**Searching for novel Sulfotyrosines (sY) in a hapYstack**

# Abstract

Table of contents

[Abstract 1](#_Toc163025738)

[1. Introduction 3](#_Toc163025739)

[1.1 What is the research challenge? 3](#_Toc163025740)

[1.2 ppm vs Dalton accuracy 4](#_Toc163025741)

[1.3 Difficulties with finding positive evidence for sulfo 4](#_Toc163025742)

[1.4 Data we can use for analysis 5](#_Toc163025743)

[2. Methods 6](#_Toc163025744)

[2.1 Data collation and aggregation by peptidoform. 7](#_Toc163025745)

[2.2 Gaussian Mixture Model (GMM) fitting 8](#_Toc163025746)

[2.3 Peptidoform assignment to mass error bins 9](#_Toc163025747)

[2.4 Data exploration for known and novel sulfotyrosines sites 10](#_Toc163025748)

[2.4.1 GO term ORA 10](#_Toc163025749)

[2.4.2 Custom terms ORA 10](#_Toc163025750)

[2.4.3 Peptidoform annotation 10](#_Toc163025751)

[2.4.4 Instrument contributions exploration 10](#_Toc163025752)

[2.4.5 MS2 evidence assessment 10](#_Toc163025753)

[3. Results 10](#_Toc163025754)

**Table of Figures**

[Fig. 1. Study design and workflow. 5](#_Toc162938165)

[Fig. 2. Peptidoform proportions across calibrated mass error bins. 8](#_Toc162938166)

# Introduction

**From Andy’s google sheet explaining background of project briefly:**

I want to see if it possible to identify sulfated peptides, based on their molecular weight (MW) profile, identified from vast amounts of tandem mass spectrometry (MS/MS) data. Fortunately for us, the data searching has already been done (I had initially planned to do lots of searches, but I think we can avoid it).

## What is the research challenge?

* Sulfation has a mass of 79.956815 Daltons <http://www.unimod.org/modifications_view.php?editid1=40>
* Phosphorylation has a mass of 79.966331 Da: <http://www.unimod.org/modifications_view.php?editid1=21>
* Mass difference = 0.009516 Da

## ppm vs Dalton accuracy

The masses of peptides with phospho or sulfo are thus very similar, and phospho is much-much more common to find. The latest MS instruments can be accurate to around 2 ppm (parts-per-million) for measuring peptidoform mass. ppm is a relative unit of measurement, it is easily understood by comparing to percentage, which is parts-per-hundred. If we measure a molecule with mass 1601.132 Daltons, with an accuracy of +/- 1%, it is mathematically straightforward to understand this. ppm is the same but with a denominator of 1 million instead of 1 hundred.

This means that the accuracy we can measure depends on the mass of a molecule. We often set our software to allow for candidates to be +/- 10ppm, so this gives rise to something like the following windows we look for:

|  |  |  |
| --- | --- | --- |
| Mass (Da) | Upper | Lower |
| 1000 | 1000.01 | 999.99 |
| 2000 | 2000.02 | 1999.98 |
| 3000 | 3000.03 | 2999.97 |

A typical peptidoform we might observe is probably say v. aprox 2000-4000 Daltons. If we set our search software to look for peptide + mass of phospho with a +/- 10ppm window, then *if it had actually be sulfated*, it would be identified positively (but wrongly, as a phosphopeptide rather than a sulfopeptide).

## Difficulties with finding positive evidence for sulfo

It is very common in proteomics to enrich for phosphopeptides, using immobilized metal ion affinity chromatography (IMAC) e.g. TiO2 (titanium dioxide). Alternative metals can also be used e.g. Fe, Zi,

It is much argued over, but we think there is evidence for around 80K human phosphosites: <https://pubs.acs.org/doi/full/10.1021/acs.jproteome.2c00131> in current databases. Mostly these are on serines, then threonines, and fewer on tyrosines.

Sulfation is much more rare, UniProt only has evidence for 50 proteins with sulfotyrosine sites: <https://www.uniprot.org/uniprotkb?query=%28proteome%3AUP000005640%29+AND+%28ft_mod_res%3Asulfotyrosine%29>

Pat and Claire Eyers won a BBSRC grant to identify lots of sulfotyrosine sites - <https://gtr.ukri.org/projects?ref=BB%2FS018514%2F1>

This has proved tough going. Leonard Daly has worked on this and found it really difficult i) to enrich for sulfopeptides instead of phosphopeptides, and ii) when you fragment a peptide by MS/MS, the sulfation side change mostly falls off, so you cannot prove that it was present on any particular amino acid in the sequence.

They recently published a protocol where they had limited success identifying sulfation sites, with a very specialised protocol: <https://pubs.acs.org/doi/10.1021/acs.jproteome.3c00425>

Sulfated proteins are usually secreted, but unknown if only on secreted proteins, and seem to be best enriched with Zirconium.

## Data we can use for analysis

Our collaborator Eric Deutsch at ISB Seattle creates large PTM builds - like human phospho 2022 build: <https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/buildDetails?atlas_build_id=537>

He has searched a huge number of public ProteomeXchange Database (PXD) MS/MS datasets that were enriched for phosphopeptides (although very few, if any, using Zirconium, and also few secretomes). So we are really looking for a needle in a haystack, and perhaps the paper will “prove” finally that you cannot identify sulfopeptides from general phospho-enriched data sets - this would be completely fine, and publishable.

The publicly available data files do not have sufficiently granular information, so Eric will package them up and send them to us in zip files - he has sent two PXD builds already in zipped pepXML format.

# Methods

An overview of the analytical pipeline is presented in Fig. 1.

A screenshot of a document

Description automatically generated

Fig. 1. Study design and workflow.

## Data collation and aggregation by peptidoform.

This study employed the largest publicly available human phosphoprotein tandem mass spectrometry build on Peptide Atlas - Human Phosphoproteome 2022-04 (**Fig. 1A**). The build includes data from 129 MS/MS datasets spanning 365 experiments for a total of 19,852 MS runs, which have contributed to the identification of 264,244 peptides derived from 9,810 canonical proteins. This build was selected due to its enrichment in phosphopeptides, which may include sulfopeptides misidentified as phosphopeptides. The spectral data were extracted and converted from ‘.pep.xml’ to ‘.tsv’ format in Python. False discovery rate thresholding was applied based on PeptideProphet q < 0.01 followed by recalibration of the thresholded data and subsequent filtering to only include peptides with assigned phosphoserine (pS), phosphothreonine (pT), and phosphotyrosine (pY) post-translational modifications (PTMs).

Based on the peptide sequence, and type and number of PTMs assigned to it, a non-strict peptidoform ID was generated for every peptide-spectrum match (PSM), as illustrated in **Fig. 1B**. This resulted in 1,747,947 unique phosphorylated peptidoform IDs across the build. Finally, the recalibrated mass error data across all datasets was aggregated by peptidoform ID.

## Gaussian Mixture Model (GMM) fitting

Prior to GMM fitting data were filtered to include only robustly detected peptidoforms across the build by taking forward peptidoforms detected in 3 or more experiments (n = 413,010 peptidoform IDs). To ensure sufficient data is used to confidently model the calibrated error gaussian mixtures, peptidoform IDs were further filtered to include IDs with 90 or more PSMs, resulting in 83,047 peptidoform IDs taken forward for analysis.

GMMs with 1, 2, or 3 components were fit on the recalibrated mass error PSM data for each peptidoform. The reason for including up to three components were to explore the following scenarios in search of sulfation misidentified as phosphorylation events:

* + - One component: Peptidoforms that are correctly identified as phosphorylated would have a normal distribution cantered around a 0 Da shift. The mass error for peptidoforms that are always singly sulfated and misidentified as phosphorylated would be expected to follow a normal distribution with median falling around a -0.0095 Da shift. Similarly, doubly sulfated peptidoforms (if any) would exhibit a Da shift at ~ -0.019 Da.
    - Two components: Two gaussians would be an optimal fit for scenarios where the peptide is sometimes phosphorylated and sometimes sulfated (both 0 Da and -0.0095 Da medians present). Alternatively two components would also cover scenarios where a peptidoform can be doubly or singly sulfated and doubly or not sulfated.
    - Three components: Models with three components would cover rare scenarios where single, double, or no sulfation are possible for the same peptidoform or any of the previous scenarios mixed with random misidentifications of other PTMs causing a different shift in mass.

The best fitting number of components was individually selected for each peptidoform based on Bayesian Information Criterion (BIC) scores (lower scores are better). To avoid overfitting the data, the smallest number of components were prioritised unless the BIC score of a model with a larger number of components was 10 or more points smaller.

## Peptidoform assignment to mass error bins

Based on the scenarios described in Section 2.2, 10 mass error bins were pre-defined spanning -0.0275 Da to 0.0275 Da. The width of each bin was set to 0.005 Da, with singly sulfated peptidoforms expected to have at least one gaussian component’s mean fall in the -0.0125 to -0.0075 Da shift bin. For every peptidoform ID, the areas under the curve (AUCs) for the best fit GMM was computed based on the GMM’s probability density functions. Peptidoforms were assigned to mass error bins if the sum of all AUCs for that peptidoform’s GMM over that bin was >= 15%. These settings ensured less stringent assignment to bins following the very stringent data filtering up to this point.

## Data exploration for known and potential novel sulfotyrosines sites

### GO term ORA

### Custom terms ORA

### Peptidoform annotation

### Instrument contributions exploration

### MS2 evidence assessment

# Results

* + - Hypothesis is that having large enough amount of data we could pick up sulfation events based on mass error shift alone.

## GMM AUC filtering highlights potentially sulfated peptidoforms.

Fig 2. – we expect sulfation to affect tyrosines but not threonines and serines.

* + - Therefore, we expect enrichment in sY in a bin to result in an increased proportion of peptides that have been assigned to contain a pY to account for misidentified sYs. By association, the fraction of peptides that contain a Y might be increased too, while S-containing and T-containing proportions shouldn’t be as strongly affected (they serve as negative controls of a sort)
    - A, B, C : We look at the proportion of all peptides that fall in a bin that have a tyrosine, threonine, or serine. The shift in proportions here is strongest in the bins to the left of the bin of interest (BOI) where we expect single sulfation events. This may have to do with our less stringent AUC filtering (15% is relatively low so we may be seeing some false negatives slip through, while true sulfated peptides might be picked up in the bins to the left; bins to the left would also include doubly sulfated peptidoforms). Note here that we have a relatively small number of peptidoforms in the bins to the left so the proportions are a bit more prone to random events.
    - D, E, F – we look at the proportion of peptides that have one of the above phosphorylated. pT and pY proportions differ a lot in bins to the left of BOI; once again -we do have a small number of peptidoforms in each of these bins but it does seem like pT assignment might be happening in peptides where sY is present and the PTM is assigned to the wrong amino acid. Promising that BOI has more than double the proportion of pY compared to the bins with the majority of the peptidoforms. Note we’re seeing the higher proportions in the DECOY bins, might be interesting to discuss why.

A screenshot of a graph

Description automatically generated

Fig. 2. Peptidoform proportions across calibrated mass error bins.

LEGEND PLACEHOLDER

All proportions are given as a fraction of the total number of peptidoforms that fall within each bin. Orange denotes the bin where we expect to detect single sulfation events, blue denotes DECOY bins where we expect to detect no sulfation.

The total numbers were as follows, from left to right: 12; 27; 52; 443; 34,495; 81,629; 35,625; 1,333; 642; 633

A. Serine-containing peptidoforms; B Threonine-containing

C. Tyrosine-containing

D. Peptidoforms with assigned phosphoserine

E. Peptidoforms with assigned phosphothreonine

F. Peptidoforms with assigned phosphotyrosine

**Conclude: Some evidence we may be seeing sYs but not very strong at this point. Look deeper into BOI and bins to the left and manually inspect the Mass error histograms to see if they look convincing (Fig. 3).**

Fig. 3 – example histograms in BOI; needs some rearranging

* + - Here we have examples of histograms that are False positives (A), known sY-containing peptidoforms (C, F), and potential novel sY-containing peptidoforms (B, D, E). We also have examples of singly sulfated, doubly sulfated, singly sulfated or phosphorylated mix. Discuss some of the interesting case studies in a little bit of detail reporting number of peptidoforms for some of the more commonly detected protein IDs in that bin. Point to Table 1 for all of the ones that seem interesting.

A screenshot of a graph

Description automatically generated

Fig. 3. Example histograms of peptidoforms with AUC > 15% in BOI

**A)** Ribosomal protein S6 kinase alpha-3; False positive peptidoform example

**B)** Sulfotransferase 2A1, Sulfated only example; 1 peptidoform for this protein; potential novel info on sulfation; **C)** Vitronectin, known sY protein: Doubly or singly sulfated peptide; total of 10 peptidoforms of Vitronectin found in BOI and bin -1; 2 peptides (and incl shortened versions of 1), both of which contain known sY residues; **D)** Osteopontin – secreted; Mostly sulfo, some phospho PSMs; 5 peptidoforms derived from 2 peptides; potential novel sulfation

**E) Calumenin – potential novel sulfation; There is a few novel ones not on here. Might be better to add but fist see if any of the MS2 are convincing?**

* **Potentially split into known sY example histograms and potential novel sY histograms?**

**To Add : nASEEEPEYGEEIK\_n230\_1\_S167\_1\_Y243\_1**

**Secretogranin, known sY**

**TABLE 1 – likely sY-containing peptidoforms (based on histogram AND protein cellular localisaton**

PLACEGOLDER; table I shared with Eric but formatted a bit better.

## Enrichment analyses

**We expected to see an enrichment in terms linked to cytokines, secreted proteins, or transmembrane proteins as these are known to be sulfated.**

GO-term ORA (Fig. 4)

* + - This figure is difficult to read, I have tried replotting but not enough space on a page really (that I could think of) – may need to convert to a table style figure similar to what Emily presented last Tuesday.
    - Take home is that some of the terms we see enriched in BOI and bin to the left (BOI-M1) make sense in the context of sY (sulfur compound binding, extracellular, integrin binding, and Golgi-related terms) but we have very few genes in the analysis. In the DECOY bins we see some enrichment results but they are not linked to sulfation – might be interesting why we see these terms in the + 0.01 Da shift region.

A screenshot of a graph

Description automatically generated

Fig. 4. Go Term ORA results by bin

PLACEHOLDER

* + - Because sY is less studied and not strictly associated with many GO terms, we explore the use of custom terms designed to pick up potential sY events (transmembrane proteins; proteins known to contain a sY; secreted prtoeins; Golgi-related proteins) and a term to serve as negative control (all other proteins that are not linked to any of these terms. (Fig. 5)
    - Of the known sY-containing proteins in UniProt (~51), across all bins only 6 are represented by peptidoforms that have passed all filtering stages. Nevertheless, there is a significant enrichment of known sY protein IDs in BOI and BOI-M1; Important to note we have quite a few peptidoforms for Vitronectin, one of the 2 known sY proteins in BOI. No peptidoforms for these fall in the DECOY bins.
    - Secreted proteins are strongly enriched in BOI-M1 – perhaps two sulfation events are more common in secreted proteins, and these are thus more enriched in the bin to the left of BOI.
    - Although not significant, Golgi-related proteins were detected at ratios higher than the background levels (dashed line) and than the DECOY bins in both BOI and BOI-M1.
    - Transmembrane protein ratio was also higher in BOI-M1 but not significantly.
    - Proteins unlikely to contain a sY (basically all other proteins) were significantly enriched in the DECOY bins, which cover a mass erro shift in the direction opposite to that of expected sY.

Therefore, we can find known sY using this method, and we may be a ble to find potential novel sY sites. However, the number of proteins that have passed the stringent filtering applied required to find sY is quite small and thus sY misidentification as pY is likely a rare event and not a massive confusing factor for proteomic studies.

A screenshot of a graph

Description automatically generated

Fig. 5. Custom Term ORA results by bin

LEGEND PLACEHOLDER

**Make sure to mention actual number of genes in text in the context of number of peptides and number of peptidoforms too.**

**What about the other 4 known sY that we are detecting? How do their histograms look like? Are they sulfated portions of the peptide at all? \*\*\***

\*\*\* still need to look into last questions on here \*\*\*

## Case studies – can we find MS2 evidence in the PSMs associated with known or likely sY-containing peptidoforms?

* + - This part needs more work; My understanding is that because adding the sulfation sites explains slightly less of the total ion current than having the PTM-stripped USI, this is good evidence of true sulfo

EXAMPLE PICKED BY ANDY AND ERIC:

A screenshot of a computer

Description automatically generated

Fig. 6. Example MS2 spectra of convincing sY

## Additional point of interest – instrument contribution to sY discovery?

Do we need higher resolution instruments to pick apart sY and pY?

A screenshot of a computer screen

Description automatically generated

This shows number of PSMs contributed by the different instruments used across the build across:

Row 1: all data based on number of spectra id’d – from Eric’s metadata published with the phosphobuild

Row 2: all peptidoforms in bin of interest (BOI) (\*and to the left) where we expect to get singly and doubly sulfated peptidoforms

Row3: peptidoforms assigned as likely sY candidates base don calibrated error histogram and subcellular localisation or being known sY

Row4: known sY peptidoforms (Vitronectin and **Secretogranin; ~12 peptidoforms total, to double check!)**

Probably best as supplementary

Seems to suggest that a lot of the false positives in BOI may be coming from Q Exactive (not super surprising as it’s one of the older instruments on there)

TO DO: weave in some of the supplementary figures we have created (e.g. colour-coding histograms by PXD ID or experiment tag or instrument.

Can’t really draw massive conclusions here, could suggest the resolution as further avenue of exploration in future in discussion?