Liquid chromatography-tandem mass spectrometry (LC-MS/MS) is the most popular way to study the proteome. In bottom-up LC-MS/MS, proteins are digested into peptides and each peptide/amino acid sequence is fragmented along its peptide bonds, resulting in a unique series of mass-to-charge (m/z) and intensity values, otherwise known as a mass spectrum. After identification of peptides in a sample, proteins can be inferred[1]. While most proteomics search tools today try to match spectra to peptides by comparing them to a database of theoretical fragment m/z values and picking the database peptide with the most matched fragments, many do not assign fragment intensity values to the theoretical spectra. Because some fragments are more intense than others, excluding intensity from the theoretical spectra misses out on information that could be used for better discrimination of high confidence peptides.

Fragment intensity values are employed in an alternative peptide searching method to the traditional database search called “clustering.” As the name suggests, clustering algorithms calculate similarity (such as cosine similarity and dot product) between spectra to cluster them into groups, circumventing the use of “theoretical spectra.” These groups are iteratively merged with other groups up to a predefined similarity threshold, at which point each group is made up of one or a few peptide sequences[2,3]. Clustering speeds up database searches by combining redundant spectra, and it can summarize spectra from disparate experiments[3]. However, if we are tasked with finding the cluster most similar to a new spectrum, the spectrum must be compared to all other clusters with a similar peptide mass. To remedy this slow comparison step, other studies have vectorized spectra in n-dimensional m/z space and binned them to reduce the number of comparisons. A new spectrum can be vectorized and compared with neighbors in the same bin. Fast lookup to find which bin the spectrum belongs to can be performed with locality sensitive hashing[4] or binary index trees[5]. Despite the faster searching algorithms, these clustering methods are unsupervised, meaning they do not guarantee that peptides with high spectral similarity or those in the same bin are truly the same peptide, potentially leading to false identifications; homogeneity of clusters is determined after the fact.

Supervised approaches provide an attractive alternative to unsupervised clustering. Embedding through deep learning[6] (DL) is one such approach that has been adapted to proteomics. Rather than using stagnant peptide representations based on fragment m/z bins, embedding learns a new space by forcing points from the same peptides together while maximizing distance between different peptides, using a Siamese neural network[7]. Not only is embedding more specialized than regular clustering, but it is easy to embed a new peptide in the space by feeding it through the learned deep neural network. Peptide embedding is effective for summarizing and searching large public repositories of MS/MS spectra[8,9], but the learned embedded peptide space has yet to be fully explored. Our understanding of embedding of post-translational modifications (PTMs), chemical modifications on proteins that modulate their function, is particularly shallow. One good candidate PTM to study is phosphorylation, since there is an abundance of phospho-enriched studies, and because it plays a critical role in many diseases[10]. To investigate phosphorylated peptide embedding and to use it to our advantage, we propose the following aims:

Specific Aim 1: Determine where phosphorylated peptides are mapped relative to unmodified counterparts. Word embedders reveal that word vectors can be added in meaningful ways (king – man + woman = queen)[6]. Using pretrained embedding spaces[8,9] and a phosphorylation space we train ourselves, we will determine whether phosphorylated peptides are randomly or predictably distributed in space, perhaps correlated to different fragmentation patterns[11,12]. We will also see if PTMs adhere to vector addition rules, such as if methionine oxidation and phosphorylation “vectors” can be added to find a peptide with both modifications. Finally, if we identify patterns in the embedded space, then we will use a DL model to predict the spectra of a phosphorylated peptide, given the embedding of its unmodified counterpart.

Specific Aim 2: Distinguish phosphorylated from sulfated peptides. Phosphorylation and sulfation modify the same amino acid residues and have very similar masses, making them nearly impossible to distinguish under most circumstances. Using a Siamese neural network trained on data from ProteomeXchange, using unmodified-phosphorylated MS2 spectrum pairs as negative examples and unmodified-sulfated pairs as positive examples, we aim to distinguish true phosphorylation events from sulfation (“fraudulent phosphorylation”) events. We will compare our model to common similarity metrics (cosine similarity, etc.) to determine which is better at distinguishing between the two PTMs.

A. Background and motivation

The most popular and high-throughput method for proteome discovery is liquid chromatography-tandem mass spectrometry (LC-MS/MS). After digestion of the proteins in a sample, the resulting peptides are fragmented by a neutral gas in a collision cell, producing a unique series of fragment ions spanning the mass-over-charge (m/z) range and differing in intensities. To match MS2 spectra to sequences, database search is performed, in which each spectrum is compared to a set of theoretical candidate spectra and the highest similarity peptide is reported[1]. Theoretical databases once only contained expected m/z values based on calculated masses of individual fragments, and each theoretical fragment ion was equally intense. Now, with millions of experimental spectra available in public repositories, proteomics algorithms can finally leverage intensity information to better match peptides when m/z information alone is insufficient to achieve unambiguous identification.

One way to use intensity information is clustering. In clustering, spectra with high similarity are put into a group, often representing one or a few peptides. While clustering has been successful at summarizing spectra across experiments, it does not produce pure clusters of only one peptide[2,3]. This is because the similarity metrics applied by clustering algorithms are unsupervised. That is, if two spectra have a similarity above a certain threshold, they are grouped together even if they are known to be from different sequences. Rather than using static similarity measures, metric learning creates its own flexible similarity function through supervised learning by looking for commonalities within a group[13]. Metric learning on MS2 spectra has been applied for peptide clustering by the tool GLEAMS[8], which uses a Siamese neural network (SNN) to learn a low dimensional embedding/vector for a spectra. Spectra from the same peptide group closer in space than unrelated peptides, even if those peptides are of the same mass or are often confused for one another by a database search. Despite the improvement of metric learning over predefined similarity metrics, embeddings learned by GLEAMS lack interpretability. For example, how are peptides of the same sequence but different charges related? How are modified peptides related to their unmodified counterparts? These questions have yet to be explored. Studying this embedding space may illuminate the black box of peptide embeddings.

The motivation behind this project is to leverage embedding information for better peptide identification. Word embedders can learn arithmetic relationships between related words[6]. While word2vec and GLEAMS use different training algorithms, it may be interesting to see if inducing a change like addition of a PTM can cause a consistent shift of peptides in our embedding space. An important application in proteomics would be using unmodified peptide embeddings to better predict modified peptide spectra. Spectral prediction for unmodified peptides is already very accurate[14,15], but much less research has been done for PTM spectral prediction, where modified peptides constitute a large part of unidentified or wrongly identified spectra. Furthermore, embeddings have not yet been tested to differentiate PTMs with similar fragmentation patterns and masses. Since GLEAMS performed well on unmodified peptides, this is the natural next step and may greatly benefit our interpretation and prediction of modified peptides, the “dark matter” of the proteome.

B. Significance

Writing software to better annotate data from proteomics experiments will lead to more accurate interpretation of biological systems at the protein level, which can lead to discovery of new disease biomarkers. This project focuses on post-translational modifications, which alter protein shape and function.

First, this project aims to approach MS2 spectrum prediction from a new angle. While the prediction model Prosit[14] also learns a lower-dimensional, latent representation of peptides, this latent space was never explored. This project aims to connect the two worlds of dimensionality reduction through embedding and spectral prediction. The model from aim 1 can almost be thought of as an autoencoder[16], a technique known for taking an input image, learning a lower-dimensional representation, and then decoding the representation back to the original input. Both aims also require the availability of a spectrum from an unmodified peptide before working with modified peptide spectra. While this does impose the limitation of only working with previously seen unmodified peptides, we may assume that if a modified peptide has been detected, its unmodified form will also be present. Typical MS2 spectrum predictors do not require this *a priori* information, but we hypothesize that an unmodified peptide counterpart will often be available and that it will help make better predictions by shifting the focus away from learning the unmodified spectrum and focusing just on how the PTM modifies the base spectrum. Aim 2 is specifically important because isobaric PTMs are notoriously difficult to differentiate (i.e., sulfopeptides are misidentified as phosphopeptides, and structurally different

isobaric glycans may stay ambiguous). Though rules exist to tell sulfo- and phosphopeptides apart, no tool – to our knowledge – exists to automatically annotate them. Overall, this project may help biologists studying the phosphoproteome to assign identities more confidently to spectra.

C. Research Strategy

C.1. Aim 1: Determine where phosphorylated peptides are mapped relative to unmodified counterparts

C.1.1. Introduction

Embedding is a technique to map peptides to *N*-dimensional space, where *N* is much smaller than the number of m/z bins originally representing an MS2 spectrum. Peptide embeddings have shown that spectra can cluster by m/z, but positions of PTMs have yet to be explored[8]. Because neural networks (NNs) learn a non-linear function to map spectra to their embeddings, theoretically this information could be useful in the opposite direction, where embedding information can be used to predict the original spectra that were mapped there. While predictions for unmodified peptides are now commonplace for spectral library generation[14] and PSM rescoring[17,18], predictions for modified peptides are available but not as commonly used due to lower accuracy.

One PTM that has generated interest from both the proteomics and deep learning communities is phosphorylation. Phosphorylation is a reversible PTM. The balance of kinase and phosphatase activity determines the proportion of proteins that are phosphorylated and active, and this has implications for activating signaling pathways[10]. Furthermore, many diseases can arise due to kinase (in)activity, leading to changes in these pathways[10]. Biologists can use LC-MS/MS to reveal disease-specific differences in phosphorylation levels on different proteins, generating further hypotheses and experiments. Because so much work has been done on phosphorylation, sufficient data has been generated to train models for a variety of predictive purposes, ranging from physicochemical properties[19] to phosphosite predictions[20].

C.1.2. Data sources

Table 1 lists the phosphopeptide (pSTY peptide) datasets we can use for model development. Spectra from various high-resolution instruments and fragmentation methods allow for learning of general pSTY fragmentation patterns. High-energy collisional dissociation (HCD) and collision-induced dissociation (CID) make up the bulk of the data. Tens of thousands of pSTY peptides and PSMs provide a large dataset with which to train and test our models. Synthetic peptides also increase our confidence in sequence and PTM localization due to the availability of ground truth identities. Importantly, high representation of the three most common phosphosites allows our models to be applicable to most phosphorylation events.

ID	Instrument	Fragmentation	Organism	Synthesis	Notes
PXD009449	Orbitrap Fusion Lumos	HCD/CID/ETD/EThcD/ETciD	Synthetic (human)	176 phosphopeptides synthesized; only Y	37,692 PSMs with pFind3
PXD000138	LTQ Orbitrap Velos	HCD/ETD	Synthetic (human)	>100,000 phosphopeptides and unmodified counterparts	57,830 phosphopeptides detected with MASCOT
PXD000612	Q Exactive	HCD	HeLa cells		>50,000 distinct phosphorylated pept des with MaxQuant
PXD007058	Orbitrap Fusion ETD	HCD/EThcD/ETciD	U2OS cells/synthetic	191 phosphosites	Multiple resolut ons; Orbitrap/ion trap for MS2; thousands of PSMs detected with Andromeda/MASCOT
PXD000474	LTQ Orbitrap Velos; Q Exactive	HCD/CID/ETD	HeLa cells/synthetic	20 synthetic	Thousands of PSMs
PXD012433	TripleTOF 6600	CID	K562/synthetic	62 synthetic	Thousands of PSMs

Table 1. All proposed phosphorylated datasets. All datasets are deposited in ProteomeXchange (PXD) and performed phosphopeptide enrichment or examined synthetic peptides.

C.1.3. Preliminary data

Word embedders such as word2vec are lauded for their ability to learn word representations that are arithmetically meaningful, but they were not trained to do that. Rather, word2vec's unexpected skill is a byproduct of its goal to predict words likely to be in the same sentence as a target word[6]. It is unknown if SNNs can behave similarly to these word embedder algorithms. While SNNs can perform metric learning to separate dissimilar peptide pairs, the loss functions used do not explicitly contain terms to separate them in a particular direction; rather, they only aim to increase the distance between the pairs, regardless of direction. To our knowledge, this would be the first examination of how pairs in an SNN are directionally related.

If our embedding space shows a non-random relationship between modified-unmodified pairs, the learned embeddings may prove useful for spectral prediction. Spectral predictors pDeep2 and 3 use recurrent neural networks (RNNs) to learn how amino acids in the sequence interact to generate the observed fragment ion intensity patterns[15,21]. pDeep2/3 performs well for phosphopeptide prediction after transfer learning (TL), with 80.6% of the predicted spectra for testing set peptides showing >90% Pearson correlation with observed spectra. This shows that phosphopeptide spectral prediction is possible.

C.1.4. Methods and rationale

Phospho-enriched datasets will be analyzed in Fragpipe, using MSFragger[22] for database searching, Percolator[23] for PSM rescoring and false discovery rate (FDR) control, and PTMProphet[24] for phosphosite localization and false localization rate (FLR) control. Only PSMs with a qval < 0.01 will be passed to PTMProphet, and only PSMs with a mean best probability statistic > 0.9 will be selected for SNN training.

In the development of GLEAMS, the only PTM included on peptides was oxidized methionine (oxM), ignoring the one study using TMT 6-plex labeling. As a first pass at finding PTM vectors, we can examine pairs of unmodified-oxidized peptides. We present two ways to visualize vectors between the pairs:

1. Perform dimensionality reduction on thousands of spectrum embeddings with UMAP[25]. Plot the resulting vectors from unmodified to modified peptide on a polar plot. Create another polar plot with vectors between pairs of same-charge unmodified peptides with a mass difference equal to the oxM mass shift, serving as a null distribution. Perform Kolmogorov-Smirnov tests to determine if there exists a statistically significant difference between the distributions of vector directions of the two polar plots.
2. Perform dimensionality reduction on the vectors between pairs with UMAP. Do this for both unmodified-oxM pairs and the null pairs described above. Cluster with an unsupervised algorithm such as HDBSCAN[26] to see if clusters can differentiate unmodified-oxM pairs from null pairs.

If oxM vectors are non-randomly distributed in the embedding space, it may be possible that phosphorylation is meaningfully positioned too. We can do TL on the GLEAMS SNN to embed phosphorylated spectra, taking as positive pairs only those spectra from peptides with the same sequence, charge, and phosphosites (and potentially fragmentation method). Then we can adapt the two steps above for the pSTY peptides to see if the SNN learned to differentiate phosphosite localization. We can evaluate the accuracy of the embeddings by embedding peptides with the same sequence but different phosphosites and calculating pairwise distances. We expect, for example, that phosphoserine (pS) spectra of the same sequence are closer to each other than pS and phosphotyrosine (pY) peptides. If there exist vectors that relate unmodified-pSTY pairs, we can use K-nearest neighbors search through the Facebook AI Similarity Search library[27] to determine how close “unmodified embedding + oxM vector + pSTY vector” is to the true embedding of this doubly modified peptide (Fig 1). Finally, should there be a non-random relationship between unmodified and phosphorylated embeddings, we can train a NN to predict pSTY spectra. The NN will be comprised of two modules: 1) the SNN to provide the input unmodified peptide’s embedding and 2) a RNN that will take the phosphopeptide sequence. Comparison of distributions of similarity between predicted and observed spectra can be made against pDeep2[21].

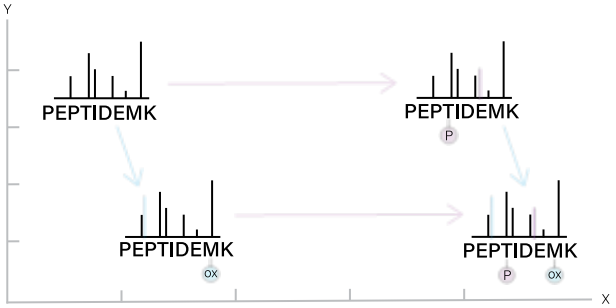

Figure 1. An example of a learned 2D embedding space. Purple arrows indicate the phosphorylated vector, blue arrows the oxidation vector

If no relationship exists between unmodified-modified pairs, more complex architecture in a “direction-enforcing NN” may help enforce a pattern (Fig 2). pDeep2, which was trained to predict b/y/neutral loss ion intensities for an input peptide, can be adapted with TL to instead learn a low-dimensional, *n*-length representation of phosphorylation events. The shorter vector can be concatentated onto the GLEAMS 32-dimensional (32-D) embedding of the unmodified peptide. A transfer-learned version of GLEAMS, referred to for now as pGLEAMS, can be trained to predict a (32+*n*)-D embedding of pSTY peptides. For instance, a spectrum from PEPTIDESK produces a 32-D GLEAMS embedding, which is concatenated to an *n*-D embedding learned by a RNN for input sequence PEP(p)TIDESK, where “(p)” is the phosphorylation event, in this case localized to T. A spectrum from PEP(p)TIDESK produces a (32+*n*)-D pGLEAMS embedding, and contrastive loss[28] can pull their embeddings closer. Meanwhile, if a PEPTIDE(p)SK spectra is fed into pGLEAMS with PEP(p)TIDESK fed into the RNN, the loss function should push these peptidoforms apart. Importantly, this approach does not explicitly enforce a direction for a specific phosphorylation fragmentation pattern, but because only a few dimensions (those learned by the RNN) change during training, separation of pS vs pT vs pY patterns will need to be efficient. This need for efficiency may lead to predictable vectors. If the embeddings from this new model architecture are reliable, pSTY spectral prediction would be the next step. The NN architecture in this section no longer represents a SNN, as networks with shared weights are not used.

C.1.5. Expected outcomes, pitfalls, and alternatives

The expected outcome is a set of phosphorylation vectors that can be helpful for pSTY spectra prediction and can be used to group phosphorylation events by site-specific fragmentation patterns. Because SNNs do not explicitly enforce a direction between pairs of related data points, it is possible that these "PTM vectors" are not obtainable. A NN may be able to use the unmodified embedding for pSTY prediction despite this, albeit without the interpretability granted by the vectors. Another issue is the various fragmentation methods that can produce disparate spectra[12,29]. Considering a (CID, ETD) pair as the same may erase the meaningful differences between them if grouped to the same embedding. While our first model version might only use CID and HCD spectra, we can consider training on all fragmentation types, adding a fragmentation method variable into the input, and treating different methods as negative pairs. However, HCD/CID pSTY spectra outnumber ETD/EThcD/ETciD spectra in our dataset, threatening to oversaturate the training set. Another approach to working with less represented fragmentation types is to learn a general embedding regardless of fragmentation and use multitask[30] learning to predict multiple pSTY spectra for a single peptide representing different fragmentation modes.

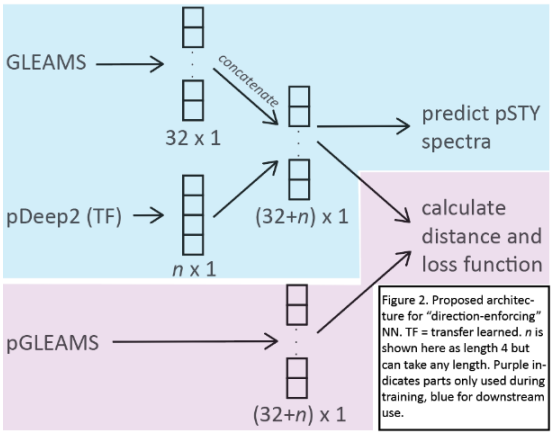

Figure 2. Proposed architecture for "direction-enforcing" NN. TF = transfer learned. *n* is shown here as length 4 but can take any length. Purple indicates parts only used during training, blue for downstream use.

C.2. Aim 2: Distinguish phosphorylated from sulfated peptides

C.2.1. Introduction

Sulfation has many functions, including activating signaling peptides[31], immune inhibition, and modulating protein-protein binding[32]. Almost 3% of tyrosines in HepG2 cells can be sulfated, mostly in secreted and membrane proteins[33,34]. Sulfation and phosphorylation serve unrelated functions. Proper PTM assignment is vital for accurate interpretation of modified proteins. Disentangling sulfation from phosphorylation is a long-standing challenge. Sulfate and phosphate groups have monoisotopic masses of 79.9568 and 79.9663 Da respectively, making it difficult to separate them by precursor mass shift alone if not using high resolution data[35]. Next, both occur on serine, threonine, and tyrosine, although sulfation is biased towards tyrosine, and phosphorylation serine[36]. Nonetheless, using site localization is not a fool-proof way of differentiating the two. Finally, many sample processing protocols have been developed for enrichment of sulfopeptides, including anti-sulfotyrosine antibodies[37] and affinity chromatography[38]. Importantly, phospho- and sulfopeptides will still appear in whole proteome data unenriched for these PTMs, so methods are needed to differentiate them in unenriched data.

Training data Testing data Extra data (training/testing)

ID	Instrument	Fragmentation	Organism	Sample processing	Notes
PXD018229	TripleTOF 5600	HCD	Chinese Hamster ovary cells, human protein expression		also searched for phosphorylation
PXD016778	Orbitrap Fusion Lumos	HCD/EThcD	Tick/synthetic	Secreted protein purification with nickel affinity chromatography; antisulfotyrosine mAB confirmed in separate sample	
PXD010964	Q Exactive	HCD	Moss		
PXD009536	Q Exactive	HCD	Stickbug		MaxQuant detected no sulfation
PXD008240	Orbitrap Fusion ETD	HCD/ETD/EThcD	Mosquito/synthetic		
PXD007614	Q Exactive	HCD	Medicago truncatula (legume)		
PXD002244/PXD003015	LTQ-Orbitrap VELOS	CID	Mouse	Cell membrane protein enrichment	no explanation for why sulfation was included as variable mod
PXD001443	LTQ Orbitrap	CID	Rat	Synaptosome purificiation	also searched for phosphorylation; no explanation for why sulfation was included as variable mod
MSV000078492	TripleTOF 5600	CID	Synthetic peptides in yeast lysate		iPRG 2012 competition, 5 sulfated peptides
Chen et al 2018	Orbitrap Fusion Lumos	CID/HCD/ETD/ETciD/EThcD	Synthetic peptides in human lysate		8 each of sulfopeptides and phosphopeptides with same sequence
PXD000865	LTQ Orbitrap Velos/Elite	HCD	Human		ProteomicsDB; Reanalysis previously done in thyroid
PXD010154	Orbitrap Fusion Lumos/ Q Exactive	HCD/CID/ETD/EThcD	Human		Human Protein Atlas

Table 2. All proposed sulfation datasets. All datasets are deposited in ProteomeXchange (PXD) or Massive (MSV) except for Chen et al. While no study specifically enriched for sulfopeptides, some sample processing steps are provided that may enrich for peptides that are more likely to be sulfated. Datasets in blue can provide PSMs for both training and testing.

C.2.2. Data sources

Table 2 lists datasets that have been searched for sulfopeptides (sY peptides) previously. All datasets have scans from at least HCD or CID. All initial training sets are from organisms besides humans, showcasing the unique use cases of this PTM. Two testing datasets are provided in Table 2. These datasets are good candidates because not only are the sY peptides synthesized, but also the manuscripts published alongside them provide a baseline for sulfation detection accuracy using search tool assignments and, in the case of the iPRG competition, manual inspection to correct pSTY assignments to sY. If more training or testing data is needed, ProteomicsDB and the Human Protein Atlas can be searched. UniProt[39] contains 57 human proteins with known sulfation sites. ProteomicsDB provides protein expression summaries. We can find in which tissues

each sulfated protein is most highly expressed and only perform database searches on those tissue-specific datasets from ProteomicsDB and the Human Protein Atlas. For example, osteomodulin with its nine sulfation sites is most highly expressed in breast and osteosarcoma cells.

C.2.3. Preliminary data

sY and pSTY peptides can fragment in drastically different ways, and our proposed method will rely heavily on MS2 fragment ion intensity differences to distinguish the PTMs. Note that while negative ion mode MS has been used extensively to characterize and differentiate sY and pSTY peptides[40,41], positive ion mode MS is more popular in proteomics and will be the focus of our project. Between MS2 spectra resulting from phosphorylation on serine/threonine and tyrosine, the latter case may produce spectra that are more distinct from sY peptide ones. Notably, in CID and HCD, sY peptides exhibit a precursor neutral loss peak much stronger than that of phosphotyrosine (pY) peptides, overwhelming the lower intensities of fragment ions. This is in accordance with the very labile nature of sulfate trioxide (SO_3)[35,40,42]. Importantly, neutral loss of meta-phosphoric acid (HPO_3) is isobaric to SO_3 (-80 Da). sY peptide fragment ions are also more likely to show the neutral loss compared to pY peptide fragment ions. The use of diagnostic ions can be informative as well; pY produces an immonium ion at 216.043 m/z, although its detection can be hampered by other peaks in the region[41]. So long as the pY spectra has fragments that span the modification site, these differences can help distinguish the two PTMs on Y. This is not always the case, however, as acidity and length of the peptide can have an effect on frequency of SO_3 neutral loss[43]. Collision energy and physicochemical properties of the peptide can also change intensity of the pY immonium ion[41,44,45].

Compared to pY peptides, pST peptides produce more intense neutral loss peaks[44]. The extent to which this neutral loss occurs is influenced by the proton mobility and amino acid sequence of the peptide. Interestingly, in addition to the -80 Da loss, pST peptides tend to lose phosphoric acid (H_3PO_4), a -98 Da neutral loss. This neutral loss can also be found in some sY peptide spectra resulting from additional loss of SO_3 with either water or ammonia[35]. Because the majority of phosphopeptides are modified on S and T, these intense neutral loss peaks will be the norm. Unmodified fragments and fragments with neutral loss also have the same m/z, making unambiguous localization difficult.

C.2.4. Methods and rationale

Searching for sulfation can be a great challenge without high resolution instruments or sY peptide enrichment. Although some of the datasets include synthetic peptides, they may not be enough for TL, which requires thousands of PSMs from tens to hundreds of peptides[21]. Therefore, we must search the remaining non-synthetic data. Without ground-truth labels like in synthetic data, we require a method to ensure we do not confuse pSTY peptides for sY ones during construction of our sY training/testing sets. However, we also do not want to use the known patterns of sY and pSTY peptide fragmentation from *C.2.3 "Preliminary data"*. For example, if we only included spectra with strong neutral loss peaks in our sY set, the model would not realize that some sY peptides may be an exception to the rule. Instead, we want our model to learn these rules (and exceptions) by itself, as well as any other patterns of fragment intensity ratios or sulfation diagnostic ions that it may find. This method of labeling training data avoids biases from prior knowledge and does not limit our model to our current understanding of sY peptide behavior. We will use the following rules to acquire sY peptide spectra with a minimal number of accidental pSTY PSMs:

1. Only accept peptides with known prior sulfation events from UniProt or literature.
2. Exclude peptides known to contain phosphosites from UniProt or literature.
3. Exclude peptides with 216.043 m/z immonium ion.
4. Exclude peptides whose PTMProphet[24] site localization reports higher probabilities for S/T than Y.

Closed searches with variable modifications of oxM and N-terminal acetylation and labile modification of sulfation will be performed with MSFragger-Glyco[46], a version of MSFragger adept at identifying peptides with labile PTMs. Spectral processing will be performed with removal of the precursor peak. Importantly, while the precursor peak will not be considered during the database search, it will still be included in the spectra used for SNN training and testing. PSMs will then be processed by Percolator[23]. Percolator rescoring can make use of many features, including one that measures agreement of experimental RT with predicted sulfation RT[19]. Only PSMs with a qval < 0.01 will be further processed by the five exclusion criteria above to form our final sY data.

Phosphorylated PSMs from Aim 1 can be reused here. Unmodified PSMs can be obtained from the searched data (if a modified peptide is found, its unmodified form will likely also be found) from peptide-centric search engines such as PepCentric (unpublished work from our lab), or from MS2 spectrum prediction models[15,47].

For TL of our SNN, we will retrain the last layer of the GLEAMS twin convolutional NNs to cluster sulfated peptides closer to unmodified peptides. While it is more intuitive to deem a sY peptide as a "fraudulent phosphorylation" event, we take unmodified-sulfation pairs as positive pairs. This is because sY spectra are more likely to show unmodified fragments, making it easier to cluster them towards unmodified peptide spectra. Furthermore, phosphorylation on Y can look very different from phosphorylation on S and T, resulting in more diverse spectra. We hypothesize that it will be easier to adapt a model to cluster sY and unmodified spectra together, rather than having it cluster all the phosphorylation variations together. While we have many more negative pairs at our disposal (i.e., more phosphorylated than sulfation examples), we can downsample them to have equal numbers of positive and negative training examples.

The architecture of our SNN will be the same as GLEAMS, with the addition of a feedforward layer to calculate a weighted L1 distance between the inputs' embeddings and a sigmoid activation function to map the distance to a probability from 0 to 1. The loss function used will be binary cross-entropy. Model performance will be evaluated with two synthetic testing sets. Receiver operating characteristic (ROC) curves will be plotted and area under the curve (AUC) calculated. Two other baseline models can be trained to predict sulfation probability and can be used for comparison: 1) a logistic regression model and 2) an XGBoost model, known to be powerful on tabular data while requiring fewer training samples than NNs[48]. In both cases, the variables include cosine similarity to predicted phosphorylated spectra[21], ion intensities of unmodified/modified fragment ions, precursor ion intensity, and pY diagnostic ion intensity. These variables reflect our current understanding of MS2 spectral differences between pSTY and sY peptides and test whether our SNN's combination of NN architecture and ability to learn new differentiating spectral features can perform better. We may also consider inclusion of site localization probability to S, T, or Y[24]; probability of being a phospho- or sulfosite[20,49]; and difference from predicted RT[19,50] as variables to include in our baseline models.

Apart from ascertaining the accuracy of our SNN, we also wish to probe the model for what new patterns it finds in the sulfated MS2 spectra. While non-linear convolutional NNs are less intuitive than linear regressors, much research has been done to demystify the black box. For example, SHAP values will be indispensable for interpreting the importance of the input features[51]. Because 2,449 m/z binned fragment intensities are used as input, we can compute SHAP values for all m/z bins for a large sample of sY peptides to detect informative m/z bins i.e., diagnostic ions.

C.2.5. Expected outcomes, pitfalls, and alternatives

The expected outcome is a SNN, which we will call SulfoSNN, that predicts probability that a peptide is sulfated, given a PSM of its unmodified counterpart. Interpretability from SHAP values will help explain the model's reasoning for its decision. The success of this model depends on the existence of enough training samples. Despite the bottleneck of acquiring sulfation data, more sY peptides can be synthesized and run on the Orbitrap Fusion Tribrid in our Proteomics Resource Facility, if needed. Another alternative is to show the method works on other isobaric or near-isobaric PTMs[52], such as glycans, for which more data may be available. Another potential issue is that TL may not find appropriate weights when both retraining the embedding layer and training the fully connected layer before sigmoid activation/probability output. In that case, we can retain the original GLEAMS architecture of the embedding layer being the last layer and only retrain that. Then, we can sort the distances from unmodified PSMs of our testing samples to determine a distance threshold below which we denote a pair as sulfated, such as 1% FDR or Youden's statistic[53] to maximize the difference between true and false positive rates.

References

1. Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics* vol. 73 2092–2123 (2010).
2. Frank, A. M. *et al.* Clustering millions of tandem mass spectra. *J. Proteome Res.* **7**, 113–122 (2008).
3. Griss, J., Foster, J. M., Hermjakob, H. & Vizcaíno, J. A. PRIDE Cluster: Building a consensus of proteomics data. *Nature Methods* vol. 10 95–96 (2013).

4. Dutta, D. & Chen, T. Speeding up tandem mass spectrometry database search: Metric embeddings and fast near neighbor search. *Bioinformatics* **23**, 612–618 (2007).
5. Bittremieux, W., Meysman, P., Noble, W. S. & Laukens, K. Fast Open Modification Spectral Library Searching through Approximate Nearest Neighbor Indexing. *J. Proteome Res.* **17**, 3463–3474 (2018).
6. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. in *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings* (International Conference on Learning Representations, ICLR, 2013).
7. Chicco, D. Siamese Neural Networks: An Overview. in *Methods in Molecular Biology* vol. 2190 73–94 (Humana Press Inc., 2021).
8. May, D. H., Bilmes, J. & Noble, W. S. A learned embedding for efficient joint analysis of millions of mass spectra. (2018) doi:10.1101/483263.
9. Qin, C. *et al.* Deep learning embedder method and tool for mass spectra similarity search. *J. Proteomics* **232**, 104070 (2021).
10. Ardito, F., Giuliani, M., Perrone, D., Troiano, G. & Muzio, L. Lo. The crucial role of protein phosphorylation in cell signalingand its use as targeted therapy (Review). *International Journal of Molecular Medicine* vol. 40 271–280 (2017).
11. Degnore, J. P. & Qin, J. Fragmentation of phosphopeptides in an ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.* **9**, 1175–1188 (1998).
12. Potel, C. M., Lemeer, S. & Heck, A. J. R. Phosphopeptide Fragmentation and Site Localization by Mass Spectrometry: An Update. *Analytical Chemistry* vol. 91 126–141 (2019).
13. Kulis, B. Metric Learning: A Survey. *Found. Trends R Mach. Learn.* **5**, 287–364 (2012).
14. Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods 2019 166* **16**, 509–518 (2019).
15. Tarn, C. & Zeng, W.-F. pDeep3: Toward More Accurate Spectrum Prediction with Fast Few-Shot Learning. *Anal. Chem.* **93**, 5815–5822 (2021).
16. Rumelhart, D. E. & McClelland, J. L. Learning Internal Representations by Error Propagation - MIT Press books. *Parallel Distrib. Process. Explor. Microstruct. Cogn. Found.* 318–362 (1987).
17. Wilhelm, M. *et al.* Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. *Nat. Commun. 2021 121* **12**, 1–12 (2021).
18. Li, K., Jain, A., Malovannaya, A., Wen, B. & Zhang, B. DeepRescore: Leveraging Deep Learning to Improve Peptide Identification in Immunopeptidomics. *Proteomics* 1900334 (2020) doi:10.1002/pmic.201900334.
19. Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L. & Degroeve, S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *bioRxiv* 2020.03.28.013003 (2021) doi:10.1101/2020.03.28.013003.
20. Song, J. *et al.* PhosphoPredict: A bioinformatics tool for prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection. *Sci. Reports 2017 71* **7**, 1–19 (2017).
21. Zeng, W. F. *et al.* MS/MS Spectrum prediction for modified peptides using pDeep2 Trained by Transfer Learning. *Anal. Chem.* **91**, 9724–9731 (2019).
22. Kong, A. T., Leprevost, F. V, Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat. Methods 2017 145* **14**, 513–520 (2017).
23. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods 2007 411* **4**, 923–925 (2007).
24. Shteynberg, D. D. *et al.* PTMProphet: Fast and Accurate Mass Modification Localization for the Trans-Proteomic Pipeline. *J. Proteome Res.* **18**, 4262 (2019).
25. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2018).
26. Campello, R. J. G. B., Moulavi, D. & Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **7819 LNAI**, 160–172 (2013).
27. Johnson, J., Douze, M. & Jégou, H. Billion-scale similarity search with GPUs. *IEEE Trans. Big Data* **7**, 535–547 (2017).
28. Hadsell, R., Chopra, S. & Lecun, Y. Dimensionality Reduction by Learning an Invariant Mapping.
29. Chen, G., Zhang, Y., Trinidad, J. C. & Dann, C. Distinguishing Sulfotyrosine Containing Peptides from

their Phosphotyrosine Counterparts Using Mass Spectrometry. *J. Am. Soc. Mass Spectrom. 2018 293* **29**, 455–462 (2018).
30. Caruana, R., Pratt, L. & Thrun, S. Multitask Learning *. **28**, 41–75 (1997).
31. Matsuzaki, Y., Ogawa-Ohnishi, M., Mori, A. & Matsubayashi, Y. Secreted Peptide Signals Required for Maintenance of Root Stem Cell Niche in Arabidopsis. *Science (80-.).* **329**, 1065–1067 (2010).
32. Franck, C. *et al.* Semisynthesis of an evasin from tick saliva reveals a critical role of tyrosine sulfation for chemokine binding and inhibition. *Proc. Natl. Acad. Sci.* **117**, 12657–12664 (2020).
33. HILLE, A. & HUTTNER, W. B. Occurrence of tyrosine sulfate in proteins – a balance sheet. *Eur. J. Biochem.* **188**, 587–596 (1990).
34. HILLE, A., BRAULKE, T., FIGURA, K. von & HUTTNER, W. B. Occurrence of tyrosine sulfate in proteins – a balance sheet. *Eur. J. Biochem.* **188**, 577–586 (1990).
35. Chen, G., Zhang, Y., Trinidad, J. C. & Dann, C. Distinguishing Sulfotyrosine Containing Peptides from their Phosphotyrosine Counterparts Using Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **29**, 455–462 (2018).
36. Hanno Steen, †, Bernhard Küster, ‡, Minerva Fernandez, †, Akhilesh Pandey, §,⊥ and & Matthias Mann*, †,‡. Detection of Tyrosine Phosphorylated Peptides by Precursor Ion Scanning Quadrupole TOF Mass Spectrometry in Positive Ion Mode. *Anal. Chem.* **73**, 1440–1448 (2001).
37. AJ, H., E, D., KG, B., JA, L. & KL, M. Detection and purification of tyrosine-sulfated proteins using a novel anti-sulfotyrosine monoclonal antibody. *J. Biol. Chem.* **281**, 37877–37887 (2006).
38. G, D. B. *et al.* Analysis of sulfated peptides from the skin secretion of the Pachymedusa dacnicolor frog using IMAC-Ga enrichment and high-resolution mass spectrometry. *Rapid Commun. Mass Spectrom.* **25**, 1017–1027 (2011).
39. Consortium, T. U. *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
40. Hersberger, K. E. & Håkansson, K. Characterization of O-Sulfopeptides by Negative Ion Mode Tandem Mass Spectrometry: Superior Performance of Negative Ion Electron Capture Dissociation. *Anal. Chem.* **84**, 6370–6377 (2012).
41. Edelson-Averbukh, M., Shevchenko, A., Pipkorn, R. & Lehmann, W. D. Discrimination Between Peptide O-Sulfo- and O-Phosphotyrosine Residues by Negative Ion Mode Electrospray Tandem Mass Spectrometry. *J. Am. Soc. Mass Spectrom. 2011 2212* **22**, 2256–2268 (2011).
42. Nemeth-Cawley, J. F., Karnik, S. & Rouse, J. C. Analysis of sulfated peptides using positive electrospray ionization tandem mass spectrometry. *J. Mass Spectrom.* **36**, 1301–1311 (2001).
43. Bundgaard, J. R., Johnsen, A. H. & Rehfeld, J. F. Analysis of Tyrosine-O-Sulfation. *Methods Mol. Biol.* **194**, 223–239 (2002).
44. Potel, C. M., Lemeer, S. & Heck, A. J. R. Phosphopeptide Fragmentation and Site Localizationby Mass Spectrometry: An Update. *Anal. Chem.* **91**, 126 (2019).
45. Mogjiborahman Salek, †, Angel Alonso, ‡, R. Pipkorn, § and & Wolf D. Lehmann*, †. Analysis of Protein Tyrosine Phosphorylation by Nanoelectrospray Ionization High-Resolution Tandem Mass Spectrometry and Tyrosine-Targeted Product Ion Scanning. *Anal. Chem.* **75**, 2724–2729 (2003).
46. Polasky, D. A., Yu, F., Teo, G. C. & Nesvizhskii, A. I. Fast and comprehensive N - and O - glycoproteomics analysis with MSFragger-Glyco. *Nat. Methods 2020 1711* **17**, 1125–1132 (2020).
47. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: Neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **17**, 41 (2020).
48. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* **13-17-August-2016**, 785–794 (2016).
49. AL-barakati, H. J. *et al.* SVM-SulfoSite: A support vector machine based predictor for sulfenylation sites. *Sci. Reports 2018 81* **8**, 1–9 (2018).
50. Ma, C. *et al.* Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning. *Anal. Chem.* **90**, 10881–10888 (2018).
51. Lundberg, S. M., Allen, P. G. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions.
52. Kim, M.-S., Zhong, J. & Pandey, A. Common errors in mass spectrometry-based analysis of post-translational modifications. *Proteomics* **16**, 700 (2016).
53. Index for rating diagnostic tests - Youden - 1950 - Cancer - Wiley Online Library. https://acsjournals.onlinelibrary.wiley.com/doi/10.1002/1097-0142%281950%293%3A1%3C32%3A%3AAID-CNCR2820030106%3E3.0.CO%3B2-3.