# 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 fitting and peptidoform assignment to bins.

Pre-filtering:

Detected in 3+ experiments: 413,010 IDs AND overall more than 90 PSMs: 83,047 # NB: 3+ experiment tags sometimes is the same experiment split into multiple subsets due to being a very large dataset.

Model fitting: Fit GMMs with 1, 2, or 3 components for the calibrated mass errors of each peptidoform ID

– reason for 1 is if the peptide can only be sulfated and not phosphorylated the shifted peaks would be expected to centre around -0.095

# reason why we look for two componenents - we expect some sulfation, so centering around -0.01 and 0.  
# reason for 3 - we have observed some data at + 0.01 in initial histograms

Best fit model selection: Prioritise fewer components unless BIC score of model with more components is 10+ points smaller (rule of thumb)

# select which model fits the data best depending on nuber of components, and BIC score  
# we want the lowest BIC score but also we don't want to overfit - we have seen sometimes  
# multiple components are selected based on BIC alone where eyeballing the histogram  
# would have resulted in a lower number of components. Therefore, we prioritise a low score  
# only if it's 'significantly' lower than the score of the smallest number of components with the best socre.  
# rule of thumb - 10 points lower score is significant, but we also test by plotting AUC later

Peptidoform assignment to m/z bins

## Data exploration for known and novel sulfotyrosines sites

### GO term ORA

### Custom terms ORA

### Peptidoform annotation

### Instrument contributions exploration

### MS2 evidence assessment

# Results

**A screenshot of a graph

Description automatically generated**

Fig. 1. Peptidoform proportions across bins.

LEGEND PLACEHOLDER

All proportions are given as a fraction of the total number of peptidoforms that fall within each bin. The total numbers were as follows, from left to right: mz -0.0275 to 0.0225: ;

A. Serine-containing peptidoforms

B

C

D

E

F

T

E

X

T

H

E

R

E

**A screenshot of a graph

Description automatically generated**

T

E

X

T

A screenshot of a graph

Description automatically generated

**Cytokines known to be sulfated – we do not see that in the GO terms BUT of the known sY-containing proteins in UniProt (~51), only 6 are represented by peptidoforms that have passed all filtering stages;**

**Custom terms for known and likely novel sY might be more suitable**

A screenshot of a graph

Description automatically generated

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

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)

MS2 assessment figure – top PSM for each case study protein?

Not sure how we’d present this? E.g. side by side spectra of added sulfo vs no sulfo to peptidoform USI? A screenshot of a computer

Description automatically generated