

Identifying and quantifying the unknown: Two workflows combined to search for variant/mutation information and to quantify un-/identified spectra in MS-based proteomics.

Dominik Lux^{1,2}, Prof. Dr. Julian Uszkoreit³, Prof. Dr. Brit Mollenhauer^{4,5}, Dr. Katalin Barkovits-Boeddinghaus^{1,2}, Prof. Dr. Martin Eisenacher^{1,2}, Prof. Dr. Katrin Marcus-Alic^{1,2}

Abstract

Introduction: Today's data analysis in MS-based proteomics primarily relies on typical identification and quantification workflows. These workflows usually utilize a FASTA database where mostly only canonical sequences of proteins are provided. In such workflows a theoretical spectrum is matched to an actual measured spectrum (search engine) returning identified spectra, which further are used to quantify the found peptide/protein. While these workflows usually suffice for the selected use-cases, not all spectra are identified and only a part of all measured molecules in a mass spectrometer are quantified, thus not reported in the final data analysis. Our workflows aim to shed some light on the unknown, trying to identify non-canonical sequences and by quantifying charged molecules, regardless of whether they are identified or not.

Methods: In our lab, we developed two workflows: a sophisticated FASTA generator (1) and a MS1-based quantification workflow (2).

- (1) generates a peptide FASTA databases containing already digested peptides, complemented with signal-, pro- and other specially cleaved peptides as well as peptides containing variants and mutations. This is achieved by using protein graphs generated via ProtGraph, the MS2 precursors and a sophisticated traversal algorithm implemented in C++ to retrieve the peptide entries.
- (2) utilizes OpenMS to find features (FeatureFinderCentroided). Label-free matching across multiple measurements is done via MapAlignerTreeGuided and FeatureLinkerUnlabeled, using identified features as anchor points. The actual quantitative values for all features is then extracted via the ThermoRawFileParser XIC-Extraction.

These two workflows are implemented in Nextflow, with various Python scripts used as intermediate steps. Identification in this combined workflow is done via Comet and Percolator with a q-value cut-off of 1%, using the human proteome database as plain text format, and allowing for up to 5 variants per peptide.

Results: We applied this combined workflow on measured CSF samples (two groups, DDA). Our results demonstrate the benefit of using a custom tailored FASTA database from workflow (1), which includes entries typically not searched for, thereby increasing the number of identified spectra in the CSF dataset. Workflow (2) shows that charged molecules can be quantified, regardless of whether an identification or even a MS2 spectrum is present. By combining and normalizing the results of each workflow, we generated a volcano plot, containing unidentified and non-canonical entries alongside of usual canonical entries, illustrating the additional data points gained by this combined workflow.

Motivation and Context

In MS-based DDA proteomics, the usual focus mostly lies on identified and quantified peptides, or proteins after inferencing. Many charged molecules are usually not used in further analysis, simply by leaving them out in common data analyses. **Figure 1** illustrates the information, which gets lost in such analyses and shows, where our work focuses on to reduce the lost information.

Figure 1: Illustration of the information which is gathered and used (proportions not correct). While most data analyses focus on the smallest subset, it also shows that there is still a lot of potential to get even more information from such MS-measurements. Our work (black arrows) focuses on increasing the identification ratio as well as using MS1-features disregarding whether there is an identification or not.

Workflow and Implementation

We first focused on increasing the identification ratio in MS2 spectra and have already developed a workflow with **ProtGraph** which is able to increase it by utilizing non-canonical sequences (**Figure 2**) in some datasets [1][2]. This workflow has been further extended with our experimental quantitative workflow (unbeQuant), which uses the OpenMS-framework to find features and the ThermoRawFileParser to retrieve the corresponding XICs/quantitative values. The final combined workflow is shown in **Figure 3**. All steps are implemented in such a way, that they can be integrated in other workflows for other use cases or to be used on their own. For Additional information and implementation details, visit our [GitHub](#)!

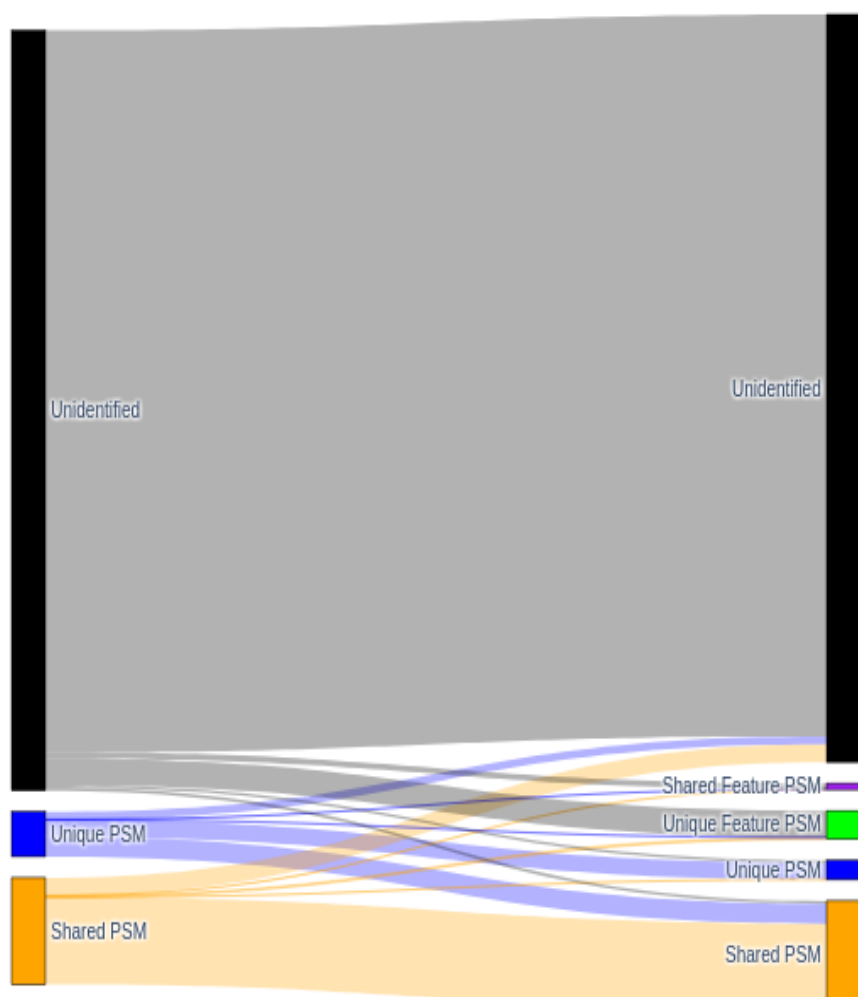


Figure 2: Difference of identified MS2 spectra across a identification with only canonical sequences in a database or supplemented with non-canonical (feature) sequences, categorized MS2 spectra by: unidentified (black), identified with a unique/shared peptide (blue/orange) as well as with non-canonical unique/shared peptides (green/purple). This result originates from a blood sample study (PRIDE: PXD028605).

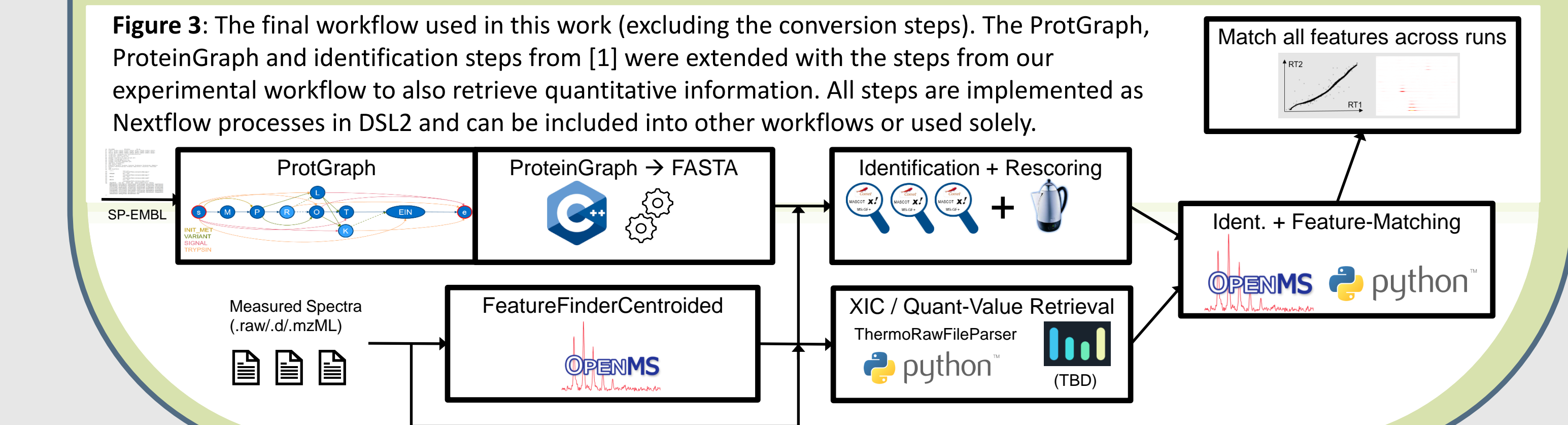


Figure 3: The final workflow used in this work (excluding the conversion steps). The ProtGraph, ProteinGraph and identification steps from [1] were extended with the steps from our experimental workflow to also retrieve quantitative information. All steps are implemented as Nextflow processes in DSL2 and can be included into other workflows or used solely.

Results

Identifying with the custom generated FASTA, containing features like cleaved peptides, variants and more, yielded an increase of the overall identification ratio of **0.01%** (around **3000** more spectra with an identification) after the FDR cutoff. Although many MS2 spectra were annotated with a feature peptide or received a better peptide candidate (around **60000** / **5%** of all identified MS2 spectra), the large FASTA database (33Gb, containing 55 million peptides) causes many entries to be shifted below the FDR-threshold of 0.01%. In **Figure 4**, we illustrate certain found features, by their count, showing that variants and conflicts have mostly been found. **Figure 5** shows the 15 proteins with the highest amount of features from found unique feature peptides. Mostly immunoglobulin proteins have been found with a conflict feature. Serotransferrin, Transthyretin, Alpha-1-antitrypsin are also present with many found variants. We also discovered the protein Secretogranin-1, and its neuropeptides: PE-11 and CBB.

Figure 4: Counted feature peptides in the CSF dataset from PSMs, depending whether it is a shared or unique PSM. This overview only contains certain hits and may count multiple features for a single PSM.

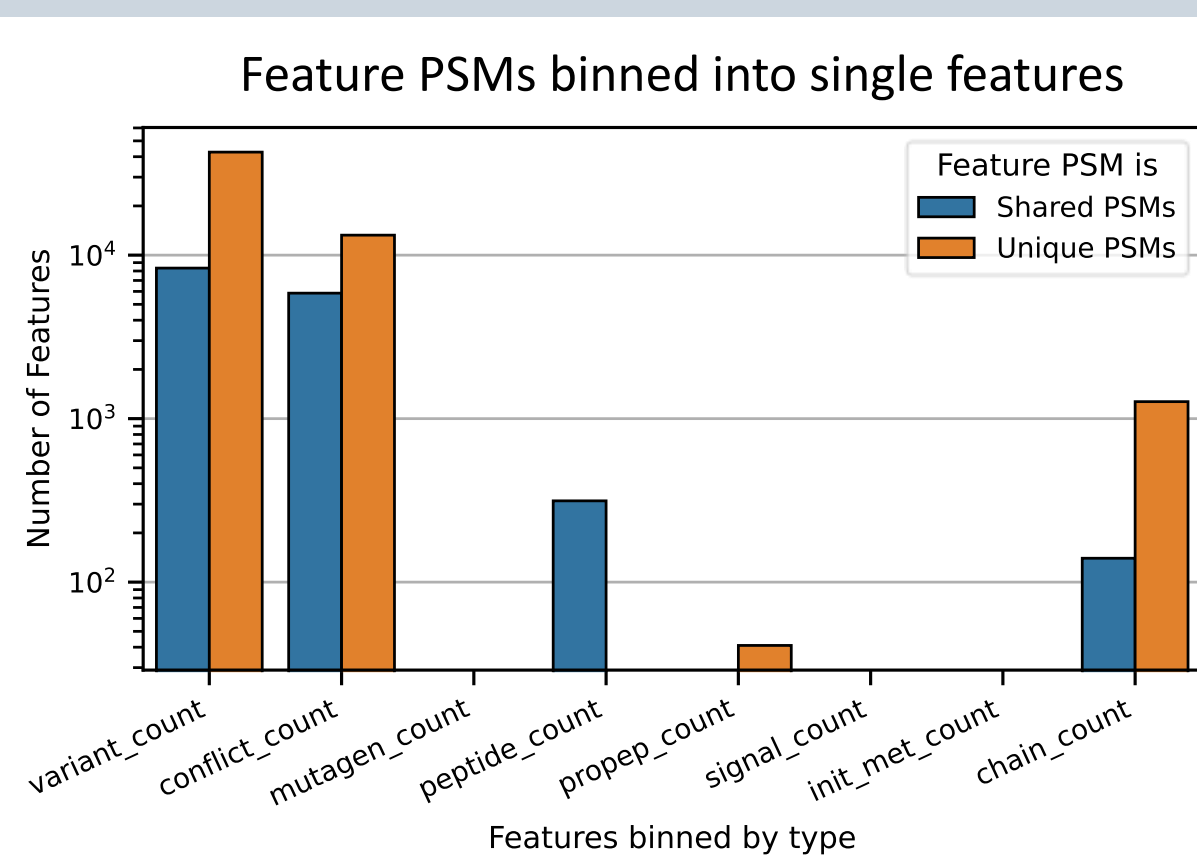


Figure 4: Counted feature peptides in the CSF dataset from PSMs, depending whether it is a shared or unique PSM. This overview only contains certain hits and may count multiple features for a single PSM.

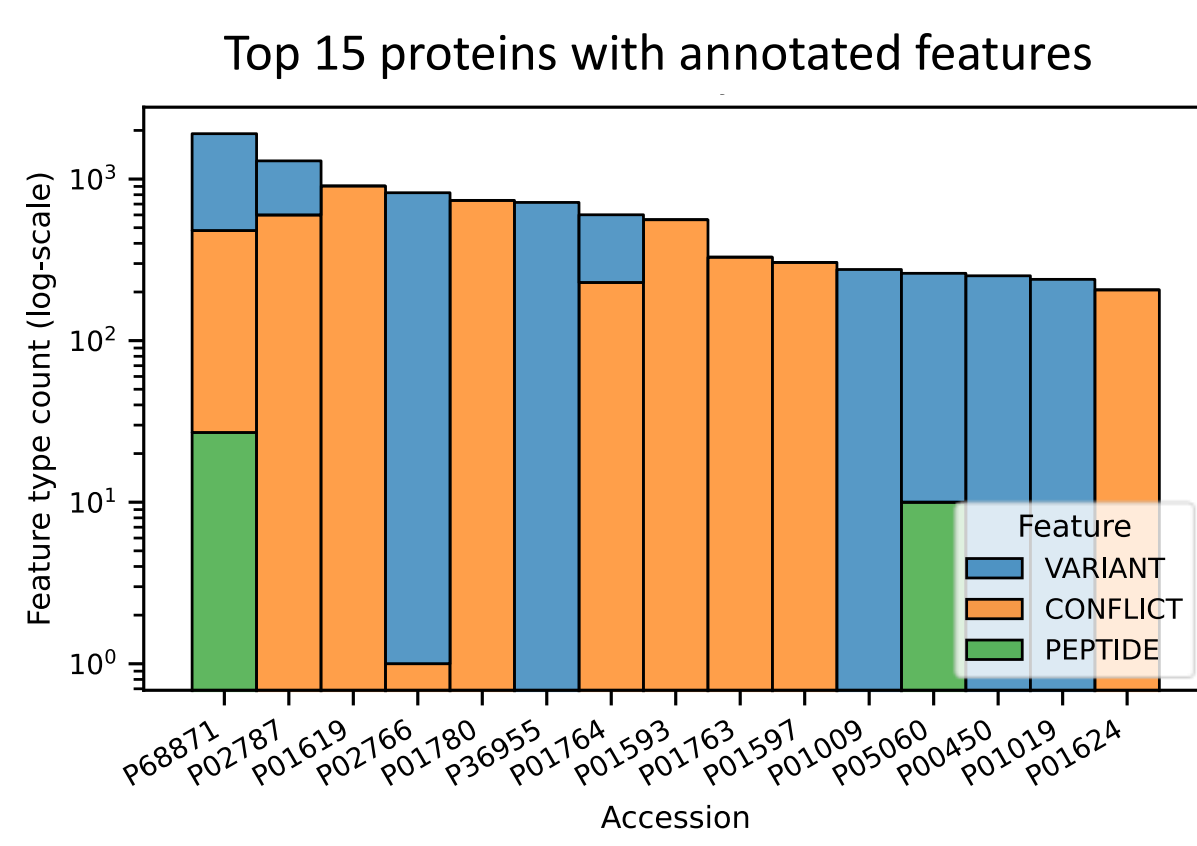


Figure 5: Top 15 proteins with annotated features, derived from unique PSMs. HBB has the most feature annotations, which is expected, since it is highly annotated in UniProt KB. Many Immunoglobulin proteins have been found with conflict features. Conflict features describe differences originating from sequence errors or of currently not described sequence polymorphisms. Proteins like Serotransferrin which are expected to be found in CSF have also been found, topped with many annotated variant features.

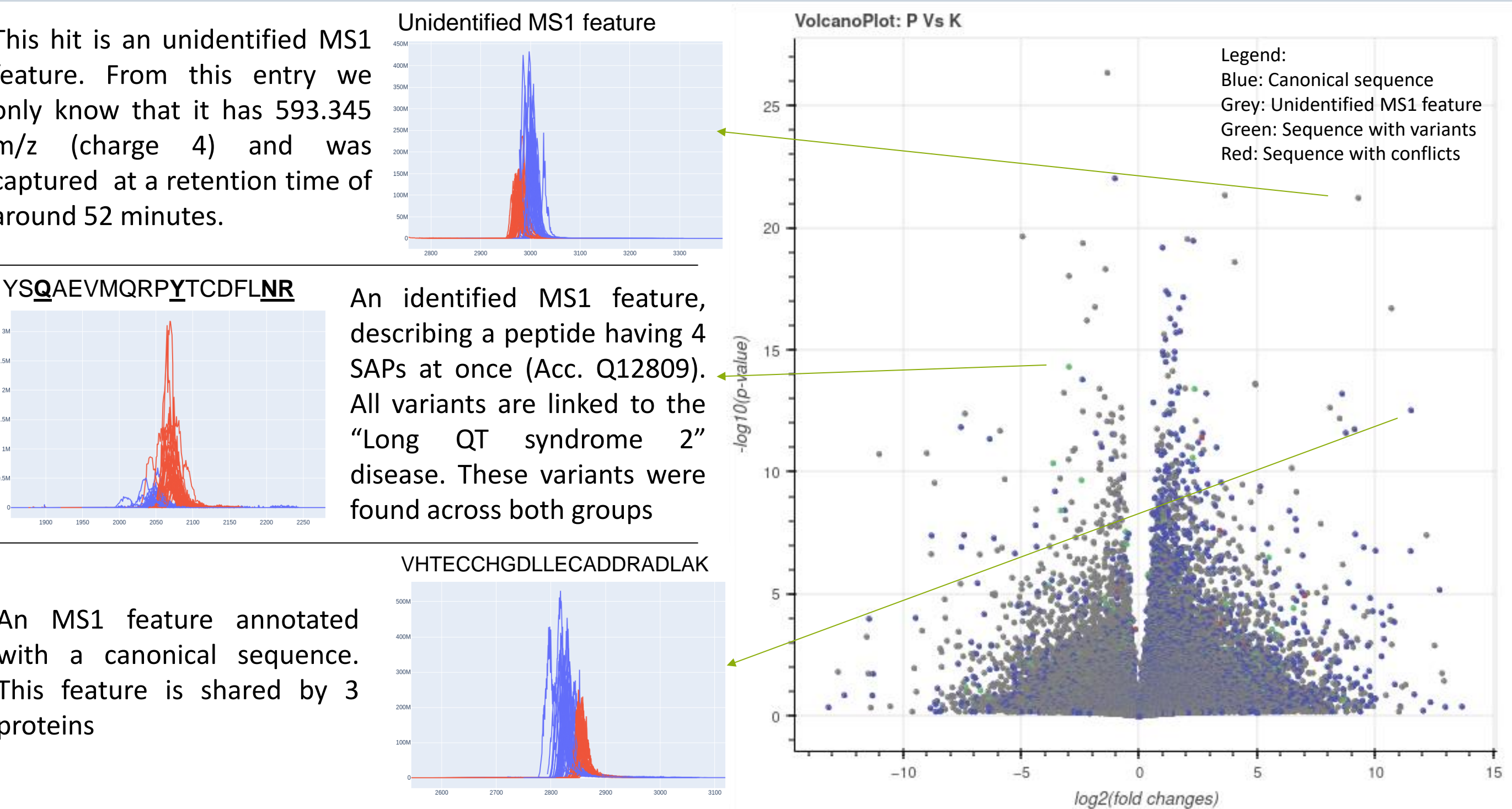


Figure 6: Volcano plot from the results of the showcased combined workflow. The volcano plot is additionally colored, where blue points describe MS1 features, containing a canonical identification, grey points describe MS1 features without identification and colored points MS1 features annotated with a variant (green) or conflict (red) feature. On the left, selected points are shown, with raw extracted XICs across all measurements and a small description of the information available on this entry.

Besides the identification results, our workflow returns MS1 features with an identification from an MS2 spectrum if present. We used all MS1 features, normalized it and used a t-test, creating a volcano plot to show most interesting hits between the healthy and diseased group, while also coloring the available identification information (**Figure 6**). The interpretation of the results is still ongoing. However, with this additional knowledge, a much more in-depth look into the CSF-dataset is already possible.

Future Work

Although the combined workflow (ProtGraph + unbeQuant) was applied on the CSF dataset, we already applied this workflow on multiple other datasets including cell models, blood, tissue and some initial tests with stool samples. We aim to provide both as general workflows applicable to a variety of DDA datasets.

Furthermore we noticed that still many MS1 features across multiple runs are not present (missing values). Due to the high amount of missing values, a p-value may not be calculated and therefore may simply not be shown in a final volcano plot. The not shown entries could be potential “black and white” entries and we also want to further explore ideas to visualize and rank them in a meaningful way. In **Figure 7**, we experimented with a view, using the ratio of missing data and the mean intensity of all (non missing features).

In future, we want to provide an explorative data view suite as a result provided by both workflows, where the results can be explored interactively, yielding information about the workflow and raw data in a meaningful way. First work has already started in this regard: a XIC extractor has been implemented which can generate XICs from the entries provided in **Figure 6** and **Figure 7** as also shown in them.

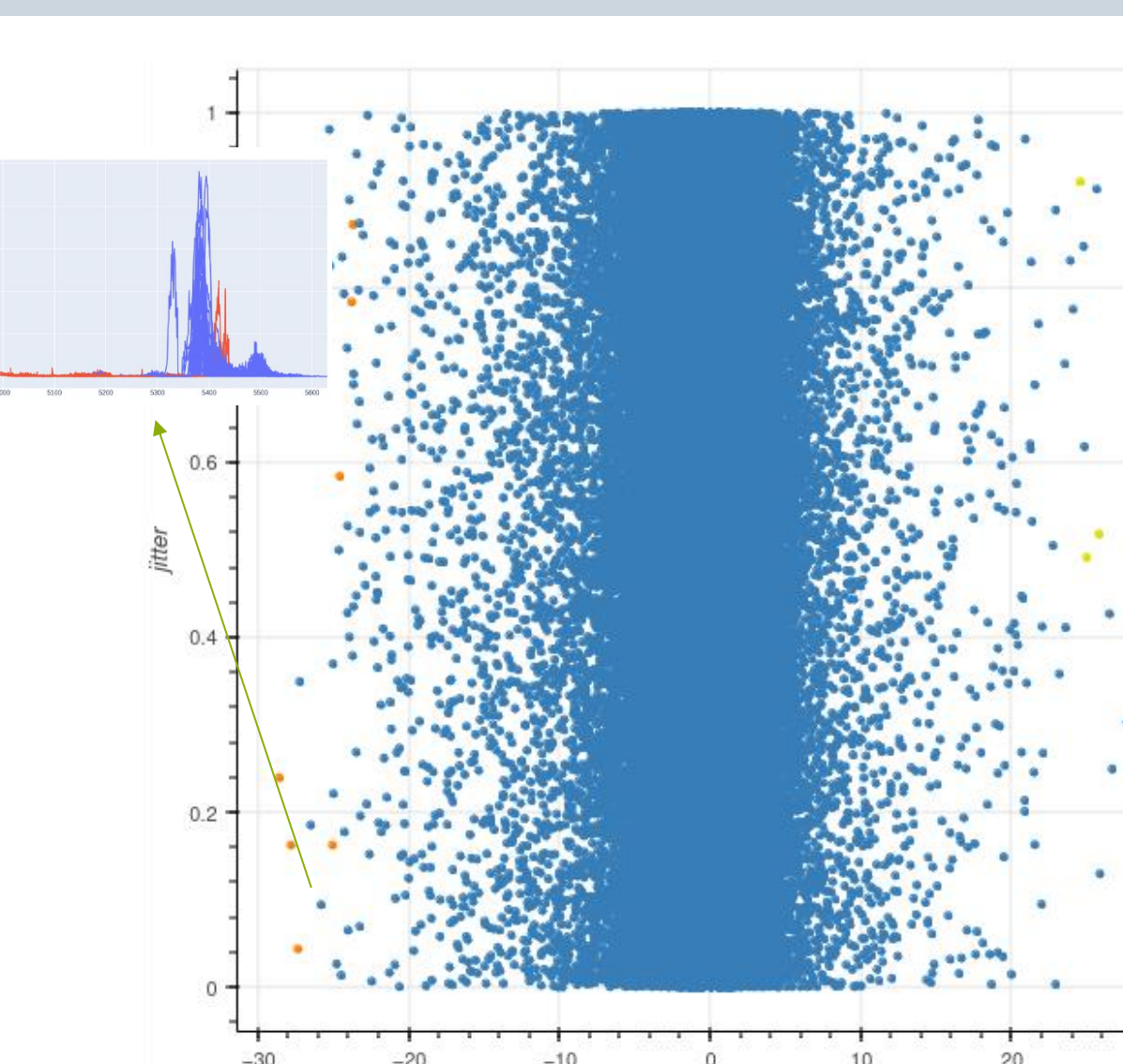


Figure 7: Plot containing all entries, which did not receive a p-value. On the x-axis, we calculated a negative or positive value depending on the quantitative ratio and for each group the ratio of missing values which was multiplied by the mean of the normalized intensity. The Y-axis contains a jitter to spread the data points.