

Computational Proteomics

Edited by

Sebastian Böcker¹, Rebekah Gundry², Lennart Martens³, and
Magnus Palmblad⁴

1 Universität Jena, DE, sebastian.boecker@uni-jena.de

2 University of Nebraska – Omaha, US, rebekah.gundry@unmc.edu

3 Ghent University, BE, lennart.martens@ugent.be

4 Leiden University Medical Center, NL, n.m.palmblad@lumc.nl

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 21271 “Computational Proteomics”. The Seminar, which took place in a hybrid fashion with both local as well as online participation due to the COVID pandemic, was built around three topics: the rapid uptake of advanced machine learning in proteomics; computational challenges across the various rapidly evolving approaches for structural and top-down proteomics; and the computational analysis of glycoproteomics data. These three topics were the focus of three corresponding breakout sessions, which ran in parallel throughout the seminar. A fourth breakout session was created during the seminar, on the specific topic of creating a Kaggle competition based on proteomics data.

The abstracts presented here first describe the three introduction talks, one for each topic. These talk abstracts are then followed by one abstract each *per* breakout session, documenting that breakout’s discussion and outcomes.

An Executive Summary is also provided, which details the overall seminar structure alongside the most important conclusions for the three topic-derived breakouts.

Seminar July 4–9, 2021 – <http://www.dagstuhl.de/21271>

2012 ACM Subject Classification Applied computing → Bioinformatics

Keywords and phrases bioinformatics, computational mass spectrometry, machine learning, proteomics

Digital Object Identifier 10.4230/DagRep.11.6.1

1 Executive Summary

Lennart Martens (Ghent University, BE)

Rebekah Gundry (University of Nebraska – Omaha, US)

Magnus Palmblad (Leiden University Medical Center, NL)

License © Creative Commons BY 4.0 International license
© Lennart Martens, Rebekah Gundry, and Magnus Palmblad

The Dagstuhl Seminar 21271 “Computational Proteomics” discussed several important developments, challenges, and opportunities that are emerging in the field of computational proteomics. Three core topics were set out at the start, and these were discussed at length throughout the seminar.

These three topics were: (i) the fast evolving use of advanced machine learning approaches in proteomics; (ii) the challenges and opportunities offered by fast developing approaches for structural and top-down proteomics; and (iii) specific issues and computational complications in glycoproteomics.



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Computational Proteomics, *Dagstuhl Reports*, Vol. 11, Issue 06, pp. 1–13

Editors: Sebastian Böcker, Rebekah Gundry, Lennart Martens, and Magnus Palmblad



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The machine learning and glycoproteomics topics were each introduced by a dedicated lecture, which set out the current state-of-the-art and presented a tentative set of issues, challenges, or opportunities that could be explored during the seminar. The structural and top-down proteomics topic was introduced by two sequential lectures, one on structural proteomics, and one on top-down proteomics. In total, four introductory talks were thus presented at the start of the seminar. For each of the three main topics, daily Working Group sessions were organised, which took place in the morning and afternoon, with a daily late-night session scheduled each day to wrap up the day's outcomes. This structure was followed to allow maximum involvement by online participants across the various timezones in the hybrid format. The Machine Learning in Proteomics Working Group also spun out another Working Group session during the seminar, which discussed the creation of a machine learning (Kaggle-like) competition based on proteomics data.

Each of these breakout sessions was very actively attended, including by online attendees, and resulted in several interesting research ideas and potential new initiatives. The Machine Learning in Proteomics Working Group was the largest working group, and addressed a number of distinct topics during the seminar. Of particular note were the spin-out effort to establish two machine learning competitions based on proteomics data and challenges to engage the broader machine learning community, and the extensive discussions on the optimal way to represent mass spectrometry data for downstream machine learning.

The Glycoproteomics Working Group was very actively attended, and discussed an exciting set of topics. A first highlight among these topics was provided by the extensive and detailed discussions with the Machine Learning Working Group regarding the potential of, and road towards, the use of state-of-the-art machine learning approaches in glycoproteomics. A second highlight concerned the delineation of a set of high-impact opinion papers to describe the state-of-the-art of the field, and its goals, ambitions, and challenges.

The Structural and Top-Down Proteomics Working Group was very active in detailing the many challenges and opportunities in this fast-evolving field. One noteworthy challenge revolved around the detection, annotation, and biological interpretation of post-translational modifications detected by mass spectrometry. A second challenge concerned the standardization of acquired native mass spectrometry data, the minimal reporting requirements for these experiments, and the dissemination of these data.

Overall, the 2021 Dagstuhl Seminar on Computational Proteomics was extremely successful as a catalyst for careful yet original thinking about key challenges in the field, and as a means to enable downstream progress by setting important, high impact goals to work on in close collaboration. During this Seminar, new topics for a future Seminar were suggested throughout as well, indicating that this active field will continue to yield novel challenges and opportunities for advanced computational work going forward.

2 Table of Contents

Executive Summary

<i>Lennart Martens, Rebekah Gundry, and Magnus Palmblad</i>	1
---	---

Overview of Talks

Topic Introduction: Shotgun Cross-Linking Mass Spectrometry and Protein Structure Prediction <i>Michael Hoopmann</i>	4
Topic Introduction: Future Outlook & Opportunities for Top Down Structural Proteomics <i>Neil Kelleher</i>	4
Topic Introduction: Machine Learning for MS-based Proteomics <i>Lukas Käll</i>	4
Topic Introduction: Glycoproteomics Challenges and Opportunities <i>Frédérique Lisacek</i>	5

Working groups

Working Group Report: Machine Learning in Proteomics <i>Lukas Käll, Marshall Bern, Sebastian Böcker, Sven Degrove, Bernard Delanghe, Viktoria Dorfer, Daniel Kolarich, Frédérique Lisacek, Magnus Palmblad, Robin Park, Veit Schwämmle, Matthew Smith, and Mathias Wilhelm</i>	5
Working Group Report: Glycoproteomics <i>Frédérique Lisacek, Kiyoko Aoki-Kinoshita, Marshall Bern, Sebastian Böcker, Robert Chalkley, Bernard Delanghe, Viktoria Dorfer, Patrick Emery, Rebekah Gundry, Michael Hoopmann, Lukas Käll, Neil Kelleher, Joanna Kirkpatrick, Daniel Kolarich, Lennart Martens, Nicki Packer, Magnus Palmblad, Daniel Questschlich, Veit Schwämmle, Matthew Smith, Sabarinath Peruvemba Subramanian, Morten Thaysen-Andersen, Lilla Turiák, and Mathias Wilhelm</i>	7
Working Group Report: Machine Learning (Kaggle) Competitions Based on Proteomics Data <i>Magnus Palmblad, Viktoria Dorfer, and Veit Schwämmle</i>	9
Working Group Report: Structural and Top-Down Proteomics <i>Daniel Questschlich, Bernard Delanghe, Viktoria Dorfer, Patrick Emery, Michael Hoopmann, Lukas Käll, Neil Kelleher, Magnus Palmblad, Veit Schwämmle, Matthew Smith, and Mathias Wilhelm</i>	10

Participants	12
-------------------------------	----

Remote Participants	12
--------------------------------------	----

3 Overview of Talks

3.1 Topic Introduction: Shotgun Cross-Linking Mass Spectrometry and Protein Structure Prediction

Michael Hoopmann (Institute for Systems Biology – Seattle, US)

License  Creative Commons BY 4.0 International license
© Michael Hoopmann

Crosslinking and bottom-up mass spectrometry (XL-MS) seeks to aid protein structure prediction and macromolecular structure assembly. The structural prediction community, however, has been slow to adopt and incorporate XL-MS technology. A recent study of CASP13 illustrated that only 12% of participants chose to compete using XL-MS. Better adoption of XL-MS for structural prediction requires improved quality and accuracy of XL-MS results and better computational pipelines for users to incorporate XL-MS into their research. We should consider how to improve the interaction between the XL-MS and structural prediction communities and accelerate the development of methods and pipelines that better integrate these technologies into robust computational tools.

3.2 Topic Introduction: Future Outlook & Opportunities for Top Down Structural Proteomics

Neil Kelleher (Northwestern University – Evanston, US)

License  Creative Commons BY 4.0 International license
© Neil Kelleher

In this brief orientation seminar, timely topics in computational proteomics as they relate to denatured and native mode top-down proteomics are presented. This includes the detection of proteoforms, their post-translational modifications (PTMs), and their complexes. Automation platforms for data creation and real-time search, processing new individual ion (i2MS) datatypes, and integration of compositional top-down proteomics with structural proteomics are also discussed. Importantly, the prospect of a Human Proteoform Project and Atlas was proposed, framed, and discussed.

3.3 Topic Introduction: Machine Learning for MS-based Proteomics

Lukas Käll (KTH Royal Institute of Technology – Solna, SE)

License  Creative Commons BY 4.0 International license
© Lukas Käll

Currently, machine learning (ML) is revolutionizing the way we interpret data. Here I will give a brief background to classical ML. I will also point out how ML is helped by various deep learning structures that can learn feature representations of a sample point. Particularly, an encoder-decoder structure known as Transformers promises to change the way we handle sequential data, by enabling transfer learning.


Machine Learning, especially Deep Learning, requires non-trivial amounts of training data. Even though much proteomics data is available in repositories, it is not immediately accessible to ML. Röttger and colleagues (Rehfeldt 2021) recently uploaded a preprint describing how to transform proteomics LC-MS data in public repositories (e.g. PRIDE) to be used for ML.

We can also formulate some potentially relevant questions that we can ask in relation to ML in proteomics, with specific pertinence to this seminar:

- How can transfer learning reduce the need for training data in similar ML applications?
- How can ML be applied to glycomics, for example to predict chromatographic behavior and fragmentation of released glycans or glycopeptides?
- How can ML-based protein structure prediction be combined with top-down or bottom-up strategies for structural proteomics?

3.4 Topic Introduction: Glycoproteomics Challenges and Opportunities

Frédérique Lisacek (Swiss Institute of Bioinformatics – Genève, CH)


License  Creative Commons BY 4.0 International license
© Frédérique Lisacek

First, broad goals, topics of interest, and bottlenecks in the field of glycoproteomics were identified. From here, three priority areas of potential discussion were defined. These priorities include: 1) outline a white paper that will serve as a tangible outcome of the forthcoming discussions; 2) envision how machine learning could be implemented to improve glycoproteomics analysis; and 3) define strategies to increase accuracy of glycoproteomics results. This latter element is a major challenge in the field of glycoproteomics, as there is a lack of established guidelines for assessing accuracy of search results. While sample preparation, data acquisition, and multiple data search tools have become increasingly accessible to many laboratories, the lack of expertise in basic principles of glycobiology can present challenges to accurate data reporting. Several ideas for increasing accuracy in glycoproteomic results were presented, including 1) strategies to integrate knowledge of biosynthetic pathways into routine data analysis processes, and 2) the need for FDR calculations suitable for intact glycopeptides.

4 Working groups

4.1 Working Group Report: Machine Learning in Proteomics

Lukas Käll (KTH Royal Institute of Technology – Solna, SE), Marshall Bern (Protein Metrics – Cupertino, US), Sebastian Böcker (Universität Jena, DE), Sven Degrove (Ghent University, BE), Bernard Delanghe (Thermo Fisher GmbH – Bremen, DE), Viktoria Dorfer (University of Applied Sciences Upper Austria, AT), Daniel Kolarich (Griffith University – Southport, AU), Frédérique Lisacek (Swiss Institute of Bioinformatics – Genève, CH), Magnus Palmblad (Leiden University Medical Center, NL), Robin Park (Bruker – Rancho Santa Fe, US), Veit Schwämmle (University of Southern Denmark – Odense, DK), Matthew Smith (University of Texas – Austin, US), and Mathias Wilhelm (TU München – Freising, DE)

License  Creative Commons BY 4.0 International license
© Lukas Käll, Marshall Bern, Sebastian Böcker, Sven Degrove, Bernard Delanghe, Viktoria Dorfer, Daniel Kolarich, Frédérique Lisacek, Magnus Palmblad, Robin Park, Veit Schwämmle, Matthew Smith, and Mathias Wilhelm

This abstract summarises the progress made by the Machine Learning in Proteomics Working Group over the course of the entire seminar.

First, the most common scenarios of machine learning in proteomics and computational mass spectrometry were discussed, and from this starting point, future trends were envisioned, including the combination of different models for different acquisition methods, and the prediction of particular mass spectrometry and biological features such as distinction of isobaric glycans *via* retention time, prediction of enzyme activity, and disease association. Particularly, top-down mass spectrometry still faces several important challenges that could be facilitated by machine learning applications, which include prediction of charge distributions of intact proteins as well as specialized applications to decipher non-linear peptides and proteins (cross-linking, cyclic peptides or proteins, and protein ubiquitination).

Several examples of end-to-end prediction *via* machine learning, mostly through currently highly prolific deep learning approaches, were discussed. The most prominent of these were to determine peptide, charge, and modifications from fragmentation mass spectra. Moreover, it was noted that a lack of sufficiently simple use cases for non-proteomics experts are missing, which, if made available, could be used to challenge the machine learning community at large to participate in future developments and innovation. In addition, such use cases could also push existing proteomics informatics community efforts forward by allowing benchmarking studies to take place.

The combination of two common machine learning methods, namely spectral clustering and predictive machine learning, were extensively discussed. Relationships between fragment ions across, e.g., fragmentation techniques could be summarized into a generation function by using experimental and even predicted data that incorporates covariance patterns and thus the variability of the very different types of fragmentation spectra of a peptide as delivered through the various fragmentation techniques. Chimaeric spectra (which contain fragments from more than one fragmented precursor peptide) and modifications could also be easier to distinguish if such information were to be included in identification algorithms.

On the second day, six different topics were explored.

1. Embedding and clustering. Three tasks were discussed that should be achievable using a “simple” representation of a mass spectrum. These tasks were (i) make spectral clustering algorithms run much faster, (ii) improve the power for a particular application, and (iii) make spectral data more readily accessible to machine learning methods.
2. Data sets for competitions: Deep learning challenges in proteomics should be sufficiently simplified to attract the involvement of the broader machine learning community. We discussed two use cases that look into specific problems in peptide MS, and this specific sub-topic became the focus of a separate, spin-out Working Group on a proteomics-based machine learning competition. A distinct abstract is provided for this Working Group, and the interested reader is directed there for more detailed information.
3. Combining models: The discussion started on the differentiation between the development of a single model that covers multiple peptide properties *versus* the combination of multiple predictions via post-processing, and their respective use-cases. A related issue was raised in that some peptides appear to be eluting multiple times in the same chromatogram, and speculation ensued as to the associated consequences on prediction accuracy and downstream data analysis pipelines. The conclusion was that this topic deserves to be investigated in more detail going forward.
4. Metaproteomics: We discussed the different ways in which machine learning could be used in peptide and protein (family) identification, and pathway and gene ontology term enrichment analysis. It was decided that this is a very Interesting and potentially quite fertile topic, and that it will quite likely be possible to transfer machine learning approaches already developed in the sibling fields of metagenomics and metatranscriptomics to make inroads into this issue.

5. Protein inference: different protein inference strategies were discussed, with a particular focus on protein fluorosequencing. It was concluded that there still is ample room for improvement and for new methods to tackle this already well-established challenge. It was also considered at some length whether non-unique peptides (i.e., peptides that match to more than one potential originator protein) could be helpful at all in resolving protein inference, but there no consensus was reached on the utility of such peptides.
6. Reporting standards: the proposed DOME reporting guidelines for supervised machine learning were discussed in the context of mass spectrometry-based proteomics. A potential commentary on the DOME paper was outlined, which will interpret these guidelines specifically for the proteomics community.

Another topic of great interest, concerned the best method to encode mass spectra for downstream machine learning. Despite intense discussions on this topic, there was no consensus on what currently constitutes the optimal method for encoding mass spectra for machine learning. However, a number of potential improvements on existing, naive methods were suggested and discussed to move the field forward. It was also noted that spectral encoding, spectral distance metrics, and spectral clustering are all highly interrelated problems. This because every encoding implicitly suggests a distance metric and a clustering method. Clearly, this topic is worthy of more detailed study as well.

4.2 Working Group Report: Glycoproteomics

Frédérique Lisacek (Swiss Institute of Bioinformatics – Genève, CH), Kiyoko Aoki-Kinoshita (Soka University – Tokyo, JP), Marshall Bern (Protein Metrics – Cupertino, US), Sebastian Böcker (Universität Jena, DE), Robert Chalkley (University of California – San Francisco, US), Bernard Delanghe (Thermo Fisher GmbH – Bremen, DE), Viktoria Dorfer (University of Applied Sciences Upper Austria, AT), Patrick Emery (Matrix Science Ltd. – London, GB), Rebekah Gundry (University of Nebraska – Omaha, US), Michael Hoopmann (Institute for Systems Biology – Seattle, US), Lukas Käll (KTH Royal Institute of Technology – Solna, SE), Neil Kelleher (Northwestern University – Evanston, US), Joanna Kirkpatrick (The Francis Crick Institute – London, GB), Daniel Kolarich (Griffith University – Southport, AU), Lennart Martens (Ghent University, BE), Nicki Packer (Macquarie University – Sydney, AU), Magnus Palmblad (Leiden University Medical Center, NL), Daniel Questschlich (University of Oxford, GB), Veit Schwämmle (University of Southern Denmark – Odense, DK), Matthew Smith (University of Texas – Austin, US), Sabarinath Peruvemba Subramanian (University of Nebraska – Omaha, US), Morten Thaysen-Andersen (Macquarie University – Sydney, AU), Lilla Turiák (Research Centre for Natural Sciences – Budapest, HU), and Mathias Wilhelm (TU München – Freising, DE)

License © Creative Commons BY 4.0 International license

© Frédérique Lisacek, Kiyoko Aoki-Kinoshita, Marshall Bern, Sebastian Böcker, Robert Chalkley, Bernard Delanghe, Viktoria Dorfer, Patrick Emery, Rebekah Gundry, Michael Hoopmann, Lukas Käll, Neil Kelleher, Joanna Kirkpatrick, Daniel Kolarich, Lennart Martens, Nicki Packer, Magnus Palmblad, Daniel Questschlich, Veit Schwämmle, Matthew Smith, Sabarinath Peruvemba Subramanian, Morten Thaysen-Andersen, Lilla Turiák, and Mathias Wilhelm

This abstract summarises the progress made by the Glycoproteomics Working Group over the course of the entire seminar.

During the first discussions, a few overall goals were outlined for the Working Group, including the delineation of the contents of a white paper on the current state of the field of glycoproteomics, an effort to integrate with the Machine Learning Working Group, and the definition of outstanding questions related to the bioinformatics in the field.

For the white paper, a few key topics of interest were quickly identified. A first was the need to allow the evaluation of the accuracy of glycoproteomics software, also by non-experts. Another was the need to provide coherent and intuitive data visualisations of the obtained results, which are currently not readily available. A large, unmet need was also identified concerning quantification, where statistical issues such as imputation difficulties and site-specific *versus* modified peptide differential analysis have not yet been addressed. Of course, there is also search space complexity, which is an already well-known problem, with various approaches in use to tackle this issue. It may therefore be relevant to perform a systematic evaluation of the respective benefits and drawbacks of these varied approaches.

As to the integration with the Machine Learning Working Group, it is clear that machine learning is currently having a profound impact on classical proteomics, and continues to make inroads there for some of the most complex problems. It will therefore be highly interesting to connect these efforts more closely with the glycoproteomics field, as there may well be similar benefits to be had here. In this context, the ongoing development, and increased adoption, of ion mobility in state-of-the-art mass spectrometers is a possible starting point for such an integration. However, it will be necessary to consider the creation of gold-standard data sets for this, or at least benchmark data sets for validation and evaluation of such efforts, alongside the necessary large amounts of reliable data needed for model training in the first place.

When discussing the bioinformatics developments, the focus shifted quickly to the integration of known biosynthetic pathways into the automated data analysis process. Currently, any successful analysis in glycoproteomics hinges heavily on the researcher's expertise in glycobiology. It is therefore important to consider whether it would be possible to introduce the principles of glycobiology into the search engines, for instance during the construction of the search space. Another approach that could be relevant would be to construct sample-specific glycan libraries, which could have the same (or even more stringent) effect. At the same time, the limited studies performed so far on unrestricted searches indicate that their performance is not as bad as typically thought, keeping that avenue open for exploration as well.

On the second day, the example provided by the field of top-down proteomics as presented in the corresponding introductory talk was considered. Here, instead of a single white paper, three independent opinion pieces at considerable impact had been written instead. As a result, the overall white paper concept was turned into the planning of three opinion papers focused on: 1) standards for glycoproteomics, 2) the reanalysis of (at least seven) published datasets of the SARS-CoV-2 spike glycoprotein, and 3) ways to address FDR calculation in glycoproteomics.

The content of 1) would span the different ways to optimize for, and ensure generation of, high quality data, while also describing the challenges involved with some of the standards; 2) would promote the multiplicity of methods and data; and 3) would cover the broad diversity of computational issues of intact glycopeptide identification, especially scoring functions.

Furthermore, a discussion was had on the possible input from machine learning into the field, and here several possibilities were proposed. First is the prediction of (relative) retention time prediction of glycosylated peptides and/or glycans. The goal would be to use these predictors as features in either a rescoring approach, and possibly to use these for isomer resolution. Another analyte (glycopeptide or glycan) behaviour to predict would be ion mobility. Further avenues for possible machine learning input were fine-tuning of false discovery rate calculations, peak picking from raw data (as peak shapes do not follow typical peptide patterns), and fragmentation method optimisation.

The shorter session on the third day focused on the abovementioned list of issues to be discussed with the Machine Learning Working Group, covering several topics in more depth, including retention time prediction. Recent analyses of the HGI challenge data were discussed as an introduction to the topic of FDR calculation.

The final day was first dedicated to a review of the conditions for setting up community challenges. Then, in order to maintain continuity with points developed earlier, the contents of the anticipated manuscripts were detailed further. In particular, a back-to-back presentation of wet and dry glyco-lab issues was decided upon. Moreover, a vigorous discussion developed between the Glycoproteomics and Machine Learning Working Groups, with several participants of the latter joining the former. Much of the discussion focused on ways in which machine learning approaches could be used for relative retention time prediction to increase confidence in glycopeptide assignments, and how this could possibly even add a level of structural detail to the typical compositional information. Experts from both sides of the discussion asked and answered questions regarding the unique challenges associated with glycopeptides and machine learning approaches (one-to-many relationship of peptide to glycans; compositional *versus* structural considerations; features of machine learning that may enable retention time prediction independently of the variability in data acquisition strategies; solutions that work for low complexity samples may not work for high complexity). Strategies discussed include incorporation of iso-electric focusing, knock-out animal data, redundant data (glycoproteomics, glycomics, deglycosylated proteomics), and top-down proteomics data. And while a consensus on how to solve the overall problems was not achieved, it was agreed that acquisition of data which can then be used for designing and testing machine learning approaches would be an important first step.

Finally, a detailed plan was outlined for a forthcoming manuscript focused on computational issues in glycoproteomics, writing assignments were distributed, and goals for the first follow-up meeting were defined.

4.3 Working Group Report: Machine Learning (Kaggle) Competitions Based on Proteomics Data

Magnus Palmblad (Leiden University Medical Center, NL), Viktoria Dorfer (University of Applied Sciences Upper Austria, AT), and Veit Schwämmle (University of Southern Denmark – Odense, DK)

License © Creative Commons BY 4.0 International license
© Magnus Palmblad, Viktoria Dorfer, and Veit Schwämmle

This Working Group was convened as a spin-out of the Machine Learning in Proteomics Working Group, and focused specifically on the creation of machine learning competitions (as inspired by the Kaggle format) built around proteomics data. The underlying idea being that this will help enlist interest and innovation from the broader machine learning community.

In order to attract the broader machine learning community, deep learning challenges in proteomics should be sufficiently simplified. We therefore discussed in detail two use cases that look into specific problems in peptide mass spectrometry: the prediction of peptide observability (a challenge we nicknamed “SuperPeptide”), and the prediction of the triggering isotope from a fragmentation spectrum (a challenge we nicknamed: “Where did you hit me?”).

These two challenges were devised to be posted on platforms such as Kaggle, and can furthermore be advertised throughout the proteomics community *via* organisations such as the European Proteomics Association (EuPA), the Human Proteome Organisation (HUPO), the European Bioinformatics Community (EuBIC), the International Society for Computational Biology (ISCB), and the Association of Biomolecular Resource Facilities (ABRF).

4.4 Working Group Report: Structural and Top-Down Proteomics

Daniel Questschlich (University of Oxford, GB), Bernard Delanghe (Thermo Fisher GmbH – Bremen, DE), Viktoria Dorfer (University of Applied Sciences Upper Austria, AT), Patrick Emery (Matrix Science Ltd. – London, GB), Michael Hoopmann (Institute for Systems Biology – Seattle, US), Lukas Käll (KTH Royal Institute of Technology – Solna, SE), Neil Kelleher (Northwestern University – Evanston, US), Magnus Palmblad (Leiden University Medical Center, NL), Veit Schwämmle (University of Southern Denmark – Odense, DK), Matthew Smith (University of Texas – Austin, US), and Mathias Wilhelm (TU München – Freising, DE)

License © Creative Commons BY 4.0 International license
© Daniel Questschlich, Bernard Delanghe, Viktoria Dorfer, Patrick Emery, Michael Hoopmann, Lukas Käll, Neil Kelleher, Magnus Palmblad, Veit Schwämmle, Matthew Smith, and Mathias Wilhelm

This abstract summarises the progress made by the Structural and Top-Down Proteomics Working Group over the course of the entire seminar.

Early discussions in this working group focussed on how the different structural mass spectrometry techniques can be integrated with one another, but also more broadly with efforts in the wider structural biology community. In addition, needs for data formats and standardisation for cross-linking mass spectrometry and native mass spectrometry were examined. Moreover, the working group also set out to engage with the Machine Learning in Proteomics Working Group to delineate topics of mutual interest in cross-linking mass spectrometry.

A key discussion point that emerged from this overview, was the overarching theme of how the structural proteomics community should engage with the wider structural biology community. The discussion focussed primarily on cross-linking mass spectrometry and native top-down mass spectrometry strategies. One potential strategy that was explored was to join the Critical Assessment of protein Structure Prediction (CASP) experiments.

Another topic of importance to the structural and top-down proteomics communities relates to the detection of post-translational modifications, and their annotation on existing protein structures. Here, there are specific challenges as well, most notably the issue of having to distinguish between functional and bystander modifications, as both are readily observed in mass spectrometry. Another relevant issue is the determination of the stoichiometry of these modifications across proteoforms. There is also the specific case of proteins with two different conformers that are regulated by complexation or post-translational modifications. Such cases could be interesting targets for computational inference from a combination of native mass spectrometry and cross-linking mass spectrometry.

Software needs and computational challenges in native mass spectrometry and native top-down mass spectrometry were discussed in more detail. The first main topic related to the modes of software dissemination. Different, non-exclusive scenarios exist today, ranging from open-source packages over freeware tools, to for-profit software as released by small to

large companies. Specific mention was also made of the need to document available software well, and to provide adequate training opportunities and materials for end users to ensure uptake and proper use.

A delineation of similarities and differences in the acquired data and the analysis approaches employed was made between native top-down mass spectrometry and traditional top-down proteomics. The use of a combination of different types of mass spectrometry analysis for validation was explored as well. One option is to combine data from traditional bottom-up approaches (for instance, affinity purification mass spectrometry or even standard shotgun mass spectrometry), with data from cross-linking experiments, and furthermore add in native (top-down) mass spectrometry data. Conceivably, hydrogen-deuterium exchange mass spectrometry data could be included here too.

Starting points were also formulated for the standardization of data reporting for native mass spectrometry. These standards would need to take the form of standardized data formats, minimal reporting requirements, and relevant terminology in existing or bespoke controlled vocabularies. The Human Proteome Organisation's Proteomics Standards Initiative (HUPO-PSI) creates community standards in the field, but currently lacks strong representation from the native mass spectrometry community. It will therefore be important to motivate more researchers in this community to engage actively in such standardisation efforts. A related aspect is the ability to publicly disseminate native mass spectrometry data, which will require compatibility with proteomics repositories such as PRIDE/ProteomeXchange. This was followed by a lively discussion of what data will need to be recorded to allow the move from proteoform analysis to complexoform analysis.

A final topic of discussion centered on ways in which data transfer and integration from structural proteomics experiments into protein knowledgebases like UniprotKB can be optimized.

Participants

- Sebastian Böcker
Universität Jena, DE
- Viktoría Dorfer
University of Applied Sciences
Upper Austria, AT
- Lukas Käll
KTH Royal Institute of
Technology – Solna, SE
- Frédérique Lisacek
Swiss Institute of Bioinformatics –
Genève, CH
- Lennart Martens
Ghent University, BE
- Magnus Palmblad
Leiden University Medical
Center, NL
- Robin Park
Bruker – Rancho Santa Fe, US
- Daniel Questschlich
University of Oxford, GB
- Veit Schwämmle
University of Southern Denmark –
Odense, DK
- Mathias Wilhelm
TU München – Freising, DE

Remote Participants

- Jeffrey Agar
Northeastern University –
Boston, US
- Kiyoko Aoki-Kinoshita
Soka University – Tokyo, JP
- Marshall Bern
Protein Metrics – Cupertino, US
- Robert Chalkley
University of California –
San Francisco, US
- Sven Degrove
Ghent University, BE
- Bernard Delanghe
Thermo Fisher GmbH –
Bremen, DE
- Patrick Emery
Matrix Science Ltd. –
London, GB
- Rebekah Gundry
University of Nebraska –
UOmaha, US
- Michael Hoopmann
Institute for Systems Biology –
Seattle, US
- Neil Kelleher
Northwestern University –
Evanston, US
- Joanna Kirkpatrick
The Francis Crick Institute –
London, GB
- Daniel Kolarich
Griffith University –
Southport, AU
- Rune Linding
HU Berlin, DE
- Nicki Packer
Macquarie University –
Sydney, AU
- Matthew Smith
University of Texas – Austin, US
- Sabarinath Peruvemba
Subramanian
University of Nebraska –
Omaha, US
- Morten Thaysen-Andersen
Macquarie University –
Sydney, AU
- Lilla Turiák
Research Centre for Natural
Sciences – Budapest, HU
- Olga Vitek
Northeastern University –
Boston, US
- Christine Vogel
New York University, US



