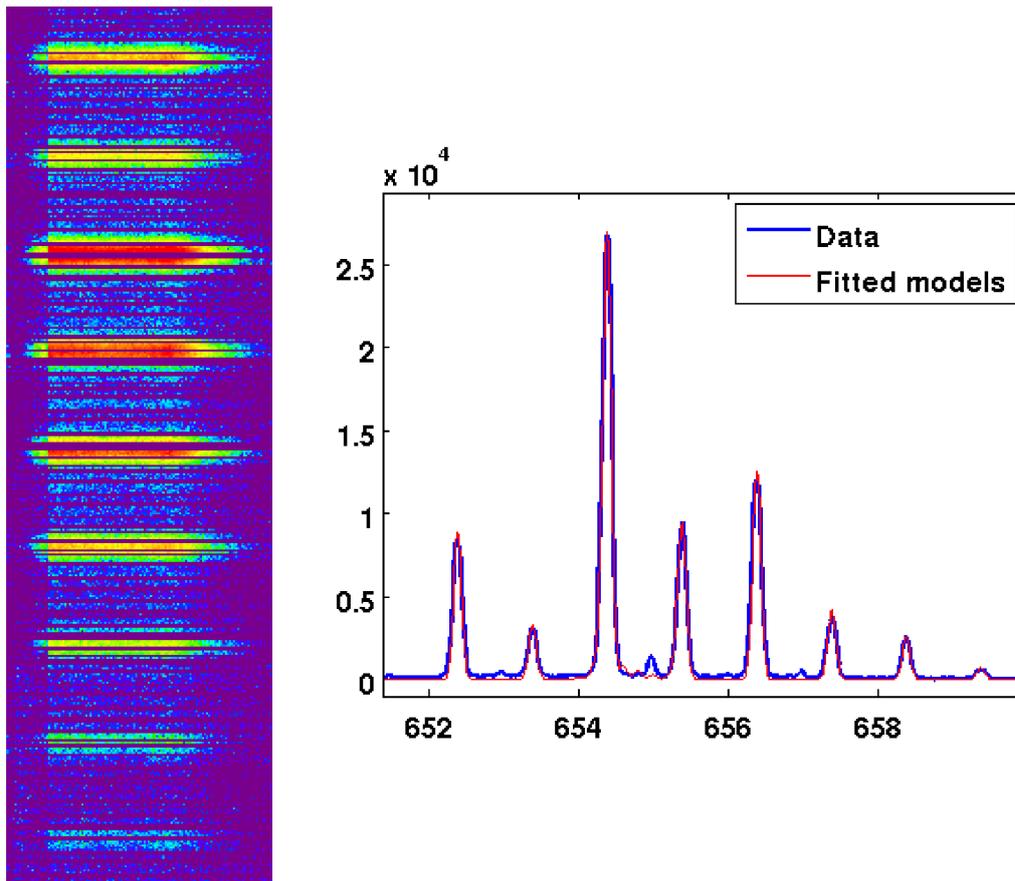


Figure 4.9: Results of peak picking on a segment of low resolution data. Left: the signal area, found by the segmentation algorithm. Right: NITPICK fit on the data. Note, how in the baseline of the spectrum NITPICK fit is not following the small noise peaks, because of the inherent sparsity of the LASSO algorithm.

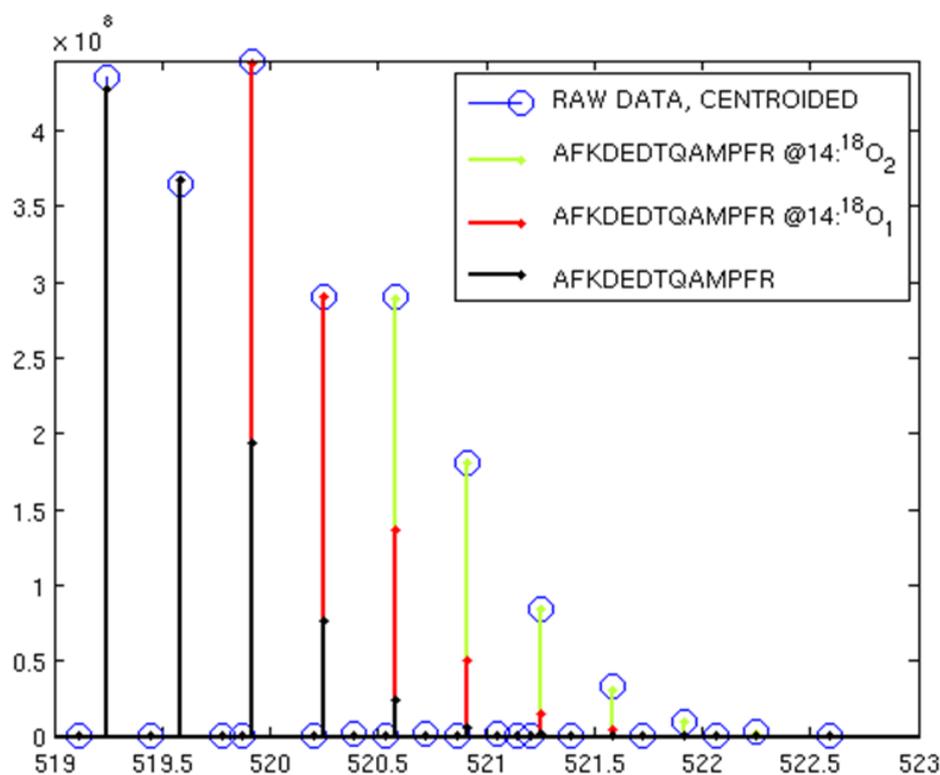


Figure 4.10: NITPICK fit of a segmented peak group in high resolution. In this case, the corresponding models come from Mascot identification results and represent AFKDEDTQAMPFR peptide and two of its forms, labeled by $^{18}\text{O}_1$ and $^{18}\text{O}_2$

ones. Figures 4.10 and 4.9 show the fit results for one peptide in low and high resolution.

All steps of the algorithm up to this point are very generic and not limited to the use case of $^{16}\text{O}/^{18}\text{O}$ labeling. Essentially, they can be applied to any kind of LC/MS proteomics data. If the data analysis only requires peak picking and quantification, the pipeline can be stopped at this step. If the user wants to analyze data for a stable isotope labeling experiment, he additionally has to provide mass differences between labeled and unlabeled peptide forms. The peak kd-tree is then queried for peak groups with peaks at the given m/z distances from each other. Finally, the software reports all such peak groups and the ratios of labeled and unlabeled peptide abundance in them.

4.3 Experiments and Results

4.3.1 Low resolution

Two datasets were used to evaluate the performance of the algorithm on low resolution data. The first one stemmed from a study on the influence of DNA damage on phosphorylation of proteins. Human KIAA0082 protein was digested in AspN and subjected to LC/MS on nanoAquity UPLC and Q-TOF (Waters). The second dataset was produced from chicken ovalbumin digested in trypsin during calibration of the instrument for the first experiment. Both datasets were acquired by the group of Prof.Dr. Lehmann in DKFZ, Heidelberg. The first dataset was used to evaluate the sensitivity of the method by comparing the number of $^{16}O/^{18}O$ peak groups found by the algorithm to those found by the human expert. Since the human expert performed the evaluation on the unsegmented (raw) spectrum, his estimate of the retention time of peptides sometimes diverged from ours. Also, some peptides with bimodal retention time profile were counted twice. To take these considerations into account, we performed two comparisons, one including and one excluding the retention time information of the peptides. For the second comparison all peptides with a given monoisotopic mass were counted as one regardless of their retention time.

Quantification results were validated by comparing the average labeled/unlabeled peptide ratios produced by the algorithm on the calibration datasets (MS1 information only) with those produced by the Mascot Multiplex tool [165]. For each dataset, we used 10% trimmed mean on the 50 most abundant peptide groups with total ion count of at least 5000 for the unlabeled peptide.

Sensitivity of the method was validated by comparing peak groups, returned by the automatic procedure, with those selected by the human expert.

Comparison type	Human expert	Our tool, in agreement with the expert	Our tool, total	Identified by Mas- cot
groups differ in m/z	212	170	588	26
groups differ in m/z and rt	303	216	1118	26

Table 4.2: Number of labeled/unlabeled peak groups, found by our algorithm and by the human expert. The last column shows how many of those were also identified by Mascot.

Quantification results were validated by comparing the average labeled/unlabeled pep-

ratio ratios, produced by the automatic procedure on the calibration datasets (MS¹ information only), with those produced by the Mascot multiplex tool [165].

True value	Mascot multiplex	Our algorithm
1	1.64	1.42
1.5	1.87	1.81
2	2.25	2.17
2.6	3.07	2.41
4	4.46	2.76
8	7.69	2.92

Table 4.3: Ratios of labeled to unlabeled peptide abundance in the peak groups found by our algorithm. First column: ratios used at sample mixing. Third column: ratios, found by Mascot multiplex tool. For each dataset, we used a trimmed mean (10%) on the 50 most abundant peptides with abundance of at least 5000 for the ^{16}O peak.

4.3.2 High resolution

For high resolution data, the method was evaluated on a mixture of tryptic digests of five commercially available standard proteins: ovalbumin (chicken), lysozyme (chicken egg white), beta casein (bovine), chymotrypsinogen A (bovine), ribonuclease B (bovine). Single proteins were digested in solution in either ^{16}O or ^{18}O water respectively and the resulting peptide mixtures were combined in ratios 1:1, 1:3, 1:5, 5:1, 1:10. Peptide mixtures were processed using an Eksigent nanoLC system coupled to an LTQ-Orbitrap mass spectrometer. One survey scan was performed at a resolution of 30000, followed by 6 MS/MS spectra of the most abundant precursor ions of each MS spectrum. The data was searched using an in-house version of Mascot 2.2 against a custom database restricted to these 5 proteins and known contaminants. All the data was acquired by Dr. Winter from the group of Prof.Dr. Steen, Harvard Medical School/Children’s Hospital Boston.

4.4 Discussion

Table 4.2 demonstrates high sensitivity of our algorithm, which was able to find most of the labeled/unlabeled groups, indicated by the human expert. As the goal of the human

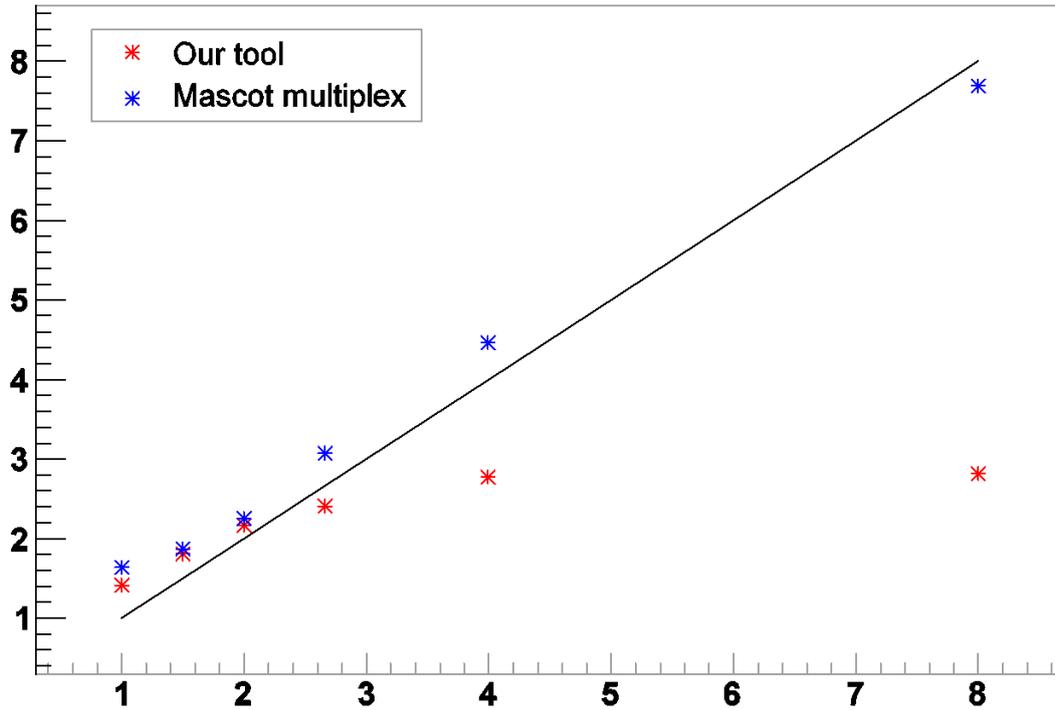


Figure 4.11: Labeled/unlabeled peptide ratios, found by Mascot multiplex tool and by our quantification procedure, compared to the true values used at sample mixing.

expert was to perform biological analysis of the data on the most interesting peak groups and not to find all such peak groups present in the data, additional peak groups found by our tool are not necessarily caused by lack of specificity. It is also interesting to note the very low Mascot identification rate for this experiment. Analysis of Table 4.2 suggests the following experimental strategy: after the first run of the experiment, find the masses of peptides which exhibit an interesting labeled/unlabeled ratio by using our automated procedure, add these mass values to MS/MS inclusion list and repeat the experiment to identify the corresponding peptides.

As for the precision of the quantification, Fig. 4.11 and Table 4.3 show, that while our tool provides very good quantification results for low ratios between labeled/unlabeled peptides, it seriously underestimates the true ratio once it gets to higher values (above four). Mascot multiplex tool does not suffer from this issue. As we only used the 50 most abundant peptides for our average estimates, this discrepancy can not be explained by noise signals weighting the average down. Closer examination of NITPICK fits for the

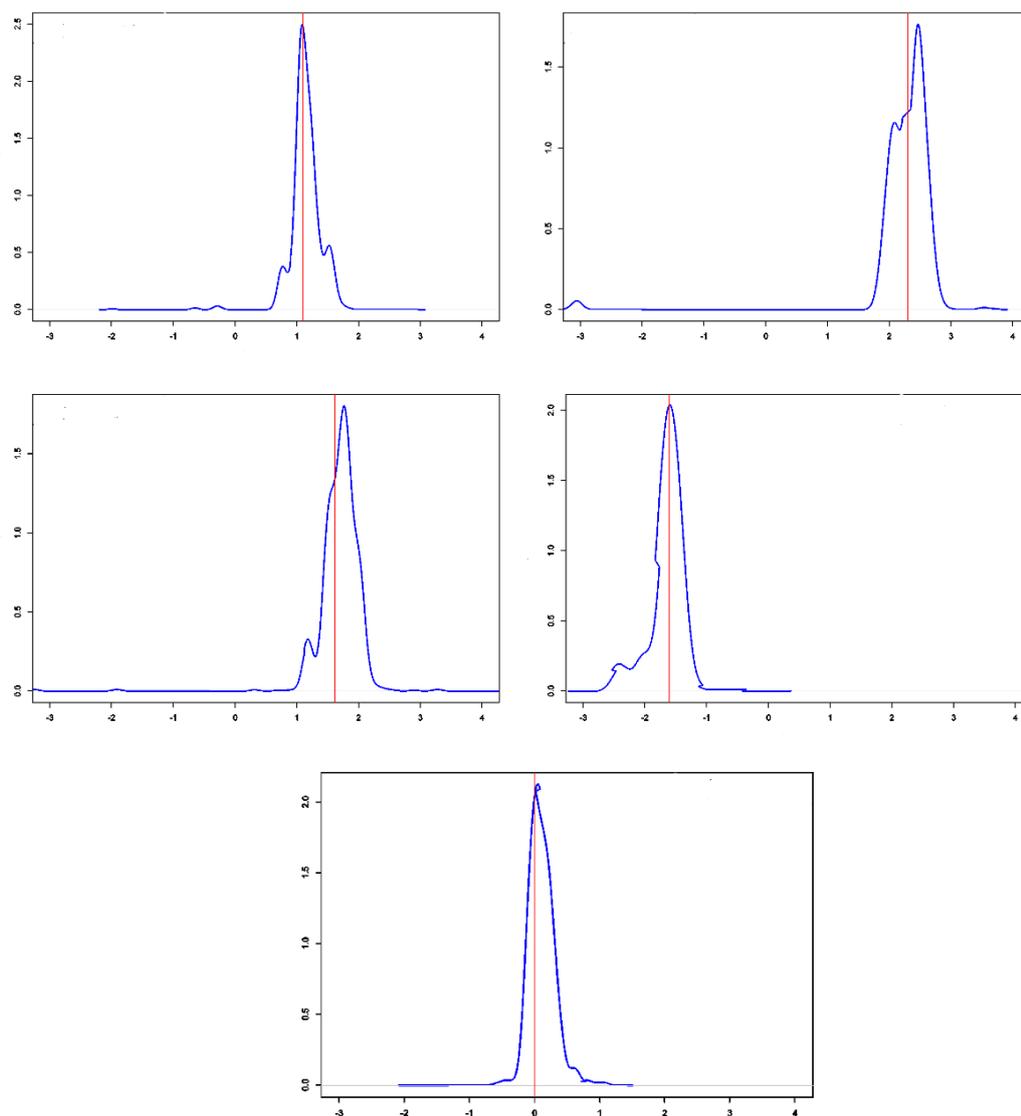


Figure 4.12: Density of the logarithm of labeled/unlabeled peptide ratios, found by our tool (blue), compared to the true values used at sample mixing (red vertical line). Top row: true ratios 1:3 (left) and 1:10 (right). Middle row: true ratios 1:5 (left) and 5:1 (right). Bottom row: true ratio 1:1. In all plots weighted density estimates were used with peptide weights based on the abundance of the unmodified cognate [25].

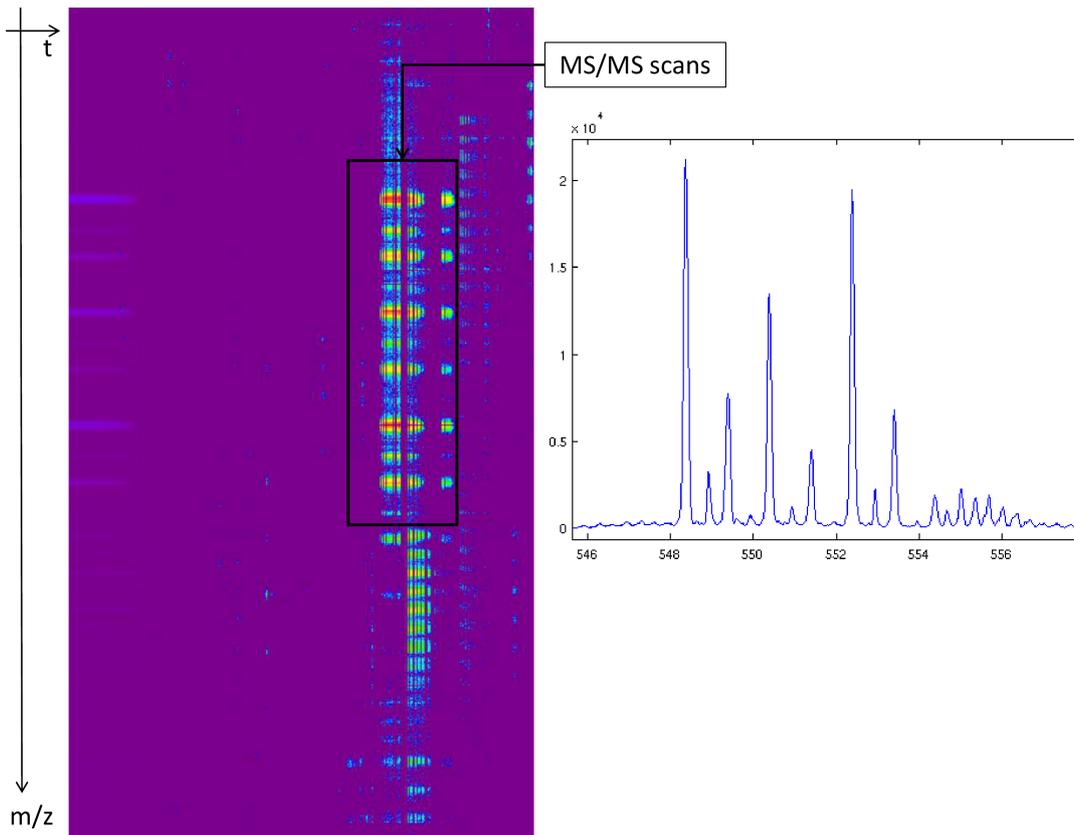


Figure 4.13: Signal of a peptide with 1 to 8 labeled/unlabeled ratio. The black arrow points to an empty region of the spectrum, showing the time period when MS/MS scans were taken. Three highest peaks correspond to unlabeled, $^{18}O_1$ and $^{18}O_2$ labeled peptides. It is quite clear, that the ratio of the areas under these peaks cannot be 1 to 8.

very abundant peptides did not uncover any lack-of-fit problems, however, it brought to our attention the fact, that the sum spectra of these peptides do not seem to contain as much signal from the labeled cognate, as one would expect after mixing labeled and unlabeled samples in 1:8 ratio. Fig. 4.13 shows one of the examined peptides. Our collaborators from Prof. Lehmann's group, who performed the experiments and acquired the data, came to the conclusion, that wrong ratio values were a property of the MS1 spectra of this experiment, caused, probably, by the overflow in the instrument. Since Mascot multiplex quantification is based on MS/MS spectra, it was not affected by this problem, but, as we observe in Table 4.2, for biological experiments Mascot identification rates can be very low.

For high resolution data, the automated quantification results follow the true ratios very well (see Fig. 4.12). These results serve as indirect evidence for the good performance of the new segmentation algorithm, as NITPICK would not be able to obtain correct quantification results if presented with incomplete peptide signal area. In fact, it would be very difficult to test the segmentation algorithm directly, since, as one can observe on Fig. 4.8, correct manual segmentation in 2D is impossible at this resolution, and re-verifying all segmented regions in 3D would be a very time-consuming task even for a mass spectrometry expert. As an alternative, one could simulate a spectrum and compare the segmentation results to the ground truth used to generate the data, but to the best of our knowledge, no realistic simulations of Orbitrap spectra are currently available.

Two other methods, close in spirit to the proposed segmentation algorithm, have already been proposed in [32] and [94]. An important distinction is that our segmentation does not have to produce consistent isotope clusters. Unlike the NITPICK regularized regression peak picking that we use, [94] employs non-negative least squares for quantification and has to rely on goodness-of-fit statistics to select the right assignment of the monoisotopic mass and charge. While this approach might be sufficient for ^{18}O labeling, it will no longer be applicable if labeling with more overlapping states is used, especially if not all those states are present at each peptide. [32] was developed for quantification of SILAC experiments and does not have to deal with overlaps in isotope patterns at all. However, it also has to find the correct charge state. In contrast, our approach is not limited by the number of labeled states or mass differences between them and can be applied to any kind of stable isotope labeling.

Chapter 5

Automated Detection and Segmentation of Synapses in Serial Electron Microscopy Images

5.1 Introduction

5.1.1 Chemical synapse and information transmission in the central nervous system

One of the major differences between neurons and other cells is their inability to divide after differentiation. With notable exceptions, such as the neurons of the olfactory system, most of 10^{11} neurons of the human brain are already present at birth [36]. The tremendous plasticity of the brain, its ability to learn, store memories and recover from injury are not based on the new neurons being formed, but on the re-wiring of the existing neural circuits.

Connectivity between neurons is provided by electrical and chemical synapses. Electrical synapses or neuronal gap junctions are formed at the very narrow (~ 1 -1.5nm) gaps between neurons, and transmit information by direct electrical coupling of the neurons. This transmission mechanism makes them faster than chemical synapses, however, they can not amplify the signal they receive. Electrical synapses form a minority of all synapses in the brain and are outside the scope of this chapter. In the following pages the word "synapse" will only be used to refer to chemical synapses.

The chemical synapse is the predominant means by which information is transferred

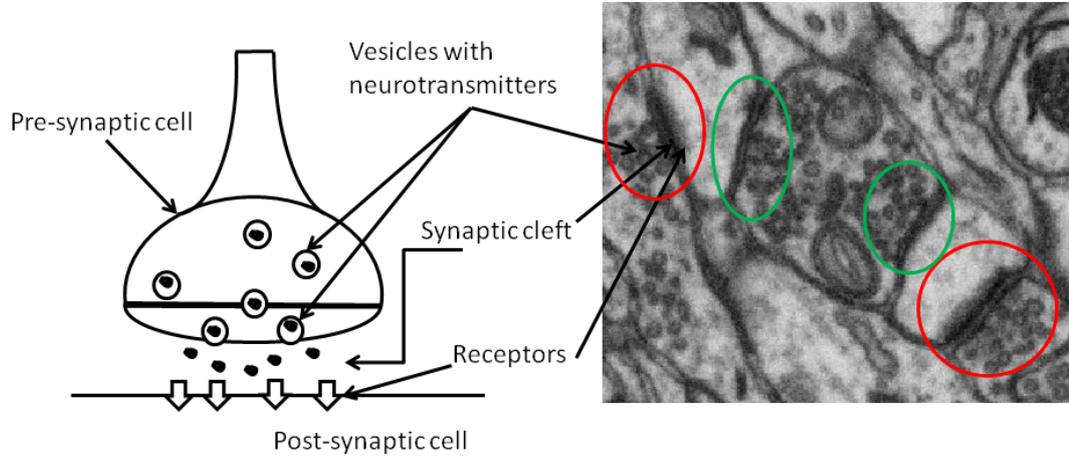


Figure 5.1: Left: synapse components. Right: excitatory (red ovals) and inhibitory (green ovals) synapses as seen in FIB/SEM images of $5 \times 5 \times 9$ nm resolution.

and stored in the central nervous system. The anatomy of a synapse is illustrated in Fig. 5.1. Briefly, the following components of a synaptic connection can be defined:

- Pre-synaptic cell, the origin of the signal
 - Vesicles with neurotransmitter molecules
- Synaptic cleft, 30-40nm wide gap between the cells
- Post-synaptic cell, the recipient of the signal
 - Neurotransmitter receptors, forming the Post-Synaptic Density (PSD)

Following [22], information transmission through a synapse can be summarized as a sequence of steps:

1. The pre-synaptic cell is depolarized and its Ca^+ channels open
2. The pre-synaptic cell releases neurotransmitters into the synaptic cleft.
3. The neurotransmitters in the synaptic cleft bind to the specific receptors in the membrane of the post-synaptic cell.
4. The post-synaptic cell opens or closes chemically gated ion channels. Other changes in the post-synaptic cell are also possible, including regulation of gene expression or metabolic changes.

Two major types of synapses are usually recognized, depending on the influence they have on the behavior of the post-synaptic cell. Release of neurotransmitters from an *excitatory* or *asymmetric* synapse increases the probability of the action potential in the post-synaptic neuron. The most common excitatory neurotransmitter is glutamate and excitatory synapses with glutamate release are referred to as "glutamatergic". Conversely, neurotransmitters of an *inhibitory* or *symmetric* synapse decrease the probability of the action potential. γ -Aminobutyric acid (GABA) is the most common inhibitory neurotransmitter and the corresponding synapses are referred to as "GABAergic". The differences between these two synapse types as they appear in electron microscope images are illustrated in Fig. 5.1, where excitatory synapses are surrounded by red ovals and inhibitory - by green ovals. Asymmetric synapses have a larger, more electron-opaque post-synaptic density and many spherical neurotransmitter vesicles. The appearance of the pre- and post-synaptic areas of inhibitory synapses is more symmetric (hence their alternative name) and the neurotransmitter vesicles have a more oval shape. It has to be noted, that even expert neurobiologists can not always distinguish these two types on electron microscopy images. The majority of synapses in the brain are excitatory, however, the exact ratio varies significantly between different parts of the brain.

Neural function and plasticity manifest through the changes in synapse size, shape and distribution. Localization and segmentation of synapses in brain images can thus provide essential information to the understanding of the neural circuitry.

5.1.2 Synapse detection

Despite the brilliant advances in light microscopy [53, 130, 10], detailed structural analysis of synapses is still only possible with electron microscopy. With serial section Transmission Electron Microscopy (ssTEM), synaptic density can be estimated by manually counting synapses within a large volume, or by stereological extrapolation from 2D images [145, 49, 96, 30]. In the early days of synapse morphometry, volumetric estimates based on extrapolation from single plane measurements were in use [95]. These were replaced by an unbiased estimate from a disector - a pair of parallel sections at a known distance, which allows for counting of arbitrary particles without making assumptions on their shape, size or orientation [145, 96]. The number of synapses in the volume between two dissecting planes is estimated as the number of synapses, visible in the first plane but not in the second. Two borders of the image (for example, top and left) are considered to be "forbidden lines" and synapses touching them are also not counted. The only assumption of the disector method is that the synapses in both planes can be identified unambiguously. While this assumption is true for synapses with the synaptic cleft perpendicular to the slicing plane, detection of synapses parallel or at a low angle to the slicing plane is much more challenging. In a recent publication devoted to this issue Kubota et al.[86] observe that as much as one third of all synapses oriented at a

low angle to the slicing plane are missed by the human annotators.

This problem has been alleviated by the progress in the electron microscopy instrumentation. The recent introduction of focused ion beam/scanning electron microscopy (FIB/SEM) [80] with isotropic resolution approaching 5 nm has now opened the door to a direct detection and segmentation of all synapses in large volumes of tissue, without the need to resort to extrapolation from paired slices. When searching for synapses, the human observer is no longer limited to the imaging plane projections of the volume, but can also explore the planes orthogonal to it. A protocol for manual synapse detection in FIB/SEM data has recently been proposed in [100]. Still, even for the best quality EM images, manual detection of synapses remains a difficult, error-prone and time-consuming task, which calls for automated protocols to overcome the tedium of manual analysis.

To detect synapses in EM images, human experts follow a set of morphological criteria: the presence of the pre- or post-synaptic densities, a visible synaptic cleft and a nearby cluster of at least two vesicles. If an automated protocol was to be based on these criteria directly, it would require a segmentation of the entire volume to find the membrane apposition sites and a full segmentation of ultra-cellular structures to detect vesicles. Although the problem of automated segmentation of neural tissue has advanced significantly in recent years, it is not yet fully solved [28, 106]. Also, automated segmentation of vesicles is nontrivial, especially at lower resolution, and has not received much attention in the literature. Rather than explicitly implementing the currently used criteria, machine learning allows to imitate the overall decisions of a human. The prediction rules are learned automatically from examples, provided in the form of annotated images (the training dataset). A meaningful measure of success is how well the automated predictions on a separate test set agree with those of the human.

Our contribution proposes an automated approach of this type and shows, through quantitative evaluation on a set of 111 synapses, that state-of-the-art machine learning methods can now achieve detection rates comparable to those of humans for asymmetric synapses in FIB/SEM data. Even though our approach does not explicitly implement the morphological criteria listed above, it finds enough evidence in the geometric features, extracted from a local neighborhood of each voxel, to mimic the decisions of the human expert. This work was performed in collaboration with the group of Dr. Knott (EPFL, Lausanne).

5.1.3 Related work

In the field of neuroscience, recent influential work along these lines has mostly focused on tracing and segmentation of neurons. Depending on the Electron Microscopy technique used to acquire the images, the methods can be divided into those treating the volume as a 3D stack and those which perform the segmentation in 2D slices and then link the segmentation results in 3D. The former methods are applied to SBFSEM or FIB/SEM data with isotropic resolution, the latter to ssTEM data with very anisotropic resolution. For SBFSEM data Andres et al [6] first use a Random Forest classifier on 3D image features to produce a voxel-wise membrane probability map and then build supervoxels on the probability map to create an over-segmentation. The excessive supervoxel edges are removed by an additional optimization step, which takes into account the possible topology of neurons and membranes. Turaga et al and Jain et al [151, 67, 68] propose to obtain a voxel affinity graph by training a convolutional neural network directly on the image intensities. For ssTEM data, another contribution of Jain et al [66] introduces a topology based metric to compare neural segmentations, which is consistent with the segmentation quality requirements from the circuit reconstruction point of view. This metric is then used as a cost function in the optimization procedure. Mishchenko in [105] uses a ridge detector to detect membranes and then interpolates small membrane gaps by anisotropic contour completion based on fuzzy logic. Kaynig et al[74] proposes a different method of gap completion by introducing a special energy term, which accounts for good continuity of membranes. Combined with Random Forest predictions for membranes in a larger energy minimization procedure, this additional term significantly improves segmentation results. Jurrus et al[70] take a very different approach based on the auto-context principle developed in [150]. They construct a serial neural network, which uses predictions in a voxel neighborhood produced by the previous network as additional features for the next network.

Since all currently available methods require manual proof-reading, several groups have developed entire software frameworks, which combine image pre-processing (registration, de-noising, contrast correction, etc), fully manual or semi-automated segmentation and interactive proof-reading [5, 59, 112, 24, 23, 119]. Fiji [92] - a general purpose biological image processing framework - also has a plug-in for 2D ssTEM image segmentation [73].

The problem of ultracellular structure segmentation has so far received less attention, however, interesting methods have been demonstrated in [91]. Lucchi et al[91] detect mitochondria in FIB/SEM image stacks by first segmenting the stack into supervoxels and then partitioning the supervoxel graph by energy minimization with unary terms including global shape cues from 3D Ray features and pairwise terms based on learned boundary appearance models. The performance of the algorithm is comparable to that of human annotators. The importance of this contribution is not limited to the problem

of mitochondria detection, as often the errors of membrane segmentation are caused by mitochondria located very close to the cell boundaries. Detecting and removing mitochondria before or simultaneously with boundary detection could improve neuron segmentation results.

For synapse detection, two automated methods have recently been proposed for fluorescence light microscopy images [61, 134]. Since these rely on fluorescent pre-labeling of all synapses, they are not applicable to EM data. For ssTEM data, Mishchenko et al [106] automatically detect synapses in the course of a large-scale semi-automated volume reconstruction effort. Once all the cell membranes are found, the synapses on them are located by a single-layer logit neural network classifier, trained on membrane intensity and width features. However, this approach relies on correct partitioning of the entire volume into cells, which is still impossible by fully automated means [28]. This chapter describes our approach to detection and segmentation of synapses in FIB/SEM data, which does not require the volume segmentation. Some considerations on the possible extension of the algorithm to ssTEM data will be provided in the Outlook section.

5.2 Methods

The input data for the algorithm consists of scanning electron micrographs of neural tissue, provided as a pre-registered image stack, and user labels on a tiny subset of the data. The labeling can be very sparse, as shown in Fig. 5.2.

5.2.1 Random Forest classifier

Random Forest is an ensemble-based method of classification, introduced by Leo Breiman in [19]. As all ensemble methods, it works by aggregating the decisions of simpler classifiers, in this case - decision trees. While an individual decision tree is prone to overfitting, a “forest” of decision trees, each of which has only seen a subset of the training data, has very favorable generalization properties (see Fig. 5.3). Aggregation of classifiers trained on bootstrapped samples of the training data is referred to as “bagging” and was also proposed by Breiman in [17]. Random Forests introduce further randomization into bagged decision tree ensembles by not only providing different training sets to different decision trees, but also forcing them to make their decisions on different sets of features at each node split. Growing a decision tree for the Random Forest can be described in the following sequence of steps, supposing the training dataset contains N points of N_f features along with their labels:

1. Randomly draw N points from the training dataset with replacement.
2. Randomly draw m_{try} features.

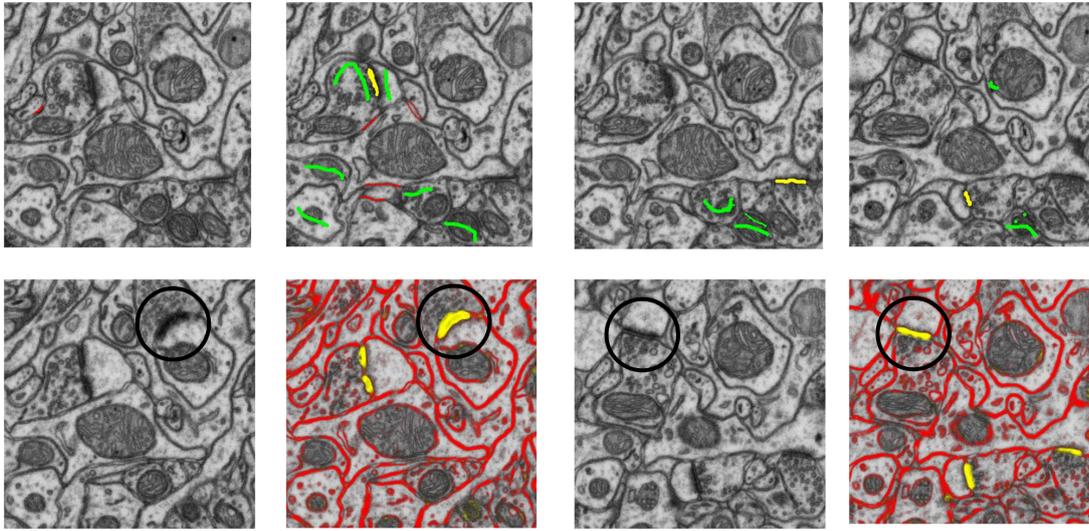


Figure 5.2: User labels and algorithm predictions. Top row: the complete set of user annotations for the first training set (20 brush strokes in total), with yellow labels for synapses, red for membranes, green for the rest. Bottom row: raw data and algorithm predictions on two other slices in the first training set. In black circles: some unlabeled synapses and their probability maps. The color intensity corresponds to the certainty in the prediction, predictions for green class are omitted for clarity.

3. According to some split criterion, choose on which of the m_{try} features to split.
4. Split the data into the left node with the points for which the split feature is less than the split value and the right node with the points for which it's larger than the split value.
5. Repeat steps 2-4 until the tree is grown to purity, i.e. the lower nodes of the trees only contain points with the same labels.

The most popular split criteria include the decrease in Gini impurity and cross-entropy. Given a node with N samples to classify into K categories of labels, define $P_j = \frac{1}{N} \sum_{i=1}^N I(x_i == j)$ - the proportion of labels of category j . Cross-entropy can then be defined as $-\sum_{j=1}^K P_j \log(P_j)$. The Gini impurity of a node and the decrease in Gini impurity resulting from a split of a node into a left and right child are defined as

$$I = \sum_{j=1}^K P_j(1 - P_j), \quad \Delta I = I_{parent} - \frac{N_{Left}}{N} I_{Left} - \frac{N_{Right}}{N} I_{Right}.$$

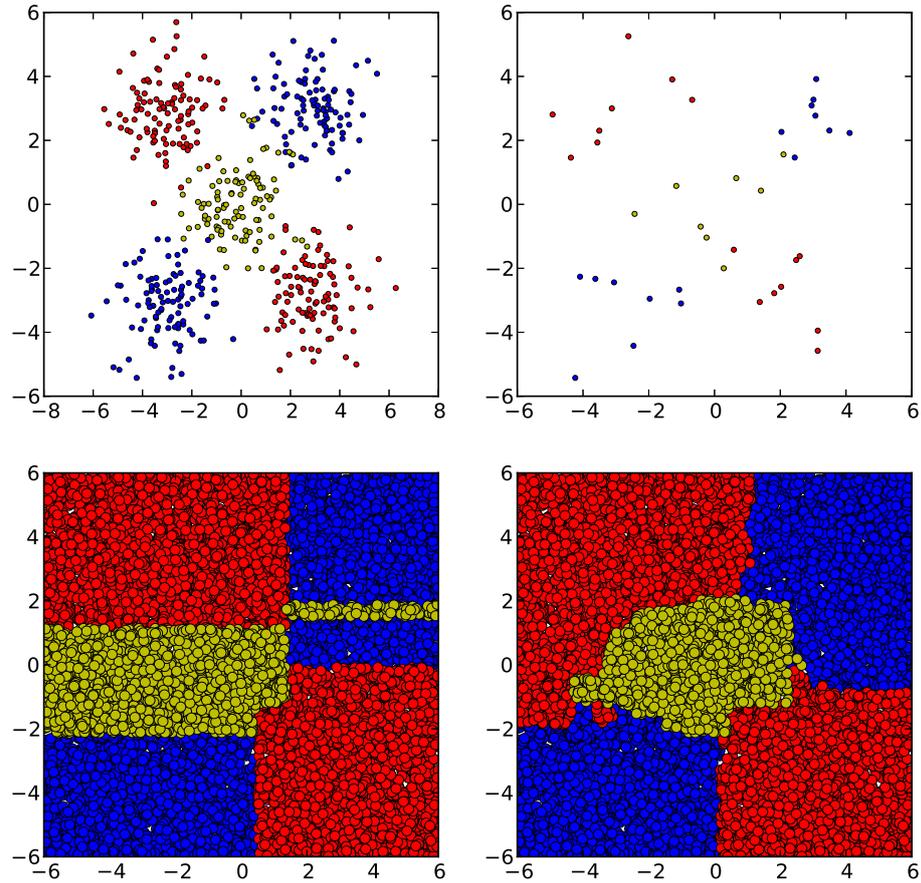


Figure 5.3: Generalization properties of a single decision tree vs. a forest. Top left: the underlying distribution of the data. Top right: the training sample. Bottom left: decision boundary between the three classes for a single decision tree. Bottom right: decision boundary of a Random Forest of 255 trees. The bottom left plot for the single decision tree shows clear signs of overfitting to the training dataset, which is not the case for the bottom right plot of the Random Forest.

The best split is then the one which provides the largest decrease in Gini impurity. Gini impurity can be interpreted in the following way: suppose that a sample in the node is classified according to the current category distribution in the node. Gini impurity then represents the misclassification rate.

Since the random draws of the samples for each tree are done with replacement, each tree only sees a subset of the training data. The rest of the training data (approximately one third) is referred to as "out-of-bag" data. Out-of-bag samples can be used to estimate the misclassification rate of the forest by letting each tree predict the labels of its out-of-bag samples and averaging the number of incorrect predictions over all trees. This estimate is quite close to the real misclassification rate, provided that the training data captures the variability of the test data well enough.

Besides finding the best split, Gini impurity measure gives rise to the following method of feature importance estimation: for each tree, consider the nodes which split on a given feature. The average Gini impurity decrease of these splits over all such nodes and all trees indicates how important the feature is for the forest. An alternative estimate of feature importance is given by permutation importance. For each tree, permute the values of feature X_j in the out-of-bag samples and compute the new out-of-bag error. The feature importance estimate is then given by the average difference between the old and new out-of-bag errors for all trees [20].

Random Forest only depends on two parameters: the number of trees and the number of features drawn at each split. Random Forest has been empirically shown to be fairly robust to their choice, and to provide very good results for a broad range of applications [52, 26, 35, 99]. Besides its excellent generalization properties, this simplicity of setup makes it especially attractive for applications geared towards non-expert users.

5.2.2 Feature selection

The standard EM protocol used to prepare the brain tissue for imaging gives high contrast not only to synapses, but also to other cellular structures, such as mitochondria. As a consequence, the classification cannot simply be based on the raw intensity values of individual voxels. Instead, more informative features are required that also encode geometrical properties of 3D voxel neighborhoods. Different features represent different properties of these neighborhoods and should be selected so as to allow for an effective discrimination of the labeled classes. For example, as synapses are darker than intracellular space, the average intensity would serve as a good feature to distinguish these two, but would not help to separate synapses from membranes or mitochondria. Edge detectors respond strongly to synapses, as the synaptic cleft is positioned between two membranes, but regular membranes and endoplasmic reticulum are also detected. Tex-

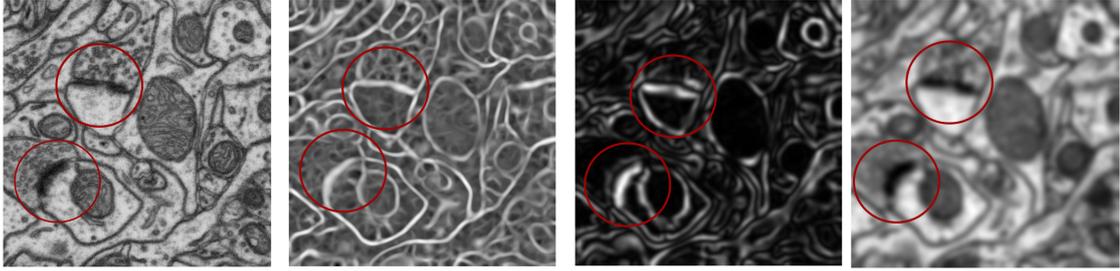


Figure 5.4: An example of feature response on a part of a stack. From left to right: raw data, first eigenvalue of the Hessian matrix, first eigenvalue of the structure tensor, intensity smoothed by a Gaussian, all with $\sigma = 5$.

ture of synapses, especially at the post-synaptic densities, is quite distinctive, however, texture features also pick up thick mitochondrial membranes. Rather than devise decision rules by hand, we use a variety of intensity, edge and texture features at different scales and apply statistical learning from a labeled training set to infer robust classification rules. Fig. 5.4 shows the feature response on a small part of a stack.

Since features have to be computed for every voxel, memory consumption has to be taken into account for large volumes. To allow running of the algorithm on a modern desktop PC rather than a high-end server without compromising classification accuracy, we performed selection of features, based on their Gini importance. The final list of 38 features is provided in Table 5.1. Although the user is free to re-adjust the list and try out new feature combinations, we do not expect it to be necessary, except for the adjustment of the neighborhood sizes to the resolution of the data. Due to boundary effects in the feature computation, the performance of the algorithm can decrease for voxels very close to the limits of the dataset, such as the voxels of the first and last scan of the stack.

Importance of 3D features

Since the the resolution of FIB/SEM microscope images is nearly isotropic in all 3 dimensions, we computed all the features in full 3D neighborhoods of each voxel, using the same σ value for all three dimensions. To test if the third feature dimension gives a noticeable effect on the Random Forest prediction quality, we also performed the training and prediction on 2D features, using the labels of Fig. 5.2 and features of Table 5.1. Fig. 5.5 shows the results of the comparison. Clearly, the 3-dimensional context of a voxel is very important for its successful classification, and using 2D features only produces a lot of false positive synapse detections. It has to be noted, that this experiment does not fully reflect the difficulty of synapse detection in anisotropic EM images,

#	Feature	# of channels	Sigma
1	Eigenvalues of the Hessian matrix Eigenvalues of the structure tensor	3, 3	3.5
2	Eigenvalues of the Hessian matrix Eigenvalues of the structure tensor	3, 3	5.0
3	Intensity smoothed by a Gaussian	1	5.0
4	Intensity smoothed by a Gaussian	1	3.5
5	Intensity smoothed by a Gaussian	1	1.6
6	Gradient magnitude and Laplacian of a Gaussian Difference of Gaussians	1, 1, 1	3.5
7	Intensity smoothed by a Gaussian	1	1.0
8	Eigenvalues of the Hessian matrix Eigenvalues of the structure tensor	3, 3	1.6
9	Gradient magnitude and Laplacian of a Gaussian Difference of Gaussians	1, 1, 1	5.0
9	Gradient magnitude and Laplacian of a Gaussian Difference of Gaussians	1, 1, 1	1.6
10	Intensity smoothed by a Gaussian	1	0.7
11	Eigenvalues of the Hessian matrix Eigenvalues of the structure tensor	3, 3	1.0

Table 5.1: Local neighborhood features, used for voxel classification. The "Sigmas" column shows the standard deviation of the Gaussians, used for smoothing the data. This parameter effectively determines the size of the necessary voxel neighborhood. For the eigenvalues of the structure tensor, the outer scale parameter was set to $\sigma/2$, for the difference of Gaussians the second Gaussian sigma was set to $0.66 \cdot \sigma$.

such as the ones produced by ssTEM, since, although the features are 2-dimensional, the z-resolution of the images is excellent and no smeared membranes or mitochondria are present.

5.2.3 Probability map thresholding

The obtained probability maps are smoothed by convolution with a Gaussian with a standard deviation of 5 voxels to avoid local discontinuities caused by noisy voxel-wise predictions. Uncertain detections are then filtered out by considering only those clusters of voxels with synapse probability greater than a given threshold and with size of at least 1000 voxels. The lower limit for the size filter was computed as the approximate volume occupied by two vesicles at the given data resolution. The probability threshold can be interactively adjusted by the user. After thresholding, only the "cores" of synapses, i.e. areas of very high synapse probability, are left. These cores underestimate the real size of synapses, so to transition from detection to a proper segmentation we relax the

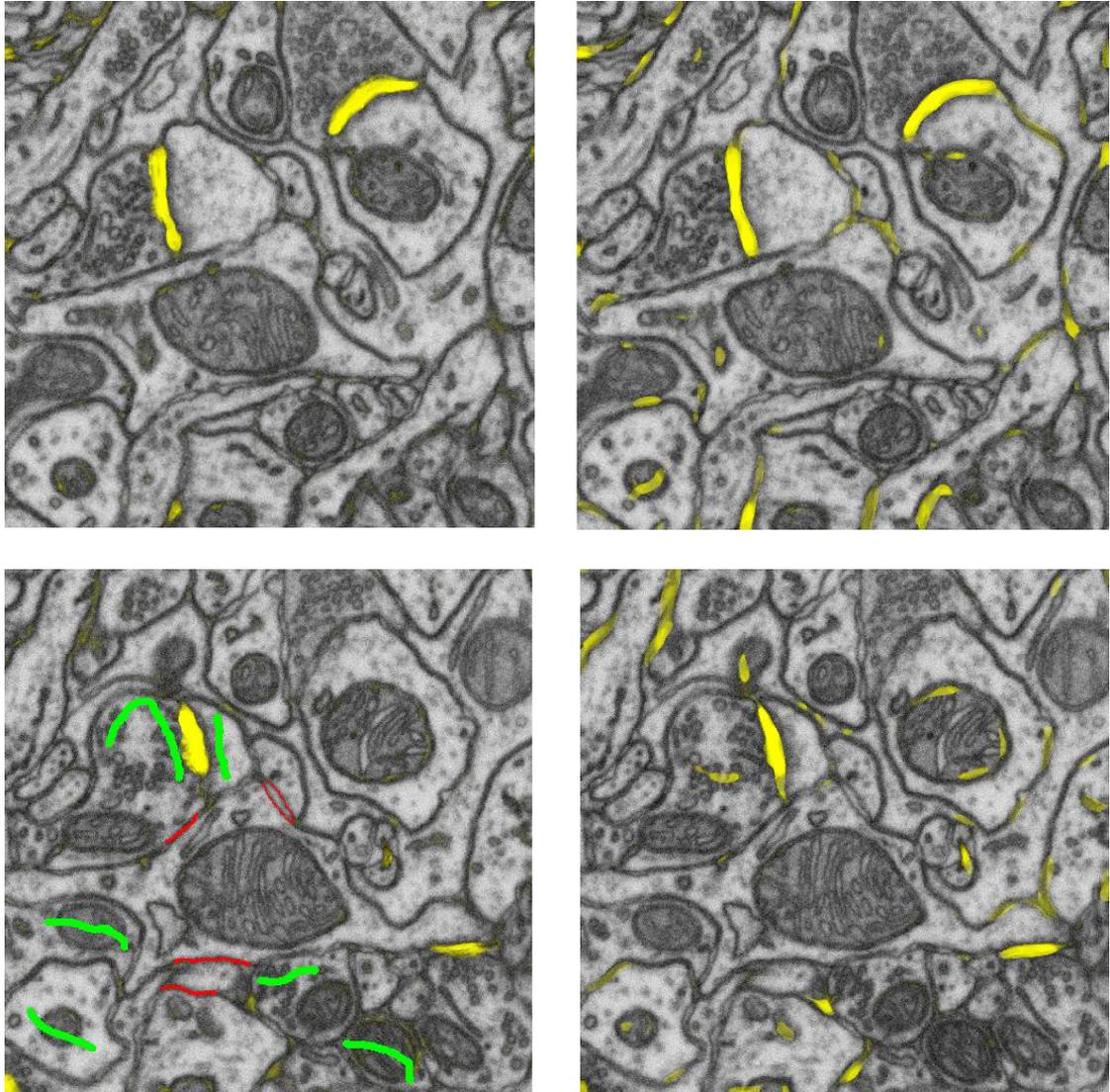


Figure 5.5: Comparison of classifier predictions using 3D(left column) or 2D(right column) features. Top two images show synapse probability maps on an unlabeled image, bottom two images - probability maps on a labeled image. Note a large number of false positive assignments to the synapse class in both cases.

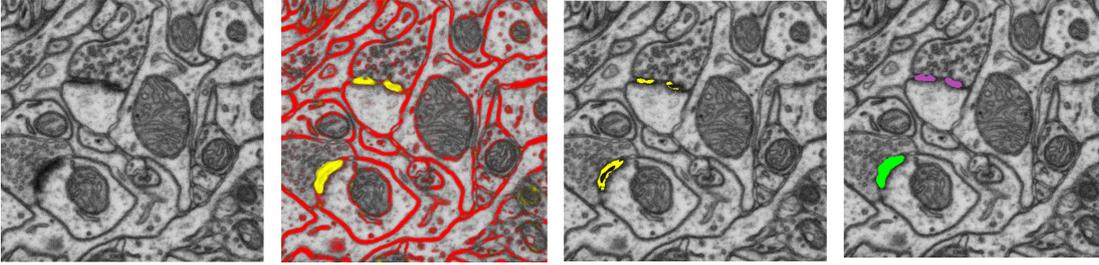


Figure 5.6: Thresholding of probability maps. Left to right: raw data; Random Forest prediction results for synapse and membrane classes; synaptic “cores” - voxels with more than 90% synapse probability; final segmentation.

synapse probability threshold to 0.5 for all voxels that are adjacent to synaptic cores. This sequence of steps is illustrated in Fig. 5.6.

5.2.4 Software

ilastik

On the software side, we build on ilastik [143]. The freely available ilastik toolkit provides an intuitive interface for classification and segmentation of 2D and 3D data and allows users without experience in machine learning to perform fairly complex tasks on their data. ilastik includes the Random Forest classifier and a set of generic image features, which can easily be expanded via the feature plug-in mechanism. The segmentation is performed by the assigning each pixel to the most probable class. In the interactive mode, it allows the user to immediately see the effect of newly added labels on the classifier’s predictions, and therefore reduces the necessary labeling time. Once the classifier has been trained on a representative subset of the data, predictions on a very large dataset can be performed off-line in batch-processing mode. ilastik is written in Python and uses c++ and Vigna image processing library [83, 82] for computationally intensive tasks.

Here we present and evaluate an extension of ilastik which includes interactively adjustable thresholding and finding of connected components, as well as a possibility to display the found objects in 3D with the help of the VTK toolkit [135]. The current synapse detection pipeline for a large dataset can be summarized as follows:

1. After loading a small subvolume of the data into ilastik, label a few synapses and some background pixels in the interactive prediction mode, until the prediction results look satisfactory. In the feature selection dialog, choose the features from

Table 5.1.

2. Interactively find a threshold on the probability map, such that most of the false positive detections disappear, but the synapses are still visible.
3. In ilastik batch processing module, perform the prediction with the trained classifier on the full volume. The prediction is performed block-wise and does not have to load the complete file in memory.
4. Run a block-wise thresholding and filtering script on the obtained full volume probability maps. The only required parameter of this script is the threshold value, however, the minimum and maximum values of the size filter can also be specified in case they have to be adjusted for different image resolution.
5. Load the results produced by the script into ilastik. The detected synapse objects can now be browsed and visualized in 3D, as shown in Fig. 5.7. Integration of ilastik with the VTK visualization allows the user to jump from a 3D object directly to its position in the image stack. The remaining false positive detections can be removed either at this step or at the next one, using the detection summary report.
6. Finally, an HTML summary report can be created for more convenient proof-reading and further analysis (Fig. 5.10).

In case only a small data stack is available, no offline processing is needed and all the computations of step 4 above can be performed by clicking a few buttons in ilastik. No step of the pipeline is specific to the task of synapse detection, which makes it applicable to the general problem of small object detection in 3D image stacks by changing only the image features.

5.3 Experiments

5.3.1 Data acquisition and generation of the gold standard

Data, used in this section, has been acquired and kindly made available by the group of Dr. Graham Knott, EPFL, Lausanne. The quantitative validation of the automated synapse detection procedure, as well as the evaluation of the human experts' error rate, was carried out on a test dataset of 111 asymmetric, presumed glutamatergic, synapses. The test dataset consisted of 409 scanning electron micrographs from layer 2/3 of the adult rat somatosensory cortex. The tissue preparation methods followed the protocol previously described in [80] and were performed in accordance with the procedures approved by the Office Vétérinaire Cantonale Lausanne (license number 2106). Briefly, the brain of an adult rat was fixed by cardiac perfusion of 2.5% glutaraldehyde, and 2% paraformaldehyde in phosphate buffer, it was then vibratome sectioned and slices

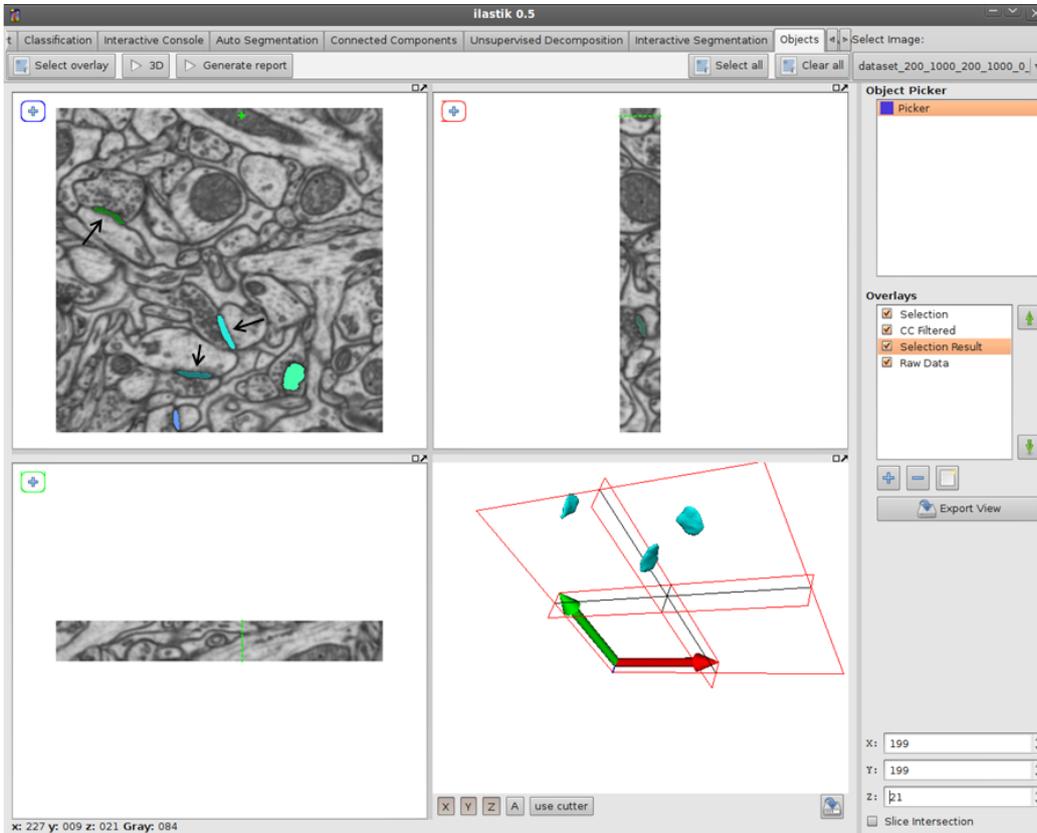


Figure 5.7: A screenshot of ilastik object browsing tab. The three synapses to which the arrows are pointing are displayed in the 3D viewer in the lower right corner.

from the somatosensory cortex were stained with buffered potassium ferrocyanide and osmium, followed by osmium, and then uranyl acetate. These stained sections were then dehydrated and embedded in Durcupan resin. The selected region was trimmed in an ultramicrotome and mounted onto an aluminium SEM stub for imaging in the FIB/SEM microscope (Zeiss NVision40), using a scanning electron beam at 1.3 kV with a current of 1 nAmp. Backscattered electrons were collected via the energy selective in-column detector (EsB) using a grid tension of 1.1 kV. The milling was achieved with a gallium ion source at 30 kV with a current of 700 pAmp. The acquired images were of 5 nm per pixel resolution with each image 1948×1342 pixels in size. The milling depth was measured at 9 nm per slice. Such high z-resolution allowed treating the data as one 3D volume of $1948 \times 1342 \times 409$ voxels instead of a collection of 2D slices.

Synapses in the dataset were manually annotated by three independent human experts

according to morphological criteria, including the presence of a pre- and post-synaptic density, as well as clustered vesicles close to the pre-synaptic membrane [81]. The human experts were researchers with experience in the analysis of electron micrographs of brain tissue and counting synapses in serial images. TrakEM2 plug-in of the FIJI framework [23] was used for the annotation. One of the experts only had four hours to label and verify the complete dataset, while the other two experts were not limited in time and took several hours longer. The annotation of each expert included positions and approximate size of detected synapses, denoted by "ball" labels from TrakEM2. Some examples of expert labels can be seen in Fig. 5.11D,5.11E,5.11F. Each expert first analyzed the dataset independently from the others and the resulting three sets of annotations were compared automatically to find all discrepancies. Since the automatic comparison procedure found differences between the expert annotations, these cases had to be re-examined jointly by all experts to establish a gold standard annotation. Synapses touching the left or top border of the image, as well as those touching the last slice of the stack, were excluded from the final count. For evaluation purposes, we also excluded synapses which had their center in the first slice of the stack, to avoid the border effects described in the next section. The resulting set of 111 synapses formed the gold standard and was used to estimate the error rates of both the original human annotations and the results obtained by the algorithm.

5.3.2 Error criteria

For the evaluation of the error rate, a synapse candidate was considered to be a false positive, if its "ball" label from the human expert or its shape segmented by ilastik did not overlap with any ball in the gold standard dataset. If such an overlap was found, the corresponding gold standard ball was removed from the set of possible matches. Conversely, a false negative detection was counted, if a ball from the gold standard did not overlap with any of the synapse candidates; if such an overlap was found the corresponding synapse candidate was removed from the set of possible matches. Human errors were additionally reverified manually, to avoid assigning a detection error in case of a geometric disagreement between labelers, i.e. when two labelers labeled the same synapse at positions so far from each other, that their "ball" labels did not overlap.

5.4 Results

5.4.1 Human experts

The expert which only had 4 hours to label and verify the synapses, missed 11 synapses and found 20 false positives. The other two experts, unlimited in time, made 2 and 3 false negative and 7 and 8 false positive detections respectively. Most expert mistakes were made for different synapses, which is in line with the observations of [59] about

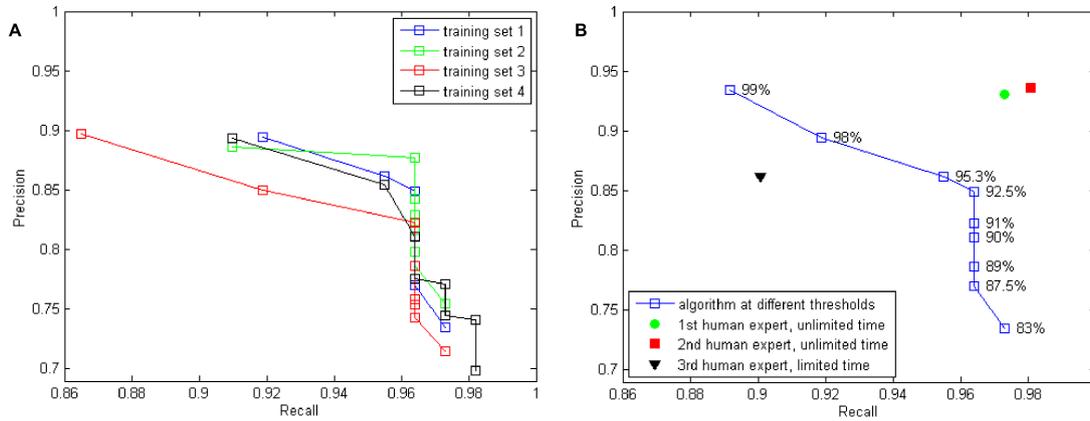


Figure 5.8: Precision and recall of the algorithm and the human experts. Recall was calculated as the (no. of true positives)/(no. of synapses in the ground truth), precision as the (no. of true positives)/(total no. of synapse candidates). **A:** Precision and recall of the algorithm results for the four different training sets. **B:** Precision and recall of the algorithm compared to the human experts with and without the time limit. The synapse probability threshold values are annotated next to the corresponding points of the curve.

attention-related errors of expert annotators of neurobiological images.

5.4.2 Automated detection

To quantitatively assess the algorithm performance and its stability with regard to the training data, four training sets were created from images acquired in the same experiment, but not overlapping with the test set. The four training sets were located in different parts of the image stack and contained approximately the same number of voxel labels. For each training set, 2-3 synapses were labeled, and for each of those synapses it was sufficient to only label it in one of the slices. Adding more labels did not improve the classification performance, as long as the already labeled set represented the data well, which can be judged, for example, by looking at the current algorithm predictions for some non-labeled synapses (Fig. 5.2, bottom row). Although the software can discriminate an arbitrary number of categories, we found three-class labeling of synapses vs. membranes vs. the rest of the tissue to produce the best results. One can also use a binary setup with synapses vs. the rest, but then the labeler has to take extra care to annotate enough membrane voxels to obtain a representative sample of the background. Adding more classes, for example, for the mitochondria, did not help the classification. Our first training set is illustrated in Fig. 5.2 and a performance comparison for the

different training sets is shown in Fig. 5.8A.

After training, the classifiers were applied to the test dataset, and thresholding with different sensitivity levels was applied to the resulting synapse probability maps. Precision and recall of the algorithm, depending on the threshold, are illustrated in Fig. 5.8 (using the training set from Fig. 5.2 for Fig. 5.8B). Recall was calculated as the (no. of true positives)/(no. of synapses in the ground truth), precision as the (no. of true positives)/(total no. of synapse candidates). The voxelwise threshold for the detection of synaptic cores was specified as the probability of the synapse class. For the training set from Fig. 5.2, the best algorithm performance was at the threshold of 98%, with recall of 0.92 and precision of 0.89. Overall, the algorithm performance is better than that of a human expert working with a four-hour time limit (0.9 recall and 0.86 precision), but worse than that of domain experts with unlimited time, who, in practice, worked on the problem on two consecutive days, though not all day long (recall of 0.97 and 0.98 and precision of 0.931 and 0.936). A comparable recall value for the algorithm (0.96) was achieved at precision of 0.85. Labeling the training set, computing its appearance features and training the classifier took approximately 15 minutes. Running the algorithm on the full test dataset took several hours, however, no user interaction was needed during this time.

A 3D view of the synapses detected by the algorithm based on the training set from Fig. 5.2 (with probability ratio threshold of 92%) is illustrated in Fig. 5.9.

The human labelers only detected synapses and specified their approximate size by the ball labels, while the algorithm segmented synapses, i.e. listed every voxel belonging to a synapse candidate. Since the real synapses are not spherical, these human annotations can not serve as voxel-level gold standard. Overall, the question of the definition of the synapse size is not yet settled and there is no protocol one can follow to compute the synapse size for FIB/SEM data. Mayhew in [95] reviews different measures for 2D slices, including the length of the membrane apposition site or the area of the pre- and post-synaptic density. Unlike mitochondria or vesicles, synapses don't have a membrane and it is therefore hard to determine exactly which voxels belong to a synapse and which to the intracellular space. Consequently, the performance of the segmentation part of the algorithm was assessed qualitatively and found to be of sufficiently high quality for detailed analysis of synapse morphology, see Fig. 5.9 and Fig. 5.10.

5.5 Discussion

The results show that with an adequate selection of appearance features, synapses are sufficiently different from other structures in neural tissue to allow for reliable automated

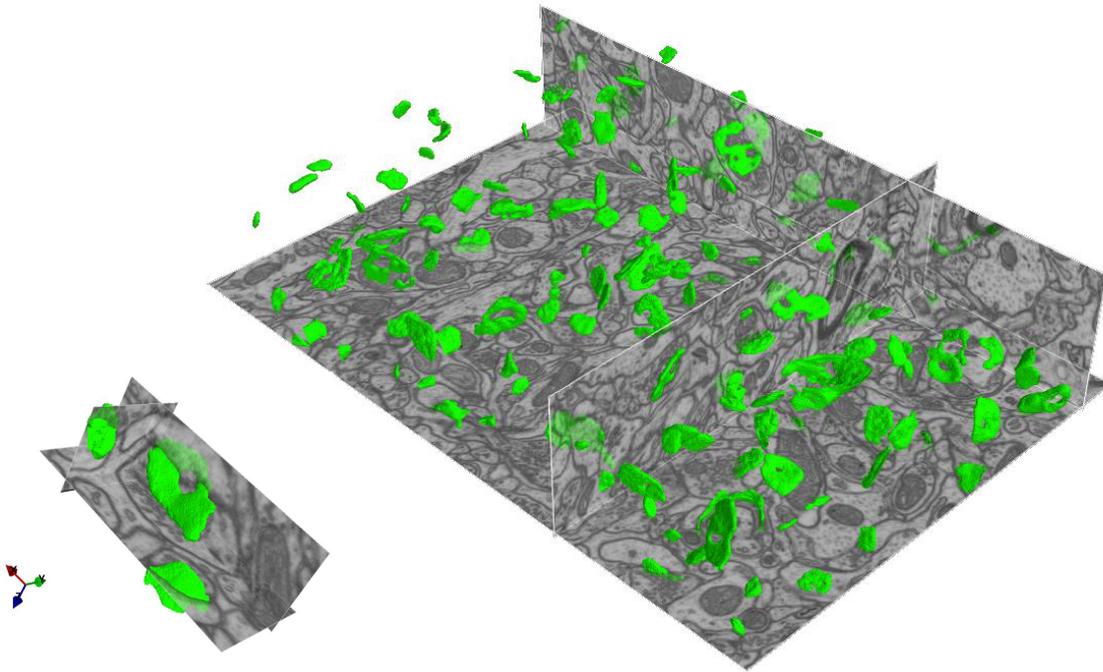


Figure 5.9: 3D visualization of the results. Top: all synapses detected by the algorithm after training on the labels from Fig. 5.2. Bottom: a close-up view of three differently oriented synapses.

detection in nearly isotropic FIB/SEM serial images. Fig. 5.11 illustrates typical false negative and false positive detections of the humans and of the algorithm, which have different causes. The false positives of the algorithm are mostly caused by myelinated membranes or very dark lines located near mitochondria (Fig. 5.11J, 5.11K, 5.11L). Similarly, most of the false negative detections also stem from synapses located very close to myelinated membranes. In the probability maps, they become connected to the large false positives caused by these membranes, and these large connected components are then filtered out based on the size criterion (Fig. 5.11G). Since ilastik provides a convenient summary report of all detected synapses (Fig. 5.10) and reduces the data from millions of voxels to just dozens of synapse candidates, the false positives for the entire stack can easily be discarded by a human in just a few minutes of additional proofreading.

5 Automated Detection and Segmentation of Synapses in Serial Electron Microscopy Images

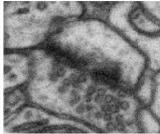
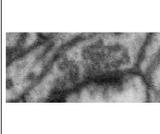
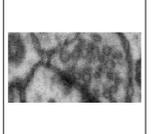
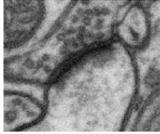
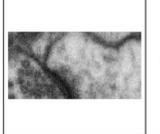
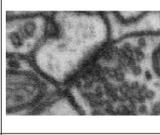
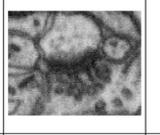
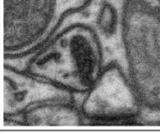
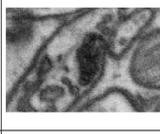
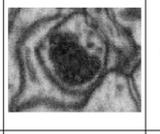
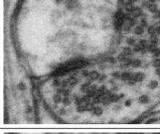
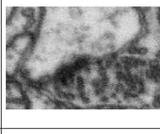
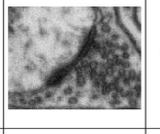
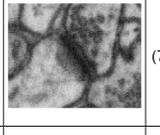
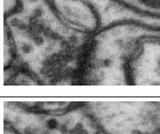
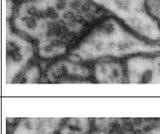
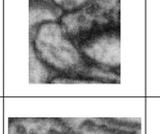
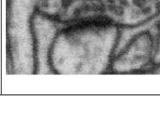
#	3d projection	XY projection	XZ projection	YZ projection	(X, Y, Z)	size in pixels
33					(666, 909, 178)	38031
34					(1087, 875, 182)	37492
35					(457, 1509, 135)	37414
36					(1070, 917, 88)	36960
37					(167, 1690, 62)	35794
38					(707, 867, 91)	34688
39					(556, 1917, 44)	33214
40					(213, 658, 59)	30086

Figure 5.10: Synapse detection summary report. Part of the summary report produced by ilastik. The fourth detection from the top (no. 36) is a false positive, which can easily be filtered out by a human expert by looking at a larger context.

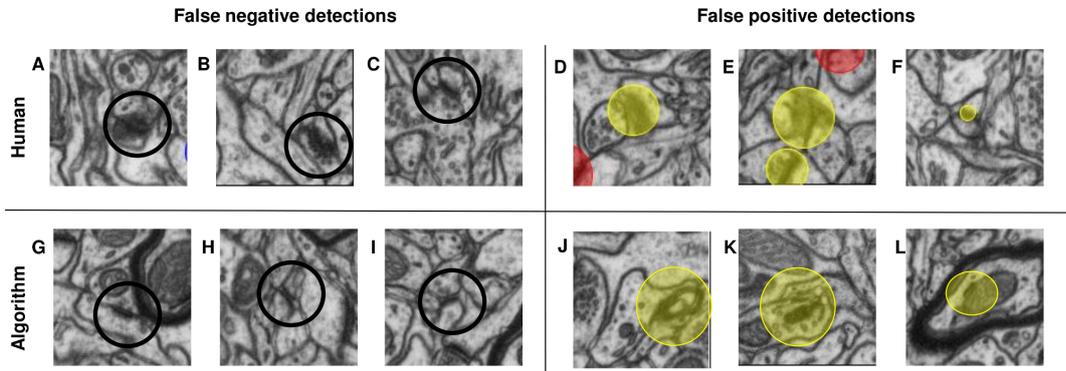


Figure 5.11: Examples of errors, committed by the algorithm and the human experts. **A, B, C**: false negative decisions of the human observers, **D, E, F**: false positive detections of the human observers, shown as yellow "ball" labels in the image center, **G, H, I**: false negative decisions of the algorithm, **J, K, L** false positive decisions of the algorithm.

For the human experts, while some synapses that were missed are accidental omissions, others serve as a good illustration of the advantages of truly 3D processing (Fig. 5.11A, 5.11B). These synapses are oriented at a low angle to the plane of imaging and do not strictly qualify as synapses according to the morphological criteria, since the synaptic cleft is not seen in the plane of imaging. Besides that, they are just hard to discern when viewing the data in native (x - y) projection only. Since the algorithm bases its decisions on geometric features computed in full 3D neighborhoods, it is not affected by synapse orientation.

As for any machine learning-based algorithm, the performance of ilastik depends significantly on how well the training dataset represents the true variability of the test data. Note also, that the images with the training labels must be large enough to allow for computation of all features from neighborhoods of the labeled voxels. The interactive learning interface of ilastik allows the user to immediately assess the algorithm performance on a subset of data and, if necessary, to modify the training labels or the threshold value. As shown in Fig. 5.8A, on our data the quality of the prediction was stable with respect to the exact choice of the training set.

We are currently working on new machine learning methods which take more spatial context into account with the aim of solving the myelinated membranes problem and achieving reliable synapse segmentation also in image stacks with low z -resolution. Some of our conclusions so far are listed in the next section.

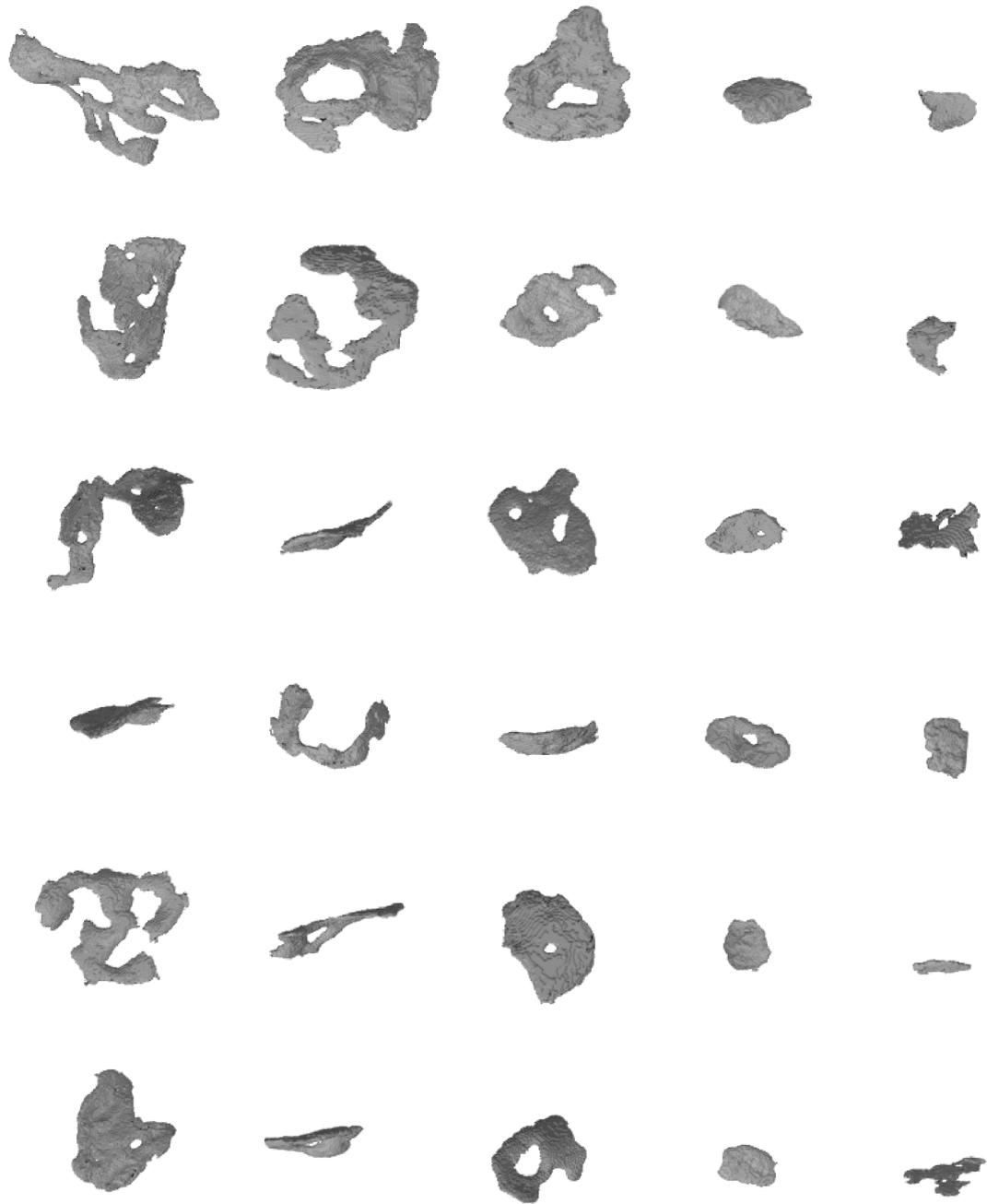


Figure 5.12: Variability of the 3D shape of synapses. These synapses were not extracted from the test dataset, but from a different dataset of a mouse brain, also produced by Dr. Knott's lab.

5.6 Outlook: synapse detection in non-isotropic images

5.6.1 Introduction

Synapse detection in ssTEM data is a much more challenging problem than synapse detection for FIB/SEM. We evaluated several strategies to tackle this problem and describe them in this subsection. All the data shown here was produced by Davi Bock et al [15] in the laboratory of Clay Reid in Harvard Medical School and was made freely available after the publication. We only use a small subset of this data for testing.

The resolution of ssTEM instruments is extremely anisotropic, with resolution differences in x , y and z reaching the factor of 20 (up to $3 \times 3 \times 60$ nm). Transmission Electron Microscopy works by transmitting electrons through ultra-thin slices of tissue. The intensity of a given pixel then represents the combined signal from all the cellular structures in the electron path. Consequently, the structures which pass at an angle to the slicing plane appear smeared and without additional context information it's impossible to tell if a particular region of the image represents a synapse, a smeared membrane, or a smeared mitochondrion border.

Another problem arises from the fact, that the ultra-thin slices have to be handled after they are cut. This introduces tears, folds and other defects, while imaging the slices separately introduces artefacts and differences in intensity and contrast. To further illustrate this point, Fig. 5.13 shows how the appearance of a synapse can change in several consecutive slices of a block of tissue.

To the best of our knowledge, Mishchenko et al [106] presented the first and only proposal for automated detection of synapses in ssTEM data. The detection is based on the observation that the membrane apposition site is thicker where a synapse occurs and is performed by training a neural network classifier on features like membrane width and intensity. The membranes in the volume have to be found in advance, either manually or by a semi-automatic approach.

5.6.2 Feature dimensionality

When looking for synapses or membranes in ssTEM data, human experts make their decisions based on the appearance of the candidate object in several consecutive slices. Automated membrane detection algorithms first detect the cell boundaries in 2D slices and then link them in 3D. Depending on the method used, the linking step can also remove the excessive or wrong 2D boundaries. The approach we successfully applied to FIB/SEM data included calculation of isotropic 3D features and thus using the 3D context directly. Consequently, as a first step to the detection of synapses in ssTEM images, we decided to check if one can ignore the anisotropy of the images and apply 3D

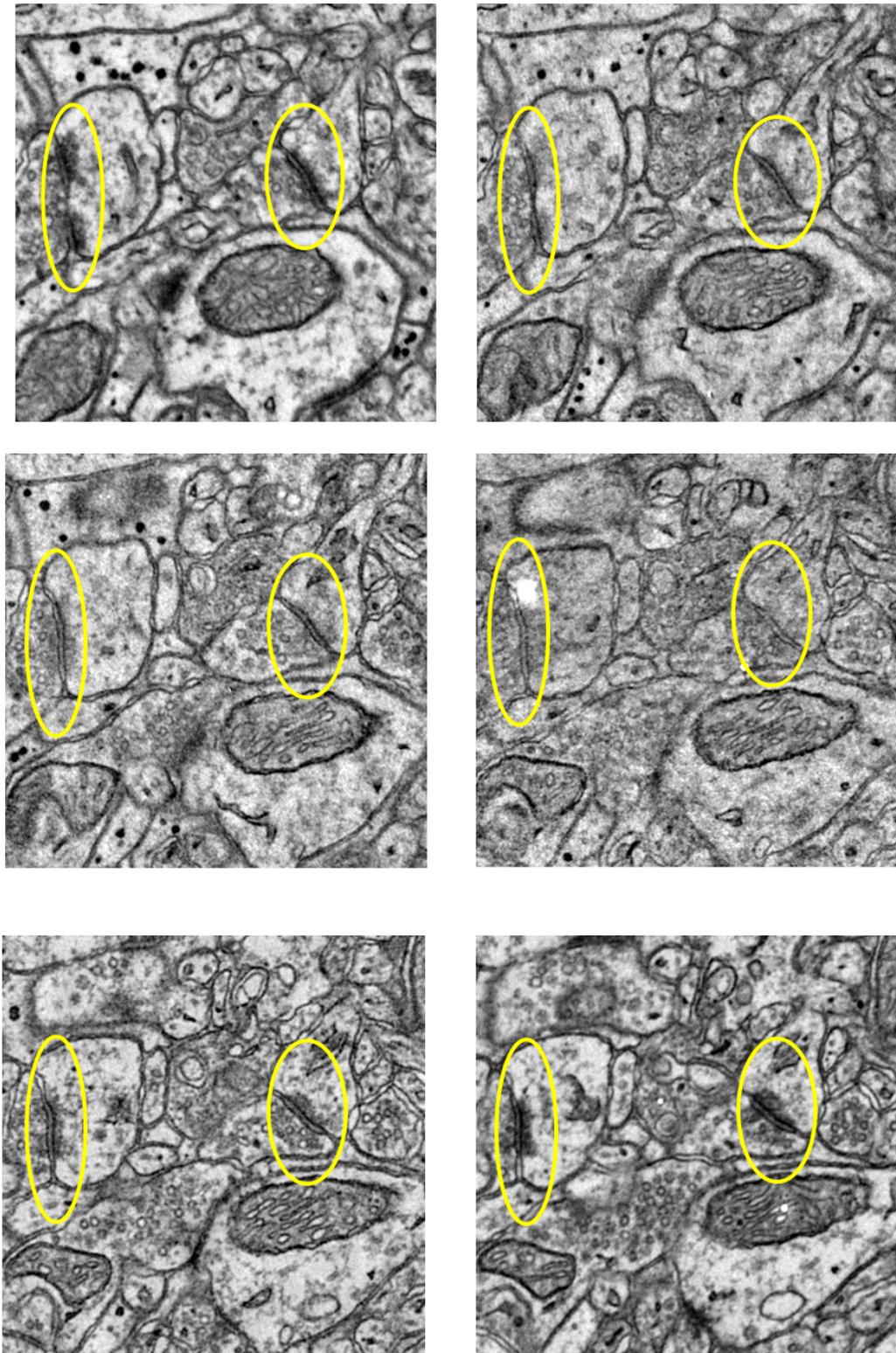


Figure 5.13: Two bright synapses (highlighted by yellow ovals) as imaged in six consecutive ssTEM slices. The order of slices is left to right, top to bottom.

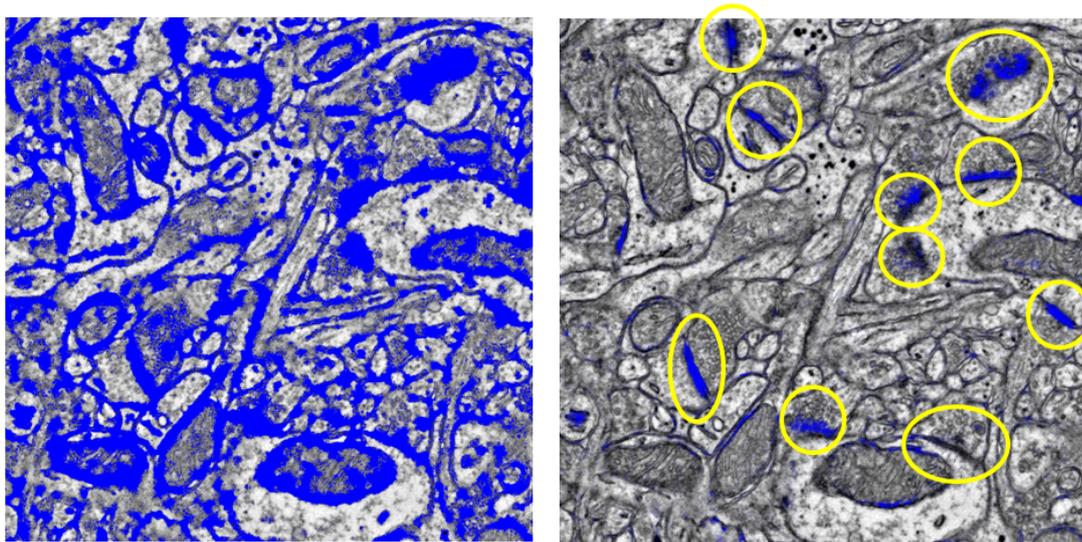


Figure 5.14: Prediction on features from Table 5.1, computed in 2D (left) or 3D(right). True synapses are indicated by yellow ovals.

filters instead of 2D filters. A comparison to pure 2D features can be seen in Fig. 5.14. It is obvious that 3D context is very important for the algorithm specificity.

5.6.3 Context and auto-context

However, even when 3D features are used, the algorithm makes a lot of false positive detections, especially on the mitochondrial membranes and endoplasmic reticulum. If one observes how the human experts perform synapse detection, it becomes apparent that humans take more context into account and re-evaluate their pixel-wise decisions based on the surroundings of a potential synapse. For example, synapses should have a cluster of vesicles nearby in at least on of the slices and can't be located on the sides of mitochondria. Incorporation of the context information has long been a subject of active research in the field of computer vision, especially in the domains of object detection and categorization and scene understanding. A common way of including additional information about the pixel neighborhood is by re-formulating the problem as a Markov Random Field or a Conditional Random Field [97]. In this reformulation, the problem of labeling each pixel in an image is presented as an energy minimization problem, where the energy is factorized into potential functions, each of which only depends on the pixel itself or on its interactions with other pixels in a neighborhood of fixed size. For example, if a pixel-wise classification is performed first, further improvement of the segmentation

can be achieved by building an energy function with unary potentials representing the classifier response in the first step and binary potentials encouraging the neighboring pixels to belong to the same class. The resulting segmentation will be more smooth than the original probability map. Binary and tertiary potentials can also include terms which give preference to known properties of the classes present in the image. For example, Kaynig et al [74] tackled the problem of membrane segmentation in ssTEM images by performing a Random Forest based prediction first and then minimizing an energy function, which included additional terms to fill the gaps in the detected membranes by encouraging membrane pixels to belong to thin elongated structures. Over the last years, the power of such modeling approaches has been demonstrated on many challenging computer vision applications [146]. Their disadvantages include the limits they enforce on the range and type of interactions that can be modelled, caused by computational costs of energy minimization and the fact that the exact form of the interactions has to be modelled from prior knowledge.

An alternative approach to include context information for the improvement of classifier results has recently been proposed by Tu and Bai [150]. Instead of modeling the interactions between pixels or pixel classes explicitly, they suggest to use classifier predictions as features for the next round of classification. The workflow of the algorithm can be summarized as follows:

1. The input data of the algorithm consists of fully labeled training images I .
2. A classifier is trained on I and some pixel-wise image features F_i , producing probability maps P_0 .
3. For each pixel of I new features F_c are added by sampling pre-defined points of P_0 . In principle, complete P_0 can be used, but training would become very slow.
4. A new classifier is trained on I and $F_i \cup F_c$.
5. Previous two steps are repeated until convergence.
6. The output of the algorithm is a chain of classifiers

The authors reason that since the original image features are still used in each round of classification, the predictions of the classifier can only improve when more context features are added. The underlying assumption of the method is that using predictions of the previous step the classifier will learn the relations and shape models of the classes implicitly. The method is not limited to any particular choice of classifier, as long as it is capable of handling a potentially very large feature space. The authors use Probabilistic Boosting Trees [149] and SVMs, but comment that Random Forests and other ensemble based methods would also be a good choice. A star pattern is used to sample

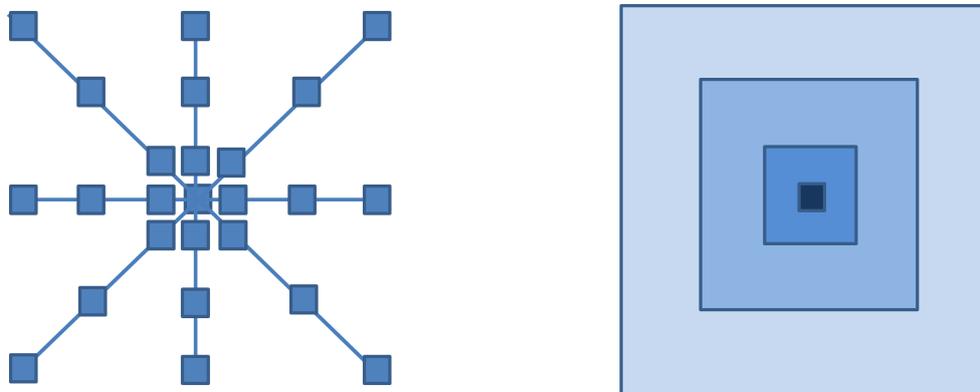


Figure 5.15: Left: structure of the neighborhood, used in [150] to sample probability maps and in [70] to sample both probability maps and raw image intensities. In case of [150], the average in a 3×3 window around each pixel of the star is also used. Right: structure of the neighborhood, used in our approach. We compute summary statistics on the pixels in square neighborhood, excluding the points which lie in interior squares.

the probability maps at each step (see also Fig. 5.15:Left), with rays of length from 5 to 200 and 21 points sampled on each ray in total. Haar features are used as image features [155]. The authors have demonstrated excellent performance of the algorithm on the following datasets: the Weizmann dataset of 328 gray scale images of horses [16], OCR dataset of hand-written words [71], the MSRC dataset for scene parsing/region labeling [140], human body parts configuration dataset [107], a dataset of MRI images for caudate segmentation and segmentation of whole brain 3D MRI images [50].

Jurrus et al [70] applied the auto-context approach to membrane segmentation in ssTEM images. They used an Artificial Neural Network (ANN) as a classifier and raw image intensities as image features. Jurrus et al use a stencil (Fig. 5.15:Left) both for sampling the raw image intensities as image features and for sampling the probability maps at later classification stages. Their results show, that introduction of context leads to a definite improvement in membrane prediction accuracy. Very recently, the same group of authors [137] have proposed an improvement of the approach [70], based on Radon features and context computed at different scales.

A major difference can be observed between our ssTEM data and the data used in

the works mentioned above. The images of all the datasets of [150] have a certain degree of rotational invariance: for body parts head is always above legs, the same is true for the Weizmann dataset of horses and especially for MRI images of the brain. ssTEM data of [70] comes from a rabbit retina and from the ventral nerve cord of a *c.elegans* worm. In both of these image stacks all the neurons follow approximately the same direction, perpendicular to the slicing plane. In contrast to the objects in these datasets, neurons and synapses in [15] do not have a preferred orientation with regard to the slicing plane or each other. Consequently, using stencil or star neighborhood to sample the probability maps in this case would lead to overfitting the classifier to the training data and would not generalize well to synapses of other orientations. Instead, we considered square neighborhoods of each pixel and computed the mean and the variance of the probability maps in the squares (see also Fig.5.15:Right), thus making the context rotation invariant in the XY plane. We further considered histograms of probability maps in the same neighborhoods. Many other statistical summary features can also be tried. Besides rotational invariance, this approach uses less features and thus simplifies the feature space and requires less labels.

5.6.4 Anisotropic features

All the features we use in Table 5.1 are based on images smoothed by a convolution with a Gaussian. Vigna image processing library has recently introduced an option to compute such features with an anisotropic kernel [82]. Such anisotropic features could be tuned to use the right amount of 3D context for synapses also for cases when the $X \times Y$ context is fairly large. Anisotropic features of this kind can be thought of as interpolating the data in the z-direction. If the objects that the learning algorithm is trying to detect are of isotropic shape (like spheres, which would look like ellipsoids in data with poor 3D resolution) or of the same orientation, introducing anisotropic features might not help the classification too much, as the classifier can learn to detect ellipsoids as easily as to detect spheres. However, synapses are much smaller in the direction perpendicular to the synaptic cleft and are thus more like disks than spheres. Also, synapses of different orientation to the slicing plane would look very different from each other in anisotropic data volumes. Interpolating the data in z-direction could then help detect those synapses even if all the labels the user gives are for synapses perpendicular to the slicing plane.

5.6.5 Experiments

Out of all the images, provided by [15], we selected two small non-overlapping sub-stacks, one of $1024 \times 1024 \times 23$ voxels for training, the other of $1024 \times 1024 \times 30$ voxels for testing. The lateral(XY) resolution of the images was below 5nm, while section thickness was below 50nm. Labeling was performed in ilastik, with new labels added interactively based on the predictions of the Random Forest on isotropic 3D features. Five classes

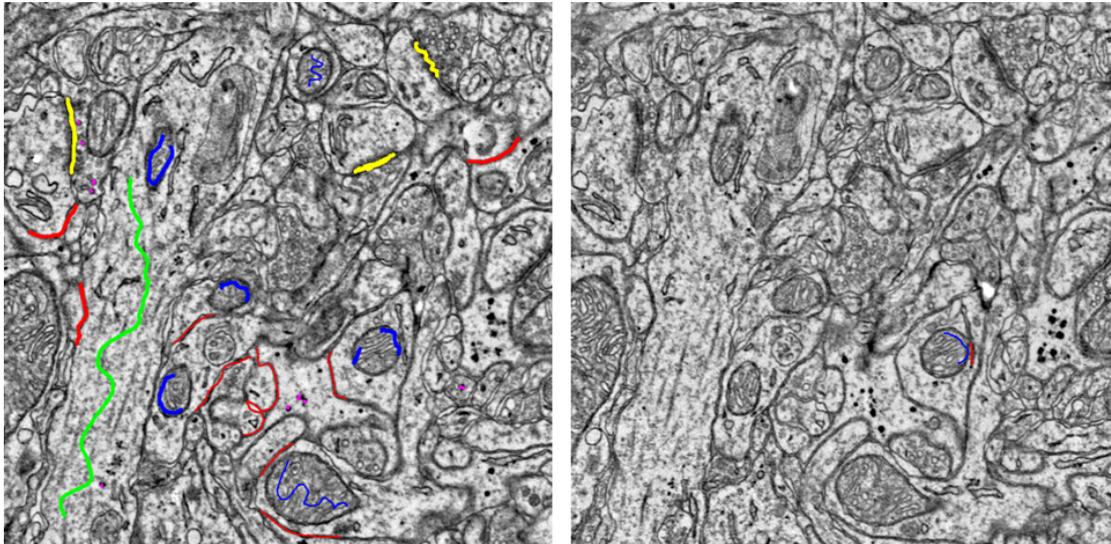


Figure 5.16: Training labels for ssTEM data. Membranes are labeled red, synapses - yellow, mitochondria - blue, vesicles - magenta and the rest green.

were used: membranes, synapses, mitochondria, vesicles and other, however, the labels were added to optimize the prediction results for the synapse class only. The training labels are shown in Fig. 5.16. For the features we used, as before, 1) intensity, smoothed by a Gaussian, 2) eigenvalues of the Hessian matrix, 3) eigenvalues of the structure tensor, 4) Laplacian of a Gaussian, 5) gradient magnitude of a Gaussian, 6) difference of Gaussians. Isotropic features were computed with $\sigma_x = \sigma_y = \sigma_z = 1, 1.6$ and 3.5 . Anisotropic features were computed with $(\sigma_x, \sigma_y, \sigma_z) = (1., 1., 1.), (1.6, 1.6, 1.6), (3.5, 3.5, 1.)$ and $(5, 5, 1.6)$. Additionally, we tested also adding $(\sigma_x, \sigma_y, \sigma_z) = (10, 10, 1.6)$ to anisotropic features, but these features did not bring any improvement to the prediction (based on visual inspection of the probability maps) and were omitted from further experiments.

The autocontext features were computed in squares of size 5, 10, 15, 20, 30 and 40. As summary statistics we tested mean, variance and 5-bin histogram of the class probability of pixels in the square, excluding the pixels which belong to the smaller square. No significant performance difference was found between using the mean and variance or the histogram, consequently, the Results section only shows probability maps for the predictions based on mean and variance features. We also tried to incorporate more 3D context information by augmenting the feature set of each pixel by the auto-context features of the pixel above and below it. Surprisingly, these features did not improve the prediction. Four iterations of auto-context were used.

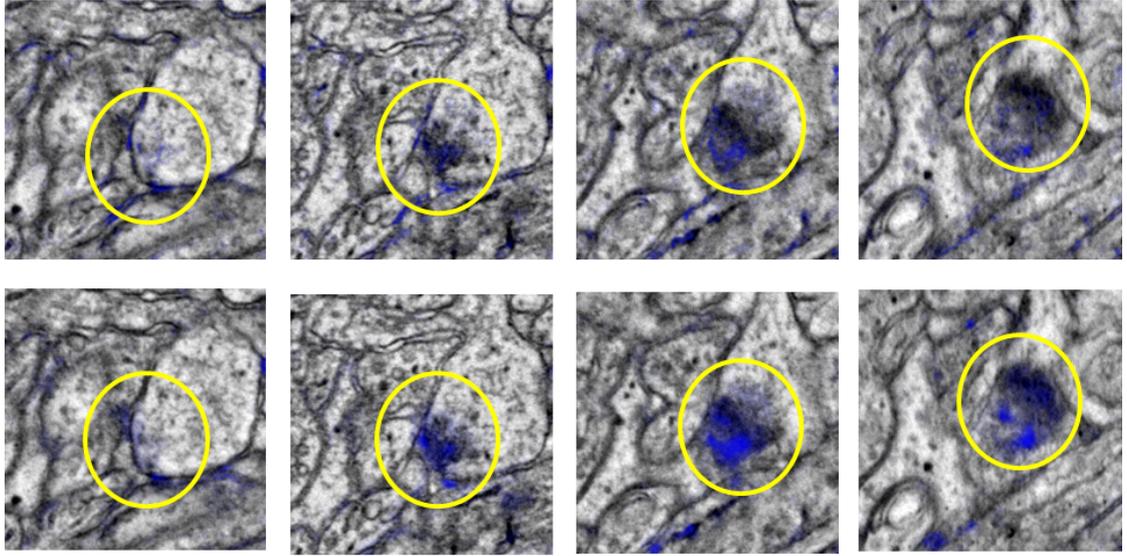


Figure 5.17: Predictions for a synapse, almost parallel to the slicing plane. Top: Random Forest on isotropic features. Bottom: Random Forest on anisotropic features. The synapse is highlighted by the yellow oval.

5.6.6 Preliminary results and Discussion

Isotropic vs. anisotropic features

Fig. 5.17 shows predictions for a synapse almost parallel to the slicing plane. This synapse is very difficult to detect and isotropic features give almost no indication of its presence (Fig. 5.17:Top). On the other hand, prediction on anisotropic features finds some evidence for this synapse, although it is not nearly as certain as for the synapses perpendicular to the slicing plane (Fig. 5.17:Bottom).

Fig. 5.18 shows another example of such a synapse (yellow), side by side with a synapse, almost perpendicular to the slicing plane (red). This example is not as difficult as the one in Fig.5.17: the synaptic cleft is also not visible, but the pre- and post-synaptic densities are visible across more slices. However, Random Forest on isotropic features does not detect it (Fig. 5.18:Top), while anisotropic features give a clear indication of its presence (Fig. 5.18:Bottom). The other synapse, almost perpendicular to the slicing plane, is detected by both methods, however, the prediction is also more smooth on the anisotropic features.

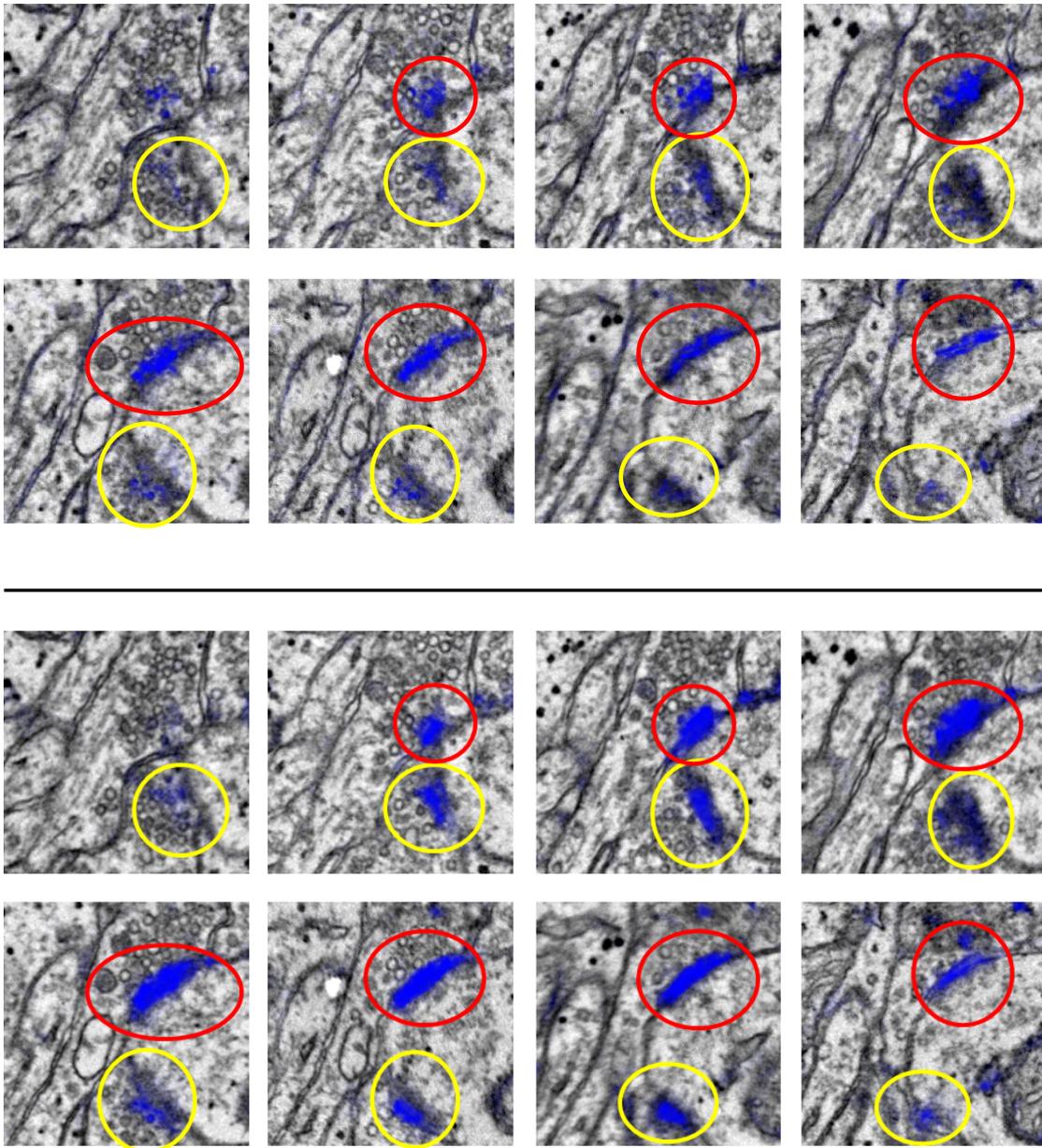


Figure 5.18: Predictions for two synapses, one at a low angle to the slicing plane and difficult to detect (yellow), the other almost perpendicular to the slicing plane and easy to detect (red). Top: Probability map produced by Random Forest on isotropic features. Bottom: on anisotropic features.

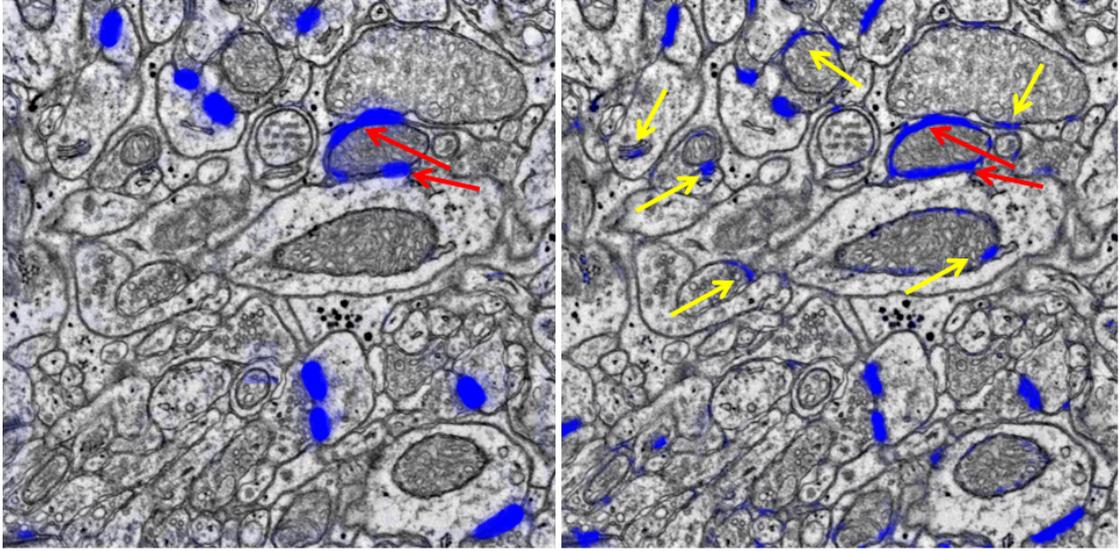


Figure 5.19: Comparison of prediction results with(Left) and without(Right) auto-context features. Yellow arrows on the right part of the Figure point to some groups of pixels falsely labeled as synapses, when no auto-context features were used. Red arrows point to false positive synapse predictions, which were not corrected by using auto-context features.

Auto-context

Fig. 5.19 shows the effect of auto-context features on the Random Forest probability maps. Besides larger groups of falsely labeled pixels, indicated by yellow arrows on Fig. 5.19:Right, many smaller false positive errors were also corrected. However, the largest false positive detection on the border of a mitochondrion indicated by red arrows, is present in both cases.

A typical false negative detection of the Random Forest with auto-context features is shown on Fig. 5.20. The small synapse indicated by the yellow oval was found in the first iteration, when no auto-context features were used, however, it got lost after these features were added. The synapse indicated by the red oval is actually found in the later slices.

False positive detections of the algorithm are usually caused by mitochondrial membranes (Fig. 5.19, red arrows) or endoplasmic reticulum (Fig. 5.21).

Slices of different appearance prove to be very challenging for the classifier. Even without auto-context features, it fails to detect synapses in the slices which look too different from the ones used for training (Fig. 5.22).

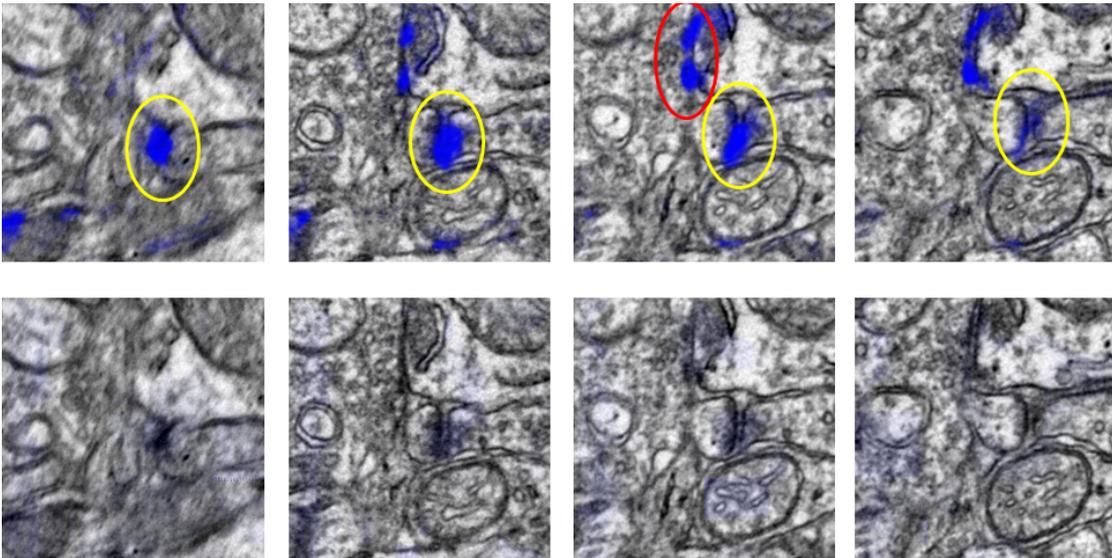


Figure 5.20: A typical false negative detection of the Random Forest with auto-context features. Left: predictions with auto-context features. Right: predictions without auto-context features. Small synapse in the yellow oval was correctly detected by using image features only (Right) and missed after adding auto-context features. The synapse in the red oval will be detected in the next slice (not shown).

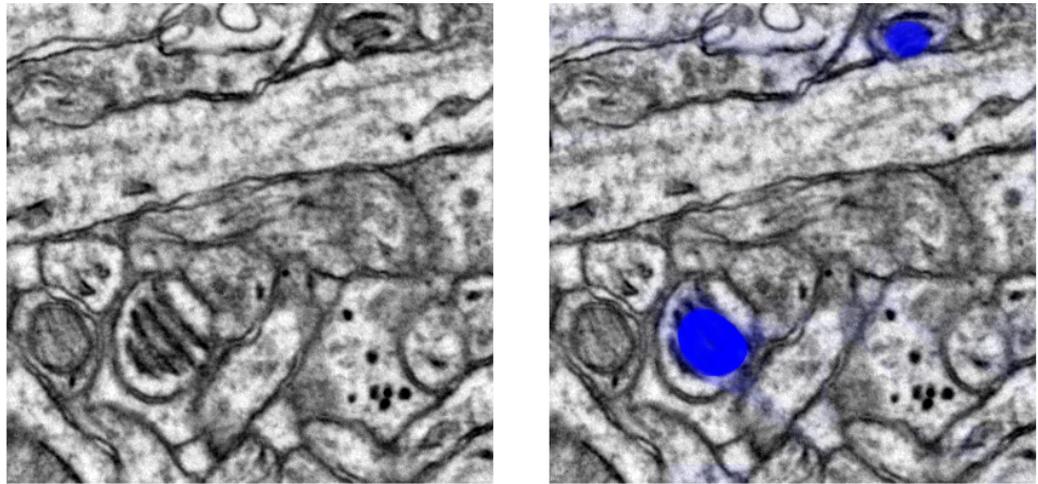


Figure 5.21: A typical false positive detection of the Random Forest. Left: raw data. Right: synapse probability map, produced by Random Forest with auto-context features. These two instances of, probably, endoplasmic reticulum, were also falsely detected using only image features.

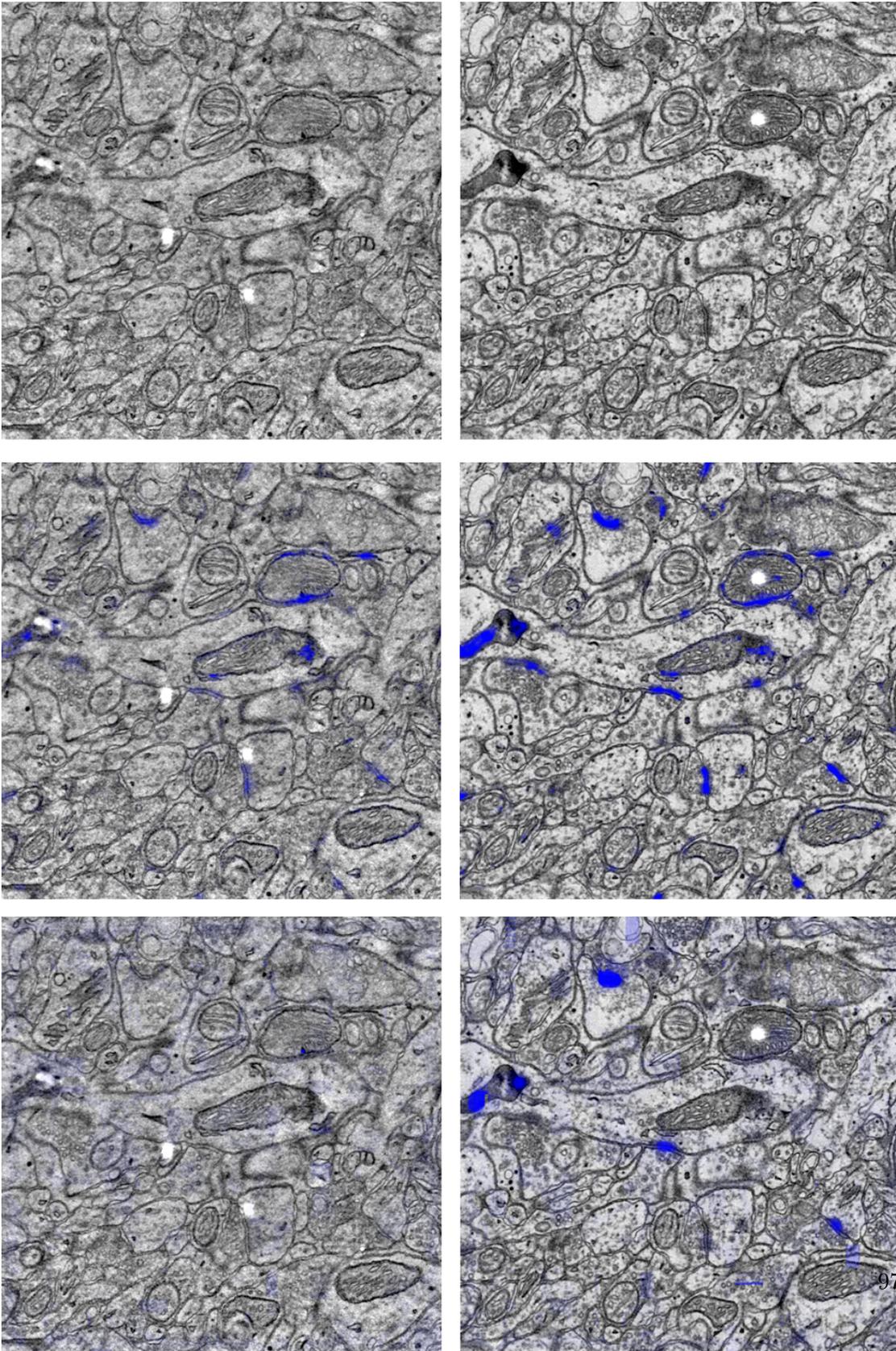


Figure 5.22: An example of classifier prediction on two slices of different appearance. Top: two slices of raw data. Middle: probability maps, produced by the Random Forest, using anisotropic image features only. Bottom: also using auto-context features.

Chapter 6

Final Discussion and Outlook

6.1 Quantification of Bimodal Isotope Peak Distributions in H/D Exchange Mass Spectrometry

In Chapter 3 we introduced a method for quantification of bimodal isotope peak distributions, based on the HeXicon software. It was validated by a pulse-labeling experiment, which studied conformational changes induced by ATP incubation. However, the suitability of HeXicon is not limited to pulse-labeling experiments, for which bimodal isotope distributions indicate the coexistence of alternative conformations and allows the determination of the interconversion rate between the alternative conformations. It could also be used for continuous-labeling experiments to distinguish EX1 and EX2 exchange mechanisms. The results for EX1 regimes would look reverse as compared to our results, since the bimodal distribution of the isotope peaks in EX1 exchange is characterized by the appearance of high exchanged species, while in our pulse-labeling experiments low exchanged species are observed after increasing pre-incubation time in the presence of ATP. More difficult would be the analysis of a mixed EX1-EX2 exchange behavior. In such an exchange regime the distribution of the isotope peaks would be bimodal with changing ratios between low and high exchanged species and concomitant shift of one or both maxima of the bimodal distribution to higher m/z values. While HeXicon would still estimate the deuteration distribution correctly, the subsequent analysis of kinetics will be more challenging.

In the presented approach the bimodal distributions are found based on simple pattern matching. Should the method be extended to other labeling strategies as described above, this part of the algorithm can be replaced by a learning-based approach.

In summary it can be stated that HeXicon with the described modifications is very

suitable for detecting bimodal isotope distributions. False negatives and false positives generally originate from co-eluting peptides with similar m/z causing merging isotope distributions after deuteration. A poor signal to noise ratio may also contribute to erroneous estimation of bimodal deuteration distributions. While the first problem will be reduced in instruments with higher resolution, the second requires improved sensitivity or modified sample handling. The ability of HeXicon to automatically extract peptides with a bimodal isotope distribution may lead to a reversal of the workflow. Instead of first establishing the sequence coverage by MS/MS experiments, one could first perform an HDX experiment, screen the data with HeXicon searching for specific features, and subsequently identify only the interesting peptides by a targeted inclusion list MS/MS experiment. Such an approach would be particularly advantageous for very large proteins or a complex of several proteins.

6.2 Automated Quantification of $^{16}\text{O}/^{18}\text{O}$ -labeled LC/MS Data

In Chapter 4 we presented an automated quantification algorithm for $^{16}\text{O}/^{18}\text{O}$ labeling experiments, which works both in high and low resolution and is not limited to identified peptides. Sensitivity of the algorithm was tested on an experimental biological dataset and its consistency and precision were demonstrated on calibration data and predefined mixtures. The algorithm includes two spectrum segmentation approaches, one for low resolution data, which is based on image processing methods, and one for high resolution data, based on efficient sparse data representation. In both cases peak picking in the spectrum segments is performed by regularized regression. As the isotope distribution models can be provided either from peptide identification results or from approximate average models, the method is not limited to identified peptides and can also be used for exploratory data analysis. In fact, neither the segmentation algorithm nor the peak picking exploit any properties specific to $^{16}\text{O}/^{18}\text{O}$ labeling experiments and the mass difference between the labels is only used at the very last step to form the labeled/unlabeled peptide peak groups. The core of the algorithm is thus not limited to $^{16}\text{O}/^{18}\text{O}$ labeling and can be adapted for other stable isotope labeling experiments with minimal changes.

A possible future development direction would be to include an additional post-processing step for non-identified peptides. Currently, the algorithm reports quantification results for all peak groups with correct inter-peak distances. Some of these peak groups may include noise. Training a classifier on peak groups, identified by Mascot as labeled by ^{18}O , could significantly improve the specificity of the method without requiring further user interaction for providing the training labels.

6.3 Automated Detection and Segmentation of Synapses in Serial Electron Microscopy Images

In Chapter 5 we introduced a learning-based method for detection and segmentation of synaptic contacts in electron microscope images of neural tissue. For nearly isotropic images acquired on a FIB/SEM microscope, the performance of our method is comparable to that of manual analysis done by experienced neuroscientists. For images with poor z-resolution we modified our approach by adding auto-context features and achieved promising results.

Besides our approach of extracting the segmentation from probability maps by a user-selected threshold, one can also consider an alternative: the segmentations produced by different threshold values can be combined into a component tree. Currently, we cut the tree at two levels: first at the user defined threshold for filtering and then at the 0.5 threshold for segmentation. It could also be possible to find a single adaptive threshold by another classification step, based on object-level features. This additional classification step would also help correct the remaining detection errors. However, it would not be practical to completely replace our interactive thresholding by this approach. Currently, we only require the user to label 2-3 synapses, but to perform object-level learning the algorithm would need many more object-level annotations. While individual voxels of synapses are quite similar in terms of their close neighborhood features, there is a great variability between synapses as 3D objects (see also Fig. 5.12) and a lot of labels will be required to capture it.

On the other hand, while the voxel-wise features depend on the image properties of the acquired stack, the object-level features reflect the biological properties of synapses and should be transferable between datasets. The following workflow could be imagined: after the current sequence of steps is performed and the user proofreads the algorithm detections on a test dataset, the results of the proofreading are taken as positive and negative examples for a second round of training, this time on the object level. This second classifier can then be applied to other parts of a larger stack or even to other stacks of the same resolution and improve its voxel-wise classification results without any additional cost in user effort.

The synapse detection pipeline of Chapter 5 has so far only been proven to work for asymmetric synapses. The main difficulty of tuning it for the detection of symmetric synapses comes from the fact that there are not enough symmetric synapses in the dataset we used for testing to make any quantitative conclusions on the quality of the detection. A test dataset for symmetric synapses would either have to be much larger, which is not possible with the current field of view limitations of FIB/SEM mi-

croscopy, or be acquired in a different part of the brain, more rich in inhibitory synapses.

Our experiments on ssTEM data show, that automated synapse detection in data with low z-resolution is possible. A combination of anisotropic image features with auto-context features provides good results on easier synapses. Anisotropic features alone provide very good sensitivity, also for the difficult synapses, oriented at a low angle to the slicing plane. The effect of using auto-context features is similar to regularization, efficiently removing most of the false positive detections. However, it also removes the less certain predictions of the difficult synapses. We have not yet fully optimized the neighborhood sizes of the features, which might help balance sensitivity of anisotropic features and specificity of the auto-context. Besides, summary statistics other than mean, variance and histogram of the pixel neighborhoods could also make auto-context-based prediction more sensitive. Interesting features to explore include smoother density estimators instead of histograms and histograms with number of bins increasing with the size of the neighborhood. Rotation invariant features based on spherical harmonics could also prove useful.

The auto-context features used in Chapter 5 were always computed in 2D neighborhoods of pixels. We tried to incorporate 3D context by including auto-context features of the pixels above and below the current one. These additional features did not improve classification results, which is surprising, considering how superior 3D image features are compared to 2D features. The reason might be that the total number of features becomes too large and more training labels are needed to correctly define the decision surface in very high dimensional feature space. Two more options for 3D context exist: computing the context features directly in 3D or performing prediction based on the auto-context in the current slice and in the slices above and below in alternation. The latter case is similar to the common strategy for membrane detection in ssTEM: detection in 2D followed by linking in 3D.

The training annotations in Chapter 5 were based on the interactive feedback of the classifier, aiming to optimize the prediction of the synapse class. Both [150] and [70] use fully labeled images for training. While we would prefer to avoid requesting such extensive annotation from the user, additional labels, improving also the prediction of other classes, would perhaps be beneficial for the auto-context learning the correct inter-class relationships.

The effect of variability of contrast and intensity between different slices of the stack could be reduced by selecting a training stack with similar variability and annotating it on a range of different slices. We also did not yet explore image filters which could help make the slices more similar.

6.3 Automated Detection and Segmentation of Synapses in Serial Electron Microscopy Images

While the field of view of FIB/SEM microscopes is constantly improving, it is not nearly as large as what can be achieved with ssTEM imaging. ssTEM microscopy remains a very popular technique in neural circuit reconstruction and could greatly benefit from automation of synapse detection, which we hope to achieve by exploring the strategies listed above in a future study.

We expect the proposed tool to be useful not only for synapse counting, synapse density estimation or estimation of synapse-to-neuron ratio, but also for the ongoing efforts in the reconstruction of neural circuits [21, 58, 28, 68, 106, 59]. Our approach of detecting synapses without a prior volume segmentation into cells would be especially attractive in cases when circuit reconstruction is done by only following the "skeletons" - the center line of the cells, as presented in [59].

Frequently used Abbreviations

BIC	Bayesian Information Criterion
Da	Dalton
EM	Electron Microscopy
ESI	Electrospray Ionization
FIB/SEM	Focused Ion Beam/Scanning Electron Microscopy
FN	False Negatives
FP	False Positives
HDX or HX	Hydrogen Deuterium Exchange
LC/MS	Liquid Chromatography Mass Spectrometry
MALDI	Matrix Assisted Laser Desorption/Ionization
MS	Mass Spectrometry
MS ²	synonym for tandem MS or MS/MS
SBFSEM	Serial Block Face Scanning Electron Microscopy
TEM	Transmission Electron Microscopy
TIC	Total Ion Chromatogram
TOF	Time of Flight Mass Spectrometer
XIC	Extracted Ion Chromatogram

Table 6.1: Frequently used abbreviations (1).

List of Tables

4.1	Mass of atoms and natural abundances of the stable isotopes of hydrogen, carbon, nitrogen, oxygen and sulfur	40
4.2	Labeled/unlabeled peak groups found by the algorithm and by the human expert	57
4.3	Labeled/unlabeled peptide ratios found by our tool and Mascot multiplex, compared to the true value	58
5.1	Local neighborhood features used for voxel classification	73
6.1	Frequently used Abbreviations (1).	105

List of Tables

List of Figures

2.1	A short peptide example	8
2.2	A schematic view of an LC/MS proteomics experiment	9
2.3	A 2D mass spectrum and the corresponding TIC and XICs	10
2.4	A schematic view of a TOF mass spectrometer	11
2.5	An example of an isotope envelope	12
2.6	ssTEM pipeline and example	15
2.7	FIB/SEM scheme and example	17
3.1	Primary and secondary protein structure	20
3.2	Comparison of unimodal and bimodal deuteration distributions	23
3.3	Continuous and pulse labeling experimental schemes	25
3.4	Improvement of the early stopping criterion	28
3.5	Alignment by parametric time warping	30
3.6	Two examples of deuteration distribution with Gaussian peaks	31
3.7	An example of a non-Gaussian deuteration distribution	32
3.8	Another example of a non-Gaussian deuteration distribution	33
3.9	Deuteration distribution of the same peptide at different charges	33
3.10	Sequence coverage of peptic peptides of HtpG	34
3.11	New peptide, found to be bimodal by HeXicon	36
3.12	Analysis of false positive candidates	38
4.1	Spectrum with the unlabeled and labeled signals overlap	41
4.2	Low resolution segmentation	46
4.3	An example of the Orbitrap spectrum	47
4.4	Sparse segmentation algorithm scheme	48
4.5	KD-tree	49
4.6	Centroiding along m/z dimension	51
4.7	Denoising of a high-resolution spectrum	53

4.8	Connecting isotope clusters	54
4.9	Low resolution fit	55
4.10	NITPICK fit in high resolution.	56
4.11	Comparison of Mascot multiplex and proposed approach	59
4.12	Results on high resolution data	60
4.13	Signal of a peptide with 1 to 8 labeled/unlabeled ratio	61
5.1	Synapse components and different synapse types	64
5.2	One of the training annotations used for automated synapse detection	69
5.3	Generalization properties of a single decision tree vs. a forest	70
5.4	Example of feature response	72
5.5	Comparison of classifier predictions using 2d and 3d features	74
5.6	Thresholding of probability maps.	75
5.7	A screenshot of ilastik object browsing tab.	77
5.8	Precision and recall of the algorithm and the human experts.	79
5.9	3D visualization of the synapse detection results.	81
5.10	Synapse detection summary report.	82
5.11	Examples of errors, committed by the algorithm and the human experts.	83
5.12	Variability of the 3D shape of synapses.	84
5.13	Synapses in six consecutive ssTEM images	86
5.14	2D and 3D features comparison	87
5.15	Stencil neighborhood vs patch	89
5.16	Training labels for ssTEM data	91
5.17	Predictions for a synapse, almost parallel to the slicing plane	92
5.18	Comparison of isotropic and anisotropic features	93
5.19	Comparison of prediction results with and without auto-context features	94
5.20	A typical false negative detection of the Random Forest with auto-context features	95
5.21	A typical false positive detection	96
5.22	An example of classifier prediction on two slices of different appearance.	97

List of Publications

Journals

- A. Kreshuk, C.N. Straehle, C. Sommer, U. Koethe, M. Cantoni, G. Knott, F.A. Hamprecht (2011). Automated detection and segmentation of synaptic contacts in nearly isotropic serial electron microscopy images, *PLoS ONE* 6(10): e24899. doi:10.1371/journal.pone.0024899
- A. Kreshuk, M. Stankiewicz, X. Lou, M. Kirchner, F.A. Hamprecht, M.P. Mayer (2010). Automated detection and analysis of bimodal isotope peak distributions in H/D exchange mass spectrometry usign HeXicon, *International Journal of Mass Spectrometry*, Special Issue on H/D Exchange MS, 302(1-3):125-131.

Peer-reviewed Conference Proceedings

- A. Kreshuk, C.N. Straehle, C. Sommer, U. Koethe, G. Knott, F.A. Hamprecht (2011). Automated Segmentation of Synapses in 3D EM Data, *2011 IEEE International Symposium on Biomedical Imaging: From Nano To Macro*, Chicago, USA.

Extended Abstracts

- M. Hanselmann, J. Röder, U. Köthe, B.Y. Renard, A. Kreshuk, R.M.A. Heeren, F.A. Hamprecht (2010). Active Learning for Efficient Labeling and Classification of Imaging Mass Spectrometry Data. 58th ASMS Conf. on Mass Spectrometry and Allied Topics, Salt Lake City, Utah, USA.
- A. Kreshuk, M. Kirchner, B.Y. Renard, D. Winter, B.X. Kausler, X. Lou, M. Hanselmann, J.A.J. Steen, H. Steen, W.D. Lehmann, F.A. Hamprecht (2010). Automatic Relative Quantification for High-Resolution LC/MS 16/18O Labeling

Experiments. 58th ASMS Conf. on Mass Spectrometry and Allied Topics, Salt Lake City, Utah, USA.

- Kausler BX, Kirchner M, Kreshuk A, Renard BY, Hahne H, Kuster B, Steen H, Hamprecht FA (2010) Resolution as a function of m/z for TOF, FT-ICR, and Orbitrap: predicted and confirmed. 58th ASMS Conf. on Mass Spectrometry and Allied Topics, Salt Lake City, Utah, USA.
- Kreshuk A, Kirchner M*, Renard BY*, Winter D, Steen H, Steen JAJ, Lehmann WD, Hamprecht FA. (2009). Automatic Quantification of $^{16}/^{18}\text{O}$ -Labeled LC/MS Data. 57th ASMS Conf. on Mass Spectrometry and Allied Topics, Philadelphia, Pennsylvania, USA.
- B.M. Voss, B.Y. Renard, A. Kreshuk, M. Hanselmann, U. Köthe, H. Steen, J.A.J. Steen, M. Kirchner, F.A. Hamprecht. (2009). Simultaneous Multiple Alignment for LC/MS Peak Lists. 57th ASMS Conf. on Mass Spectrometry and Allied Topics, Philadelphia, Pennsylvania, USA.

* contributed equally

Bibliography

- [1] R. Adams and L. Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):641–647, 1994.
- [2] Ayelet Akselrod-Ballin, Davi Bock, R Clay Reid, and Simon K Warfield. Accelerating image registration with the Johnson-Lindenstrauss lemma: application to imaging 3-D neural ultrastructure with electron microscopy. *IEEE Transactions on Medical Imaging*, 30(7):1427–1438, July 2011. PMID: 21402511.
- [3] Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Essential Cell Biology*. Garland Science, March 2009.
- [4] Claus Andersen, Stefano Gotta, Letizia Magnoni, Roberto Raggiaschi, Andreas Kremer, and Georg Terstappen. Robust MS quantification method for phosphopeptides using $^{18}\text{O}/^{16}\text{O}$ labeling. *BMC Bioinformatics*, 10(1):141, 2009.
- [5] James R Anderson, Bryan W Jones, Jia-Hui Yang, Marguerite V Shaw, Carl B Watt, Pavel Koshevoy, Joel Spaltenstein, Elizabeth Jurrus, Kannan UV, Ross T Whitaker, David Mastronarde, Tolga Tasdizen, and Robert E Marc. A computational framework for ultrastructural mapping of neural circuitry. *PLoS Biology*, 7(3):e1000074, March 2009. PMID: 19855814 PMCID: 2661966.
- [6] Björn Andres, Ullrich Köthe, Moritz Helmstaedter, Winfried Denk, and Fred A. Hamprecht. Segmentation of SBFSEM volume data of neural tissue by hierarchical classification. In Gerhard Rigoll, editor, *Pattern Recognition*, volume 5096 of *LNCIS*, pages 142–152. Springer, 2008.
- [7] E Appella, E A Padlan, and D F Hunt. Analysis of the structure of naturally processed peptides bound by class I and class II major histocompatibility complex molecules. *Exs*, 73:105–119, 1995. PMID: 7579970.

- [8] Marcus Bantscheff, Birgit Dumpelfeld, and Bernhard Kuster. Femtomol sensitivity post-digest 18O labeling for relative quantification of differential protein complex composition. *Rapid Communications in Mass Spectrometry*, 18(8):869–876, 2004.
- [9] Marcus Bantscheff, Markus Schirle, Gavain Sweetman, Jens Rick, and Bernhard Kuster. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry*, 389(4):1017–1031, 2007.
- [10] Mark Bates, Bo Huang, Graham T Dempsey, and Xiaowei Zhuang. Multicolor super-resolution imaging with photo-switchable fluorescent probes. *Science (New York, N.Y.)*, 317(5845):1749–1753, September 2007. PMID: 17702910.
- [11] Wolfgang Baumeister. Electron tomography: towards visualizing the molecular organization of the cytoplasm. *Current Opinion in Structural Biology*, 12(5):679–684, October 2002. PMID: 12464323.
- [12] Matthew Bellew, Marc Coram, Matthew Fitzgibbon, Mark Igra, Tim Randolph, Pei Wang, Damon May, Jimmy Eng, Ruihua Fang, ChenWei Lin, Jinzhi Chen, David Goodlett, Jeffrey Whiteaker, Amanda Paulovich, and Martin McIntosh. A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics*, 22(15):1902–1909, 2006.
- [13] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.
- [14] Tom G. Bloemberg, Jan Gerretzen, Hans J.P. Wouters, Jolein Gloerich, Maurice van Dael, Hans J.C.T. Wessels, Lambert P. van den Heuvel, Paul H.C. Eilers, Lutgarde M.C. Buydens, and Ron Wehrens. Improved parametric time warping for proteomics. *Chemometrics and Intelligent Laboratory Systems*, In Press, Corrected Proof.
- [15] Davi D. Bock, Wei-Chung Allen Lee, Aaron M. Kerlin, Mark L. Andermann, Greg Hood, Arthur W. Wetzell, Sergey Yurgenson, Edward R. Soucy, Hyon Suk Kim, and R. Clay Reid. Network anatomy and in vivo physiology of visual cortical neurons. *Nature*, 471(7337):177–182, March 2011.
- [16] Eran Borenstein. Combining top-down and bottom-up segmentation. *In Proceedings IEEE Workshop on Perceptual Organization in Computer Vision, CVPR*, 4:46, 2004.
- [17] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.
- [18] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. 10.1023/A:1010933404324.

-
- [19] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. 10.1023/A:1010933404324.
- [20] Leo Breiman and Adele Cutler. Random forests - classification description. http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.
- [21] Kevin L Briggman and Winfried Denk. Towards neural circuit reconstruction with volume electron microscopy techniques. *Current Opinion in Neurobiology*, 16(5):562–570, October 2006. PMID: 16962767.
- [22] Alan Geoffrey Brown. *Nerve cells and nervous systems: an introduction to neuroscience*. Springer, August 2001.
- [23] A Cardona. TrakEm2: an ImageJ-based program for morphological data mining and 3D modeling. In *Proceedings of the 1st ImageJ User and Developer Conference*, pages 18–19, Luxembourg, 2006.
- [24] Albert Cardona, Stephan Saalfeld, Stephan Preibisch, Benjamin Schmid, Anchi Cheng, Jim Pulokas, Pavel Tomancak, and Volker Hartenstein. An integrated micro- and macroarchitectural analysis of the drosophila brain by Computer-Assisted serial section electron microscopy. 8(10), October 2010. PMID: 20957184 PMCID: 2950124.
- [25] Brian Carrillo, Corey Yanofsky, Sylvie Laboissiere, Robert Nadon, and Robert E. Kearney. Methods for combining peptide intensities to estimate relative protein abundance. *Bioinformatics*, page btp610, November 2009.
- [26] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168, Pittsburgh, Pennsylvania, 2006. ACM.
- [27] Dmitri B Chklovskii. Synaptic connectivity and neuronal morphology: two sides of the same coin. *Neuron*, 43(5):609–617, September 2004. PMID: 15339643.
- [28] Dmitri B Chklovskii, Shiv Vitaladevuni, and Louis K Scheffer. Semi-automated reconstruction of neural circuits using electron microscopy. *Current Opinion in Neurobiology*, 20(5):667–675, October 2010.
- [29] David Clifford, Glenn Stone, Ivan Montoliu, Serge Rezzi, François-Pierre Martin, Philippe Guy, Stephen Bruce, and Sunil Kochhar. Alignment using variable penalty dynamic time warping. *Anal. Chem.*, 81(3):1000–1007, 2009.
- [30] Richard E. Coggeshall and Helena A. Lekan. Methods for determining numbers of cells and synapses: A case for more uniform standards of review. *The Journal of Comparative Neurology*, 364(1):6–15, 1996.

- [31] Jurgen Cox and Matthias Mann. Is proteomics the new genomics? *Cell*, 130(3):395–398, 2007.
- [32] Jurgen Cox and Matthias Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12):1367–1372, December 2008. PMID: 19029910.
- [33] R. de Gelder, R. Wehrens, and J. A. Hageman. A generalized expression for the similarity of spectra: application to powder diffraction pattern classification. *Journal of Computational Chemistry*, 22(3):273–289, 2001.
- [34] Winfried Denk and Heinz Horstmann. Serial Block-Face scanning electron microscopy to reconstruct Three-Dimensional tissue nanostructure. *PLoS Biol*, 2(11):e329, October 2004.
- [35] Ramon Diaz-Uriarte and Sara Alvarez de Andres. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1):3, 2006.
- [36] John E. Dowling. *Neurons and networks: an introduction to behavioral neuroscience*. Harvard University Press, 2001.
- [37] F. Dubois, R. Knochenmuss, R. Zenobi, A. Brunelle, C. Deprun, and Y. Le Beyec. A comparison between ion-to-photon and microchannel plate detectors. *Rapid Communications in Mass Spectrometry*, 13(9):786–791, 1999.
- [38] J. E. Eckel-Passow, A. L. Oberg, T. M. Therneau, C. J. Mason, D. W. Mahoney, K. L. Johnson, J. E. Olson, and H. R. Bergen. Regression analysis for comparing protein samples with $^{16}\text{O}/^{18}\text{O}$ stable-isotope labeled mass spectrometry. *Bioinformatics*, 22(22):2739–2745, November 2006.
- [39] P Edman. Sequence determination. *Molecular Biology, Biochemistry, and Biophysics*, 8:211–255, 1970. PMID: 4950190.
- [40] Pehr Edman, Erik Högfeldt, Lars Gunnar Sillen, and Per-Olof Kinell. Method for determination of the amino acid sequence in peptides. *Acta Chemica Scandinavica*, 4:283–293, 1950.
- [41] Bradley Efron. Least angle regression. *The Annals of Statistics*, 32(2):407–499. Mathematical Reviews number (MathSciNet): MR2060166; Zentralblatt MATH identifier: 02100802.
- [42] Ingvar Eidhammer. *Computational methods for mass spectrometry proteomics*. John Wiley & Sons, 2007.

-
- [43] Paul H. C. Eilers. Parametric time warping. *Anal. Chem.*, 76(2):404–411, 2003.
- [44] Jimmy Eng, Ashley McCormack, and John Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, November 1994.
- [45] S W Englander, L Mayne, Y Bai, and T R Sosnick. Hydrogen exchange: the modern legacy of Linderström-Lang. *Protein Science: A Publication of the Protein Society*, 6(5):1101–1109, May 1997. PMID: 9144782.
- [46] JB Fenn, M Mann, CK Meng, SF Wong, and CM Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, October 1989.
- [47] Catherine Fenselau and Xudong Yao. 18O₂-Labeling in quantitative proteomic strategies: A status report. *Journal of Proteome Research*, 8(5):2140–2143, May 2009.
- [48] Debra M Ferraro, Noel D Lazo, and Andrew D Robertson. EX1 hydrogen exchange and protein folding. *Biochemistry*, 43(3):587–594, January 2004. PMID: 14730962.
- [49] Y. Geinisman, H. J. G. Gundersen, E. Zee, and M. J. West. Unbiased stereological estimation of the total number of synapses in a brain region. *Journal of Neurocytology*, 25(1):805–819, 1996.
- [50] Bram Van Ginneken, Tobias Heimann, and Martin Styner. M.: 3D segmentation in the clinic: A grand challenge. In *MICCAI Workshop on 3D Segmentation in the Clinic: a Grand Challenge. (2007)*.
- [51] Christian Graf, Marta Stankiewicz, Gunter Kramer, and Matthias P Mayer. Spatially and kinetically resolved changes in the conformational dynamics of the hsp90 chaperone machine. *EMBO J*, 28(5):602–613, March 2009.
- [52] L. Guo, Y. Ma, B. Cukic, and Harshinder Singh. Robust prediction of fault-proneness by random forests. In *Software Reliability Engineering, 2004. ISSRE 2004. 15th International Symposium on*, pages 417–428, 2004.
- [53] Mats G L Gustafsson. Nonlinear structured-illumination microscopy: wide-field fluorescence imaging with theoretically unlimited resolution. *Proceedings of the National Academy of Sciences of the United States of America*, 102(37):13081–13086, September 2005. PMID: 16141335.
- [54] S P Gygi, B Rist, S A Gerber, F Turecek, M H Gelb, and R Aebersold. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, 17(10):994–999, October 1999. PMID: 10504701.

- [55] Brian D. Halligan, Ronit Y. Slyper, Simon N. Twigger, Wayne Hicks, Michael Olivier, and Andrew S. Greene. ZoomQuant: an application for the quantitation of stable isotope labeled peptides. *Journal of the American Society for Mass Spectrometry*, 16(3):302–306, March 2005. PMID: 15734322 PMCID: 2793075.
- [56] David K. Han, Jimmy Eng, Huilin Zhou, and Ruedi Aebersold. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nature biotechnology*, 19(10):946–951, October 2001. PMID: 11581660 PMCID: 1444949.
- [57] KJ Hayworth, N Kasthuri, R Schalek, and JW Lichtman. Automating the collection of ultrathin serial sections for large volume TEM reconstructions. *Microscopy and Microanalysis*, 12(Supplement S02):86–87, 2006.
- [58] Moritz Helmstaedter, Kevin L Briggman, and Winfried Denk. 3D structural imaging of the brain with photons and electrons. *Current Opinion in Neurobiology*, 18(6):633–641, December 2008. PMID: 19361979.
- [59] Moritz Helmstaedter, Kevin L Briggman, and Winfried Denk. High-accuracy neurite reconstruction for high-throughput neuroanatomy. *Nat Neurosci*, 14(8):1081–1088, 2011.
- [60] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964–972, December 2007. PMID: 18075575.
- [61] Julia Herold, Walter Schubert, and Tim W Nattkemper. Automated detection and quantification of fluorescently labeled synapses in murine brain tissue sections for high throughput applications. *Journal of Biotechnology*, 149(4):299–309, September 2010. PMID: 20230863.
- [62] Matthew Hotchko, Ganesh S. Anand, Elizabeth A. Komives, and Lynn F. Ten Eyck. Automated extraction of backbone deuteration levels from amide H/2H mass spectrometry experiments. *Protein Science : A Publication of the Protein Society*, 15(3):583–601, March 2006. PMID: 16501228 PMCID: 2249778.
- [63] Qizhi Hu, Robert J. Noll, Hongyan Li, Alexander Makarov, Mark Hardman, and R. Graham Cooks. The orbitrap: a new mass spectrometer. *Journal of Mass Spectrometry*, 40(4):430–443, 2005.
- [64] T. William Hutchens and Tai-Tung Yip. New desorption strategies for the mass spectrometric analysis of macromolecules. *Rapid Communications in Mass Spectrometry*, 7(7):576–580, 1993.

-
- [65] A Hvidt and K Linderstrom-Lang. Exchange of hydrogen atoms in insulin with deuterium atoms in aqueous solutions. *Biochimica Et Biophysica Acta*, 14(4):574–575, August 1954. PMID: 13198919.
- [66] V. Jain, B. Bollmann, M. Richardson, D.R. Berger, M.N. Helmstaedter, K.L. Briggman, W. Denk, J.B. Bowden, J.M. Mendenhall, W.C. Abraham, K.M. Harris, N. Kasthuri, K.J. Hayworth, R. Schalek, J.C. Tapia, J.W. Lichtman, and H.S. Seung. Boundary learning by optimization with topological constraints. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2488–2495, 2010.
- [67] Viren Jain, Joseph F. Murray, Fabian Roth, Srinivas Turaga, Valentin Zhigulin, Kevin L. Briggman, Moritz N. Helmstaedter, Winfried Denk, and H. Sebastian Seung. Supervised learning of image restoration with convolutional networks. In *Computer Vision, IEEE International Conference on*, volume 0, pages 1–8, Los Alamitos, CA, USA, 2007. IEEE Computer Society.
- [68] Viren Jain, H Sebastian Seung, and Srinivas C Turaga. Machines that learn to segment images: a crucial technology for connectomics. *Current Opinion in Neurobiology*, 20(5):653–666, October 2010. PMID: 20801638.
- [69] Kenneth L. Johnson and David C. Muddiman. A method for calculating 16o/18o peptide ion ratios for the relative quantification of proteomes. *Journal of the American Society for Mass Spectrometry*, 15(4):437–445, 2004.
- [70] Elizabeth Jurrus, Antonio R C Paiva, Shigeki Watanabe, James R Anderson, Bryan W Jones, Ross T Whitaker, Erik M Jorgensen, Robert E Marc, and Tolga Tasdizen. Detection of neuron membranes in electron microscopy images using a serial neural network architecture. *Medical Image Analysis*, 14(6):770–783, December 2010. PMID: 20598935.
- [71] Robert H Kassel. A comparison of approaches to on-line handwritten character recognition. <http://dspace.mit.edu/handle/1721.1/11407>. Thesis (Ph. D.)—Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science, 1995.
- [72] Bernhard X. Kausler, Marc Kirchner, Anna Kreshuk, Bernhard Y. Renard, H Hahne, Bernhard Kuster, Judith A. J. Steen, Hanno Steen, and Fred A. Hamprecht. Physically motivated computational models for Mass/Charge-Dependent resolution changes in mass spectra. *Conference of the American Society for Mass Spectrometry (ASMS)*, 2010.

- [73] Verena Kaynig, Ignacio Arganda-Carreras, and Albert Cardona. Trainable segmentation plugin. http://fiji.sc/wiki/index.php/Trainable_Segmentation_Plugin, March 2010.
- [74] Verena Kaynig, Thomas Fuchs, and Joachim M. Buhmann. Neuron geometry extraction by perceptual grouping in ssTEM images. In *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, volume 0, pages 2902–2909, Los Alamitos, CA, USA, 2010. IEEE Computer Society.
- [75] Neil L. Kelleher, Hong Y. Lin, Gary A. Valaskovic, David J. Aaserud, Einar K. Fridriksson, and Fred W. McLafferty. Top down versus bottom up protein characterization by tandem High-Resolution mass spectrometry. *J. Am. Chem. Soc.*, 121(4):806–812, 1999.
- [76] Andrew Keller, Jimmy Eng, Ning Zhang, Xiao jun Li, and Ruedi Aebersold. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol*, 1, 2005.
- [77] Arminja N Kettenbach, John Rush, and Scott A Gerber. Absolute quantification of protein and post-translational modification abundance with stable isotope-labeled synthetic peptides. *Nat. Protocols*, 6(2):175–186, January 2011.
- [78] Zia Khan, Joshua S. Bloom, Benjamin A. Garcia, Mona Singh, and Leonid Kruglyak. Protein quantification across hundreds of experimental conditions. *Proceedings of the National Academy of Sciences*, 106(37):15544–15548, 2009.
- [79] M. Knoll and E. Ruska. Das elektronenmikroskop. *Zeitschrift für Physik*, 78(5-6):318–339, 1932.
- [80] Graham Knott, Herschel Marchman, David Wall, and Ben Lich. Serial section scanning electron microscopy of adult brain tissue using focused ion beam milling. *J. Neurosci.*, 28(12):2959–2964, March 2008.
- [81] Graham W. Knott, Charles Quairiaux, Christel Genoud, and Egbert Welker. Formation of dendritic spines with GABAergic synapses induced by whisker stimulation in adult mice. *Neuron*, 34(2):265–273, April 2002.
- [82] Ullrich Köthe. *vigra - vigra: VIGRA reference manual*. <http://hci.iwr.uni-heidelberg.de/vigra/doc/vigra/index.html>.
- [83] Ullrich Köthe. *Generische Programmierung für die Bildverarbeitung*. BoD – Books on Demand, September 2000.
- [84] Anna Kreshuk, Marta Stankiewicz, Xinghua Lou, Marc Kirchner, Fred A. Hamprecht, and Matthias P. Mayer. Automated detection and analysis of bimodal

- isotope peak distributions in H/D exchange mass spectrometry using HeXicon. *International Journal of Mass Spectrometry*, 302(1-3):125–131, April 2011.
- [85] Mallela M. G. Krishna, Linh Hoang, Yan Lin, and S. Walter Englander. Hydrogen exchange methods to study protein folding. *Methods*, 34(1):51–64, September 2004.
- [86] Yoshiyuki Kubota. Important factors for the three-dimensional reconstruction of neuronal structures from serial ultrathin sections. *Frontiers in Neural Circuits*, 3, 2009.
- [87] Eva Lange, Ralf Tautenhahn, Steffen Neumann, and Clemens Gropl. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics*, 9(1):375, 2008.
- [88] Sabine Lange, Marc Sylvester, Michael Schumann, Christian Freund, and Eberhard Krause. Identification of Phosphorylation-Dependent interaction partners of the adapter protein ADAP using quantitative mass spectrometry: SILAC vs 18O-Labeling. *Journal of Proteome Research*, 9(8):4113–4122, 2010.
- [89] H. Liebl. *Ion Microprobe Mass Analyzer*. December 1967.
- [90] Xinghua Lou, Marc Kirchner, Bernhard Y. Renard, Ullrich Köthe, Sebastian Boppel, Christian Graf, Chung-Tien Lee, Judith A. J. Steen, Hanno Steen, Matthias P. Mayer, and Fred A. Hamprecht. Deuteration distribution estimation with improved sequence coverage for HX/MS experiments. *Bioinformatics*, 26(12):1535–1541, June 2010.
- [91] A. Lucchi, K. Smith, R. Achanta, V. Lepetit, and P. Fua. A fully automated approach to segmentation of irregularly shaped cellular structures in EM images. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, Beijing, China, 2010.
- [92] P. J Magelhaes, S. J Ram, and D Abramoff. Image processing with ImageJ. *Biophotonics International*, 11(7):36–42, 2004.
- [93] Alexander Makarov. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Analytical Chemistry*, 72(6):1156–1162, March 2000.
- [94] Christopher J. Mason, Terry M. Therneau, Jeanette E. Eckel-Passow, Kenneth L. Johnson, Ann L. Oberg, Janet E. Olson, K. Sreekumaran Nair, David C. Muddiman, and H. Robert Bergen. A method for automatically interpreting mass spectra of 18O-Labeled isotopic clusters. *Molecular & Cellular Proteomics*, 6(2):305–318, February 2007.

- [95] T. M. Mayhew. Stereological approach to the study of synapse morphometry with particular regard to estimating number in a volume and on a surface. *Journal of Neurocytology*, 8(2):121–138, 1979.
- [96] T. M. Mayhew. How to count synapses unbiasedly and efficiently at the ultrastructural level: proposal for a standard sampling and counting protocol. *Journal of Neurocytology*, 25(1):793–804, 1996.
- [97] Andrew McCallum, Fernando Pereira, and John Lafferty. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Departmental Papers (CIS)*, June 2001.
- [98] Bruce F. McEwen and Michael Marko. The emergence of electron tomography as an important tool for investigating cellular ultrastructure. *Journal of Histochemistry & Cytochemistry*, 49(5):553–563, May 2001.
- [99] Bjoern H Menze, B Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A Hamprecht. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, 10:213–213. PMID: 19591666 PMCID: 2724423.
- [100] Angel Merchan-Perez, Jose-Rodrigo Rodriguez, Lidia Alonso-Nanclares, Andreas Schertel, and Javier Defelipe. Counting synapses using FIB/SEM microscopy: A true revolution for ultrastructural volume reconstruction. *Frontiers in Neuroanatomy*, 3, 2009. PMID: 19949485.
- [101] Kristina D Micheva, Brad Busse, Nicholas C Weiler, Nancy O’Rourke, and Stephen J Smith. Single-synapse analysis of a diverse synapse population: proteomic imaging methods and markers. *Neuron*, 68(4):639–653, November 2010. PMID: 21092855.
- [102] Kristina D. Micheva and Stephen J Smith. Array tomography. *Neuron*, 55(1):25–36, July 2007. PMID: 17610815 PMCID: 2080672.
- [103] Olga A. Mirgorodskaya, Yuri P. Kozmin, Mikhail I. Titov, Roman Korner, Carsten P. Sonksen, and Peter Roepstorff. Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using ^{18}O -labeled internal standards. *Rapid Communications in Mass Spectrometry*, 14(14):1226–1232, 2000.
- [104] Shama P. Mirza, Andrew S. Greene, and Michael Olivier. ^{18}O labeling over a coffee break: A rapid strategy for quantitative proteomics. *Journal of Proteome Research*, 7(7):3042–3048, July 2008.

- [105] Yuriy Mishchenko. Automation of 3D reconstruction of neural tissue from large volume of conventional serial section transmission electron micrographs. *Journal of neuroscience methods*, 176(2):276–289, January 2009. PMID: 18834903 PMID: 2948845.
- [106] Yuriy Mishchenko, Tao Hu, Josef Spacek, John Mendenhall, Kristen M Harris, and Dmitri B Chklovskii. Ultrastructural analysis of hippocampal neuropil from the connectomics perspective. *Neuron*, 67(6):1009–1020, September 2010. PMID: 20869597.
- [107] Greg Mori, Xiaofeng Ren, Alexei A Efros, and Jitendra Malik. Recovering human body configurations: combining segmentation and recognition. In *Proceedings of the 2004 IEEE computer society conference on Computer vision and pattern recognition*, CVPR’04, pages 326–333, Washington, DC, USA, 2004. IEEE Computer Society.
- [108] Lukas N. Mueller, Mi-Youn Brusniak, D. R. Mani, and Ruedi Aebersold. An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J. Proteome Res.*, 7(1):51–61, October 2011.
- [109] Alexey I. Nesvizhskii, Andrew Keller, Eugene Kolker, and Ruedi Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.*, 75(17):4646–4658, 2003.
- [110] Pornpat Nikamanon, Elroy Pun, Wayne Chou, Marek Koter, and Paul Gershon. ”TOF2H”: a precision toolbox for rapid, high density/high coverage hydrogen-deuterium exchange mass spectrometry via an LC-MALDI approach, covering the data pipeline from spectral acquisition to HDX rate analysis. *BMC Bioinformatics*, 9(1):387, 2008.
- [111] Nobelprize.org. The nobel prize in chemistry 2002, October 2011.
- [112] DJ Olbris, P Winston, and Dmitri B Chklovskii. Raveler: a software for editing large segmented EM datasets. *In preparation*.
- [113] Shao-En Ong, Blagoy Blagoev, Irina Kratchmarova, Dan Bach Kristensen, Hanno Steen, Akhilesh Pandey, and Matthias Mann. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics*, pages M200025–MCP200, May 2002.
- [114] Olayinka Aduke Oyeyemi and Berkeley University of California. *Linking protein dynamics to protein function*. ProQuest, 2008.

- [115] Magnus Palmblad, Jos Buijs, and Per Hakansson. Automatic analysis of hydrogen/deuterium exchange mass spectra of peptides and proteins using calculations of isotopic distributions. *Journal of the American Society for Mass Spectrometry*, 12(11):1153–1162, November 2001.
- [116] BD Pascal, MJ Chalmers, SA Busby, CC Mader, MR Southern, NF Tsinoremas, and PR Griffin. The deuterator: software for the determination of backbone amide deuterium levels from H/D exchange MS data. *BMC Bioinformatics*, 8(1):156, 2007.
- [117] Bruce D. Pascal, Michael J. Chalmers, Scott A. Busby, and Patrick R. Griffin. HD desktop: An integrated platform for the analysis and visualization of H/D exchange data. *Journal of the American Society for Mass Spectrometry*, 20(4):601–610, April 2009.
- [118] Scott D. Patterson and Ruedi H. Aebersold. Proteomics: the first decade and beyond. *Nat Genet.*
- [119] Hanchuan Peng, Zongcai Ruan, Fuhui Long, Julie H Simpson, and Eugene W Myers. V3D enables real-time 3D visualization and quantitative analysis of large-scale biological image data sets. *Nat Biotech*, 28(4):348–353, April 2010.
- [120] 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. *Electrophoresis*, 20(18):3551–3567, 1999.
- [121] Brianne O. Petritis, Wei-Jun Qian, David G. Camp, and Richard D. Smith. A simple procedure for effective quenching of trypsin activity and prevention of 18O-Labeling Back-Exchange. *J. Proteome Res.*, 8(5):2157–2163, October 2011.
- [122] Amol Prakash, Parag Mallick, Jeffrey Whiteaker, Heidi Zhang, Amanda Paulovich, Mark Flory, Hookeun Lee, Ruedi Aebersold, and Benno Schwikowski. Signal maps for mass spectrometry-based comparative proteomics. *Molecular & Cellular Proteomics*, 5(3):423–432, March 2006.
- [123] Elaine R and Mardis. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3):133–141, 2008.
- [124] Bernhard Renard, Marc Kirchner, Hanno Steen, Judith Steen, and Fred Hamprecht. NITPICK: peak identification for mass spectrometry data. *BMC Bioinformatics*, 9(1):355, 2008.
- [125] Savarese Software Research. libssrckdtree.

- [126] Wolfgang Rist, Christian Graf, Bernd Bukau, and Matthias P. Mayer. Amide hydrogen exchange reveals conformational changes in hsp70 chaperones important for allosteric regulation. *Journal of Biological Chemistry*, 281(24):16493–16501, June 2006.
- [127] Wolfgang Rist, Thomas J. D. Jorgensen, Peter Roepstorff, Bernd Bukau, and Matthias P. Mayer. Mapping temperature-induced conformational changes in the escherichia coli heat shock transcription factor sigma32 by amide hydrogen exchange. *Journal of Biological Chemistry*, 278(51):51415–51421, December 2003.
- [128] John W. Robinson. *Focus on protein research*. Nova Publishers, October 2004.
- [129] Heinrich Roder, Gulnur A. Elove, and S. Walter Englander. Structural characterization of folding intermediates in cytochrome c by h-exchange labelling and proton NMR. *Nature*, 335(6192):700–704, October 1988.
- [130] Michael J Rust, Mark Bates, and Xiaowei Zhuang. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature Methods*, 3(10):793–795, October 2006. PMID: 16896339.
- [131] Stephan Saalfeld, Albert Cardona, Volker Hartenstein, and Pavel Tomancak. As-rigid-as-possible mosaicking and serial section registration of large ssTEM datasets. *Bioinformatics*, 26(12):i57–63, June 2010.
- [132] Rovshan G. Sadygov, Yingxin Zhao, Sigmund J. Haidacher, Jonathan M. Starkey, Ronald G. Tilton, and Larry Denner. Using power spectrum analysis to evaluate ¹⁸O-Water labeling data acquired from low resolution mass spectrometers. *Journal of Proteome Research*, January.
- [133] Markus Schirle, Marie-Anne Heurtier, and Bernhard Kuster. Profiling core proteomes of human cell lines by one-dimensional PAGE and liquid Chromatography-Tandem mass spectrometry. *Molecular & Cellular Proteomics*, 2(12):1297–1305, December 2003.
- [134] Sabine K Schmitz, J J Johannes Hjorth, Raoul MS Joemai, Rick Wijntjes, Susanne Eijgenraam, Petra de Bruijn, Christina Georgiou, Arthur PH de Jong, Arjen van Ooyen, Matthijs Verhage, L Niels Cornelisse, Ruud F Toonen, and Wouter Veldkamp. Automated analysis of neuronal morphology; synapse number and synaptic recruitment. *Journal of Neuroscience Methods*, In Press, Accepted Manuscript.
- [135] Will Schroeder, Ken Martin, and Bill Lorensen. *The Visualization Toolkit. An object oriented approach to 3D graphics*. Kitware, Inc., fourth edition edition.

- [136] Michaela Scigelova and Alexander Makarov. Advances in bioanalytical LC-MS using the orbitrap mass analyzer. *Bioanalysis*, 1(4):741–754, July 2009. PMID: 21083136.
- [137] Mojtaba Seyedhosseini, Ritwik Kumar, Elizabeth Jurrus, Rick Giuly, Mark Ellisman, Hanspeter Pfister, and Tolga Tasdizen. Detection of neuron membranes in electron microscopy images using multi-scale context and Radon-Like features. In Gabor Fichtinger, Anne Martel, and Terry Peters, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011*, volume 6891 of *Lecture Notes in Computer Science*, pages 670–677. Springer Berlin / Heidelberg, 2011. 10.1007/978-3-642-23623-5_84.
- [138] S Sherrington. Santiago Ramon y Cajal. 1852-1934. *Obituary Notices of Fellows of the Royal Society*, 1(4):424, 1935.
- [139] Andrew K. Shiau, Seth F. Harris, Daniel R. Southworth, and David A. Agard. Structural analysis of e. coli hsp90 reveals dramatic Nucleotide-Dependent conformational rearrangements. *Cell*, 127(2):329–340, October 2006.
- [140] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. TextonBoost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. In AleÅ; Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision – ECCV 2006*, volume 3951 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin / Heidelberg, 2006. 10.1007/11744023_1.
- [141] Gordon Slys, Charles Baker, Benjamin Bozsa, Anthony Dang, Andrew Percy, Melissa Bennett, and David Schriemer. Hydra: software for tailored processing of H/D exchange data from MS or tandem MS analyses. *BMC Bioinformatics*, 10(1):162, 2009.
- [142] Stephen J Smith. Circuit reconstruction tools today. *Current opinion in neurobiology*, 17(5):601–608, October 2007. PMID: 18082394 PMCID: 2693015.
- [143] C. Sommer, C. Straehle, U. Kothe, and F.A. Hamprecht. Ilastik: Interactive learning and segmentation toolkit. In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pages 230–233, 2011.
- [144] Hanno Steen and Matthias Mann. The abc’s (and xyz’s) of peptide sequencing. *Nat Rev Mol Cell Biol*, 5(9):699–711, 2004.
- [145] D C Sterio. The unbiased estimation of number and sizes of arbitrary particles using the disector. *Journal of Microscopy*, 134(Pt 2):127–136, May 1984. PMID: 6737468.

-
- [146] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer, October 2010.
- [147] Koichi Tanaka, Hiroaki Waki, Yutaka Ido, Satoshi Akita, Yoshikazu Yoshida, Tamio Yoshida, and T. Matsuo. Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, 2(8):151–153, 1988.
- [148] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, January 1996. ArticleType: research-article / Full publication date: 1996 / Copyright © 1996 Royal Statistical Society.
- [149] Zhuowen Tu. Probabilistic Boosting-Tree: learning discriminative models for classification, recognition, and clustering. In *Computer Vision, IEEE International Conference on*, volume 2, pages 1589–1596, Los Alamitos, CA, USA, 2005. IEEE Computer Society.
- [150] Zhuowen Tu and Xiang Bai. Auto-Context and its application to High-Level vision tasks and 3D brain image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(10):1744–1757, 2010.
- [151] Srinivas C Turaga, Kevin L Briggman, Moritz Helmstaedter, Winfried Denk, and (last) H Sebastian Seung. Maximum affinity learning of image segmentation. 2009.
- [152] Mike Tyers and Matthias Mann. From genomics to proteomics. *Nature*, 422(6928):193–197, March 2003.
- [153] Mathias Vandenberghe, Sebastien Li-Thiao-Te, Hans-Michael Kaltenbach, Runxuan Zhang, Tero Aittokallio, and Benno Schwikowski. Alignment of LC-MS images, with applications to biomarker discovery and protein identification. *PROTEOMICS*, 8(4):650–672, 2008.
- [154] Ashok Veeraraghavan, Alex. V. Genkin, Shiv Vitaladevuni, Lou Scheffer, Shan Xu, Harald Hess, Richard Fetter, Marco Cantoni, Graham Knott, and Dmitri Chklovskii. Increasing depth resolution of electron microscopy of neural circuits using sparse tomographic reconstruction. In *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, volume 0, pages 1767–1774, Los Alamitos, CA, USA, 2010. IEEE Computer Society.
- [155] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, volume 1, page 511, Los Alamitos, CA, USA, 2001. IEEE Computer Society.

- [156] Björn Voss, Michael Hanselmann, Bernhard Y. Renard, Martin S. Lindner, Ullrich Köthe, Marc Kirchner, and Fred A. Hamprecht. SIMA: simultaneous multiple alignment of LC/MS peak lists. *Bioinformatics*, 27(7):987–993, April 2011.
- [157] R W Ware. Three-dimensional reconstruction from serial sections. *International Review of Cytology*, 40:325–440, 1975. PMID: 1097356.
- [158] David D. Weis, John R. Engen, and Ignatius J. Kass. Semi-Automated data processing of hydrogen exchange mass spectra using HX-Express. *Journal of the American Society for Mass Spectrometry*, 17(12):1700–1703, December 2006.
- [159] David D. Weis, Thomas E. Wales, John R. Engen, Matthew Hotchko, and Lynn F. Ten Eyck. Identification and characterization of EX1 kinetics in H/D exchange mass spectrometry by peak width analysis. *Journal of the American Society for Mass Spectrometry*, 17(11):1498–1509, November 2006.
- [160] Carl A. White, Nicodemus Oey, and Andrew Emili. Global quantitative proteomic profiling through ^{18}O -Labeling in combination with MS/MS spectra analysis. *J. Proteome Res.*, 8(7):3653–3665, October 2011.
- [161] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner. The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 314(1165):1–340, November 1986.
- [162] Xudong Yao, Carlos Afonso, and Catherine Fenselau. Dissection of proteolytic ^{18}O labeling: Endoprotease-Catalyzed ^{16}O -to- ^{18}O exchange of truncated peptide substrates. *Journal of Proteome Research*, 2(2):147–152, April 2003.
- [163] Xudong Yao, Amy Freas, Javier Ramirez, Plamen A. Demirev, and Catherine Fenselau. Proteolytic ^{18}O labeling for comparative proteomics: Model studies with two serotypes of adenovirus. *Analytical Chemistry*, 73(13):2836–2842, July 2001.
- [164] Xiaoying Ye, Brian T. Luke, Donald J. Johann, Akira Ono, DaRue A. Prieto, King C. Chan, Haleem J. Issaq, Timothy D. Veenstra, and Josip Blonder. Optimized method for computing $^{18}\text{O}/^{16}\text{O}$ ratios of differentially Stable-Isotope labeled peptides in the context of postdigestion ^{18}O Exchange/Labeling. *Anal. Chem.*, 82(13):5878–5886, October 2011.
- [165] Guoan Zhang and Thomas A. Neubert. Automated comparative proteomics based on multiplex tandem mass spectrometry and stable isotope labeling. *Mol Cell Proteomics*, 5(2):401–411, February 2006.