Exploiting Charge-State Distribution to Probe Salt-Bridge Structures in Gas-Phase Phosphopeptides and Enhance Proteomics Analyses

Vladimir Gorshkov and Frank Kjeldsen

Department of Biochemistry and Molecular Biology, University of Southern Denmark, DK-5230 Odense M, Denmark

Abstract

Charging of analytes is a prerequisite for performing mass spectrometry analysis. In proteomics, electrospray ionization stands as the dominant technique for this process. Although the observation of differences in the peptide charge state distribution (CSD) is well-known among experimentalists, its analytical value remains underexplored. To investigate the utility of this dimension, we analyzed several public datasets, comprising over 250,000 peptide CSD profiles from the human proteome. We found that the CSD dimension demonstrates high reproducibility across multiple laboratories and extensive time intervals. The general observation was that the CSD dimension enabled effective partitioning of the peptide property space, resulting in enhanced discrimination between sequence or constitutional peptide isomers. Next, by evaluating the CSD values of phosphorylated peptides, we were able to differentiate between phosphopeptides that we presume adopt salt-bridge structures in the gas phase and those that do not. The reproducibility of the CSD values, in general (mean cosine similarity above 0.97 for most of the experiments), qualified CSD data suitable to train a deep learning model capable of accurately predicting CSD values (negative cosine similarity below –0.98). When we applied the CSD dimension to MS1- and MS2-based proteomics experiments, we consistently observed an increase in protein identification rate. Although the CSD dimension is not as effective a discriminator as the widely-used retention time dimension, it still holds potential for application in direct infusion proteomics.

Electrospray ionization (ESI) is by far the most dominant ionization technique used for biomolecule ionization in mass spectrometry,[1] including the extensive field of proteomics. ESI is frequently coupled with liquid chromatography-tandem mass spectrometry (LC-MS/MS) and offers efficient ionization of a broad range of biomolecules.

Under ESI conditions, peptides can adopt several protonation states. Some of these states are more likely and thus more abundant, leading to the formation of a peptide charge-state distribution (CSD). Currently, proteomics almost exclusively uses positive ion mode, i.e. ionization resulting in positively charged ions. Peptide cations are preferentially produced by protonation of the most basic sites in peptides. The gas-phase basicity (GB) of these sites is determined by various factors, including the intrinsic gas-phase basicity (GBint) of amino acids, although it is not the sole decisive parameter. Other

factors also play significant roles. For instance, the presence of chemical functional groups capable of sharing protons, known as charge solvation groups, can increase the energy required to remove a proton from a charged site, effectively enhancing the GB of that site. Additionally, the Coulomb repulsion between neighboring protons already present reduces the proton affinity of a potential protonation site. This reduction in GB can make a residue that is inherently less basic but more distant from existing protons more favorable for protonation. Combined, these factors influence the pattern of protonation which has been measured experimentally[2,3] and has been a subject of investigation since the beginning of tandem mass spectrometry[4,5]. The peptide CSD is also influenced by the peptide conformational space, which encompasses the array of possible configurations arising from different combinations of bond rotations, dihedral angles, and side chain orientations. In that way, different peptide structures can affect the accessibility of specific amino acid residues for protonation or deprotonation, thereby impacting the observed charge states. Finally, some polypeptides can form salt bridges in the gas phase, which are electrostatic interactions between amino acid residues of opposing charge and have been widely studied due to their structural and functional importance in proteins[6–9]. In the context of CSD, the formation of a salt bridge affects this parameter by neutralizing one positive charge in the bond. For phosphorylated peptides, the phosphate motif is particularly prone to forming salt bridge structures. This is because its molecular orbital system disperses its negative charge over two or more oxygen atoms. In contrast, other typical electron donors like Asp and Glu have resonance structures where the electron pair is delocalized over a smaller π-molecular orbital system[10]. Earlier important work has noted measurable compaction of phosphopeptides in ion mobility mass spectrometry[11–13]. In this relation, Clemmer and co-workers[14] determined lower intrinsic size parameters for the phosphorylation motif compared to any other amino acid. The intrinsic size parameter is a measure of the average value of each amino acid residue or modification that contributes to a peptide's cross-section. It was suggested that the presence of the phosphoryl group can play a role in internal interactions that cause an overall compacting effect on the peptide structure compared to sequences without this modification.

Peptide CSDs and their relation to peptide sequence have been the focus of two recent studies by Guan et al.[15] and Xu et al.[16]. The study of Guan et al. demonstrated the possibility of accurate prediction of the CSD using a deep learning LSTM model. The study of Xu et al. aims to uncover sequence features responsible for peptide charging by employing a more explainable logistic regression model based on peptide amino acid composition (i.e. not accounting for amino acid interactions). Both studies have indicated the potential utility of their models and findings to advance peptide identification, though definitive outcomes have yet to be presented. Additionally, neither study has considered phosphorylated peptides – a prevalent and extensively researched category of post-translationally modified peptides.

Building on these foundational studies, we pursue here to explore the broader analytical importance of CSD in proteomics and investigate in particular the CSD landscape of phosphorylated peptides. In doing so, we examined multiple public datasets, encompassing over 250,000 peptide/phosphopeptide CSD profiles from the human proteome. We emphasize its consistency, its likely capacity to distinguish phosphorylated peptides with salt-bridge structures, and its potential to enhance protein identification rates in both MS1 and MS2 proteomics experiments.

Results and Discussion

Charge-state distribution definition

Throughout this manuscript, under the term CSD, we assume the normalized (sum of all abundances equal to one) relative abundance of the peptide ions having charge (and the same protonation) states

from 1 to 6 for the same peptide sequence. Thus, the CSD of a peptide is a vector with six dimensions. Furthermore, one can use this vector to calculate the average observed charge state using the following formula: $\sum CSD \times \langle 1,2,3,4,5,6 \rangle$. Details of charge-state distribution calculation can be found in the Methods section.

Charge-state distribution is a universal property of peptides

To assess the reproducibility of the peptide CSD feature, we have analyzed several published DDA experiments (see Supplementary table for details) and extracted CSDs for all identified peptides. Focusing first on unmodified peptides allows us to limit the peptide conformational space to species composed of the 20 naturally occurring amino acid residues. The analyzed datasets contain HeLa digests analyzed at five different research laboratories over five years, employing several instrument types, resolutions, AGC targets, and sample load amounts. Supplementary Figure S1 represents the average cosine similarity between CSDs of the same peptides for all analyzed experiments. The majority of the LC-MS experiments display a striking level of similarity (average cosine similarity of 0.97 and higher) independently of the employed instrument, peptide sample load, gradient length, and other parameters. Especially, the independence from the gradient length and peptide sample load is quite fascinating. One could expect that short gradients will force more peptides to be eluted simultaneously and, thus, increase possible peptide ion suppression and distortion of the CSD during the ionization process. Larger peptide sample loads should have a similar effect. The data, however, clearly contradicts these expectations (compare CPR_QEX datasets for gradient length influence) and IMP_QEX (25 ng on column) with other LC-MS experiments (200 – 1000 ng on column). Some earlier reports[16] demonstrated changes in the CSD of certain peptides resulting from different gradient lengths. However, our data clearly show that such an observation is rather uncommon. Despite these vast differences in experimental conditions the smallest cosine similarity is 0.92 and observed only for two combinations of experimental sets. It should be noted, however, that all presented LC-MS experiments have been acquired using nano-HPLC setups employing water and acetonitrile with formic acid as chromatographic buffers on Orbitrap-based instruments from Thermo Scientific, thus, a possibility exists that a different MS instrument (i.e. mass analyzer or ion interface) or a different LC conditions (buffer composition, flow rate) will influence the CSD. The potential impact of different flow rates and solvent composition was investigated by analyzing the public data from a microflow LC system with DMSO added in the buffers (Supplementary Figure S2) – the mean cosine similarity decreased to 0.98 – 0.86. The influence of the mass analyzer has been addressed in a recent report indicating the similarity of CSDs recorded on timsTOF and Orbitrap instruments[16]. Combined these observations suggest that the CSD (given comparable LC-MS conditions) is a property of the peptide itself and, thus, could be potentially employed for peptide characterization.

Exploring the charge-state distribution properties and discriminatory power

To further explore the properties of the CSD dimension we reanalyzed a portion of a recently published dataset (PXD024364)[17] containing K562 cell lysates digested by trypsin and LysC and fragmented with higher-energy collisional dissociation (HCD). We extracted in total 1043541 CSDs from both trypsin and LysC datasets, these resulted in 234535 unique peptide CSD features. Supplementary Figures S3 a,b show the distribution of the cosine similarity between the CSDs and the standard deviation of the average calculated charge for the same peptides (for peptides detected more than once). Both figures indicate very high reproducibility of the detected CSDs, for example, over 96% of peptides have a deviation of the calculated average charge below 0.1. Table 1 summarizes the number of different observed charge states (i.e., the number of non-zero elements in the CSD vector) for detected peptides. Most peptides are detected in two or three different charge states, this observation aligns well with the earlier knowledge. Noteworthy, for approximately 42% of peptides

analyzed, a single charge state accounts for over 90% of the total ion abundance (Supplementary Figure S4). As expected, more abundant peptides tend to have a larger number of observed charge states (Supplementary Figure S3d), although, the trend is rather weak, thus, it is likely that the number of observed charge states is primarily influenced by the structure of the peptide itself. As seen in Supplementary Figure S3c the average calculated peptide charge of tryptic peptides spans a wide range from 1 to 5.5, while most of them are detected with an average charge from around 1.75 to 4.

Table 1. Observed charge states for peptides in the K562 dataset

Nr charge states	1	2	3	4	5	6
Nr peptides	19268	136843	63779	13049	1596	0
Percent	8.22%	58.35%	27.19%	5.56%	0.68%	0.00%

Xu et al[16] introduced a concept of *under- and overcharging* peptides and showed that peptides with less than three basic sites are preferably observed in the charge state above the number of basic sites (overcharging), while peptides with more than three basic sites are preferably observed in charge states below the number of basic sites (undercharging). Based on our analysis, the notion of peptides being categorized strictly into undercharged or overcharged classes seems overly simplistic. This distinction appears to stem from a fundamental relationship between the peptide's mass and its charge-bearing capacity. Given that the mean length of a tryptic peptide is 15 residues, there is a limit to how many protons it can bind. In the K562 dataset, we found that approximately 63.9% of peptides exhibit undercharging, 10.6% show overcharging, and the remainder display an average charge proximal to the number of basic sites (± 0.05). This correlation of mass, average charge, and the number of basic sites aligns with the trend outlined by Xu et al (See Supplementary Figure S5).

Since we have established that the CSDs can be reliably determined and are largely independent of the LC-MS conditions, we decided to investigate if this property is capable of distinguishing peptides. Modern MS instruments allow extremely robust and precise measurement of peptide mass, however, a significant number of peptide sequences are sequence isomers (i.e. the same amino acids in a different order), or constitutional isomers (i.e. the same composition of chemical elements that can be organized differently in the sequence, such as, for example, Q and GA). Both sequence and constitutional isomers cannot be distinguished by mass alone, since they have identical masses, and, thus, present a challenge for MS-based proteomics if the fragmentation of a peptide is incomplete or not used at all (MS/MS-free approaches). Around 6% and 52% of peptides in the K562 dataset have at least one sequence or constitutional isomer, respectively. The histogram of the cosine similarity and the average charge difference for peptides in each class is presented in Figure 1. The presented data corresponds to the minimal pairwise cosine similarity and the maximal pairwise difference in the average charge of the peptides in the group, i.e. having the same amino acid or elemental composition, thus, indicating if it is possible to differentiate at least one pair of peptides, using the corresponding metric. The shaded area indicates the region of cosine similarity below 0.95 and the average charge difference above 0.1. The values have been selected according to data presented in Supplementary Figure S1 (cosine similarity of replicated CSD estimations) and Supplementary Figure S3b (average charge state difference of the same peptide). The CSD is only slightly beneficial for the discrimination of sequence isomers (9%), while it allows significant differentiation of constitutional isomers (50%). The average charge state difference (Figure 1) presents a stronger discriminator in general. For sequence and constitutional isomers, 26% and 69% of peptides, respectively, could be distinguished. The second mode (with apex around 0.7) in constitutional isomers charge difference (Figure 1b) is particularly interesting. It indicates that a particular subset of constitutional isomers possesses an elemental configuration that changes drastically the average charge of a peptide. A

possible example of such a substitution can be an interchange of tryptophan and asparagine pair to a tyrosine and histidine pair that introduces amino acids with significantly different gas-phase basicity.

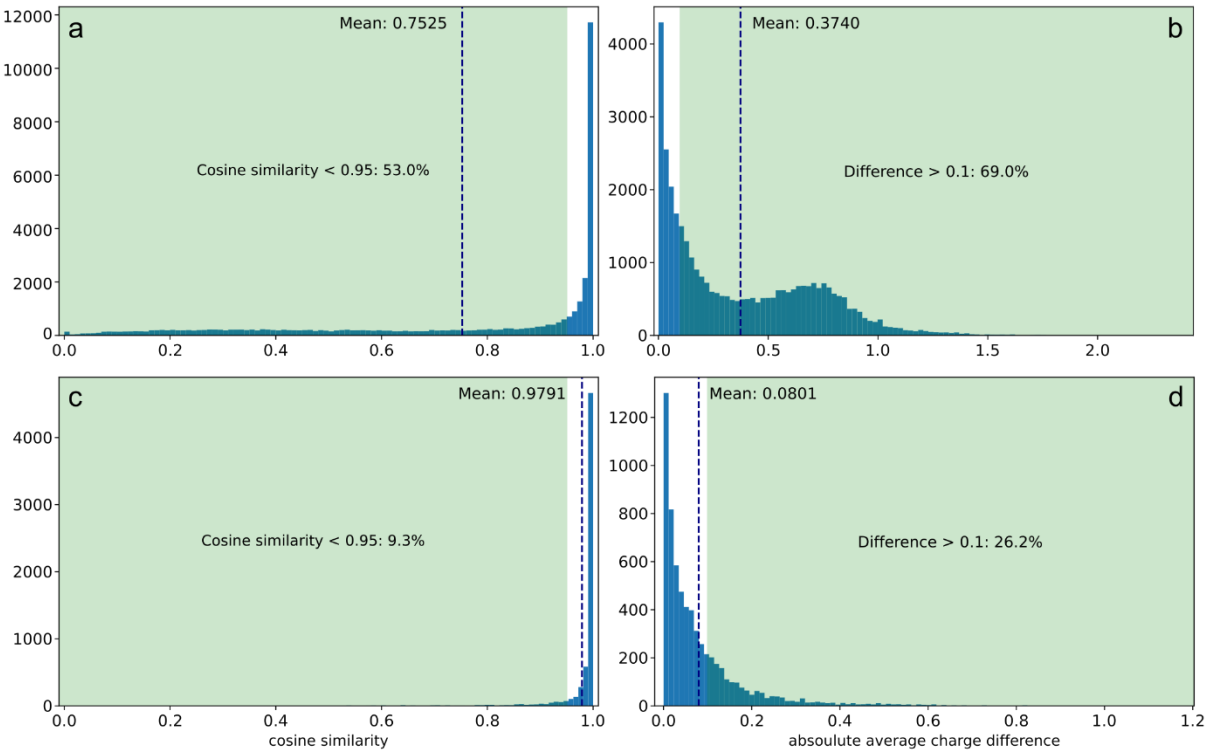

Figure 1: (a,b) cosine similarity and absolute difference in observed charge for strictly constitutional isomers (i.e. peptides with the same elemental composition, but not the same amino acid content); (c,d) the same for sequence isomers

Charge-state distribution of phosphorylated peptides

Phosphorylation of hydroxyl-containing amino acids (serine, threonine, and tyrosine) is one of the most common and widely studied post-translational modifications (PTM) of proteins. Due to the labile nature of this PTM, precise localization of the modification is still a challenging task. The phosphoric acid moiety can be readily deprotonated under common electrospray conditions and, thus, can participate in ion-ion interactions in liquid- and gas phases. Due to these features, it is expected that phosphorylation (and its position) can significantly influence the CSD of a peptide. In this section, we aim to characterize this impact.

As the first step, we obtained several datasets of phosphorylated peptides and extracted reliably identified phosphopeptides having confidently localized phosphosites (please, refer to the Methods section for details). The analyzed dataset contains LC-MS/MS experiments from four different research facilities, obtained over four years (2016 – 2019) and analyzed using several Orbitrap-based instruments. Like the results obtained for non-phosphorylated peptides, we observe very high reproducibility of CSD of phosphopeptides (most cosine correlations are above 0.97, and the lowest one is around 0.935), independent of the employed instrument, research facility, LC gradient, and other LC-MS parameters (Supplementary Figure S6).

Phosphopeptide identifications from all experiments were combined with two synthetic phosphopeptide datasets (see supplementary table for details) resulting in 531.267 CSD vectors that were merged into 42.531 unique (including phosphosite position) phosphopeptides. Supplementary Figure S7a shows the distribution of the standard deviation of the measured average charge of the

same phosphopeptides over multiple measurements (only for phosphopeptides measured at least twice). The distribution is wider than for the non-phosphorylated peptides (compared to Supplementary Figure S3b) with only 84% of the phosphopeptides showing a standard deviation below 0.1. Since it is reasonable to assume that the precision of the CSD extraction computational pipeline is the same as for the non-phosphorylated peptides, the wider distribution might indicate a lower data quality of the phosphopeptide data sets. The latter can stem from greater challenges in the identification of phosphorylated peptides (and specifically localization of phosphate moiety) compared to non-phosphorylated ones. The distribution of the average observed charge is however quite like the non-phosphorylated dataset, indicating that it is mostly influenced by the tryptic nature of the peptides in the dataset. Similarly, the relationship between the peptide mass, the observed average charge, and the number of basic sites is close to the regular (non-phosphorylated) cases (Supplementary Figure S8).

As mentioned, one of the common challenges in phosphopeptide characterization is the proper assignment of the position of the phosphoric acid moiety. It is, however, expected that phosphopeptide isoforms (phosphopeptides with the same sequence, but different positions of phosphorylation) can obtain different gas-phase structures resulting in different representations of poly-protonated ions). We could identify 1693 groups of phosphosite isomers. For the members of each of these groups, we calculated the largest observed average charge difference and smallest cosine similarity. The corresponding distributions are presented in Figure 2. While the majority of phosphopeptides demonstrate relatively small changes in the CSD quantities when the position of the phosphorylations is changed, it is possible to differentiate at least one of the isoforms from the others in about 8.6%, and 25.6% of the cases using the cosine similarity and average observed charge, respectively. The sensitivity is comparable to the case of sequence isomer peptides (see Figure 1c-d). The same properties plotted for peptides having a specific number of phosphosites are presented in Supplementary Figure S9.

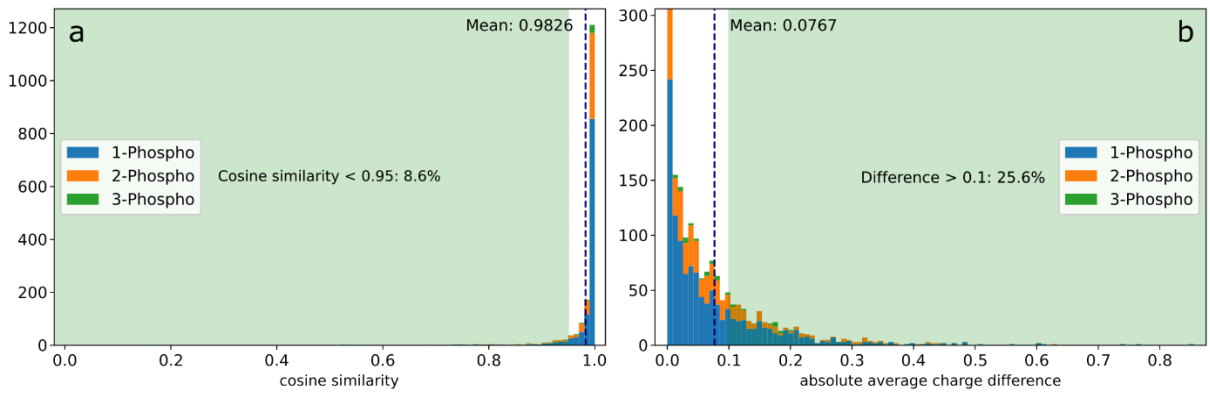

Figure 2: (a) Smallest cosine similarity and (b) largest absolute average observed charge difference between phosphopeptide isoforms. Parts of the distribution are colored according to the number of phosphosites present, and the combined quantities are calculated for the complete distribution.

To study in more detail the impact of peptide phosphorylation, we investigated the changes in CSD quantity between phosphorylated peptides and their non-phosphorylated counterparts. The CSD values for the corresponding non-phosphorylated peptides were extracted from the K562 dataset described above. The pairwise differences between phosphorylated (with all possible positions and number of PTM) and non-phosphorylated peptides are presented in Figure 3a-b. Firstly, around 30% of peptides demonstrate cosine similarity below the 0.95 threshold, confirming a significant influence of phosphorylation on the charge state distribution of peptides. Secondly, when comparing the shift in charge state between phosphorylated and the same non-phosphorylated peptides, two distinct

populations were observed. One of them, comprising 18.3% of peptides, showed an overall increase in charge (> 0.1 units) upon phosphorylation. The other distribution, representing 29.5% of peptides, exhibited a decrease (< 0.1 units) in overall charge. To investigate this phenomenon further we have extracted sequence motifs (up to seven amino acids flanking the phosphorylation site) for the two different phosphopeptide populations and analyzed which amino acids are overrepresented/underrepresented (Figure 3c-d). Only singly phosphorylated peptides were considered for this analysis to avoid possible ambiguity in phosphosite influence. Phosphorylated peptides with a higher average charge than their non-phosphorylated counterparts were found to lack basic residues in the vicinity of the phosphorylation site. Conversely, phosphorylated peptides with a reduction in average charge displayed a higher proportion of basic residues near the phosphorylation site.

The population exhibiting a negative shift in average charge state is suggested to be due to these phosphopeptides' greater propensity to adopt a salt bridge structure. This notion is further corroborated by ion mobility mass spectrometry studies on phosphopeptides, which explained the negative shift in the ion collisional cross-section (CCS) by enhanced intramolecular interactions of the phosphate motif with functional groups of the peptide.[14] This compression of the peptide conformational space and the suggested establishment of salt bridges both contribute to the observed reduction in the average charge state of phosphorylated peptides in the gas phase. This occurs because functional groups capable of stabilizing a positive charge may become less accessible for protonation in gas-phase ions. In addition, the compression of conformational space can bring charged groups within the peptide closer together, leading to significant Coulombic repulsion that hinders the acquisition of greater charge density through further protonation. Finally, the formation of a salt bridge effectively neutralizes a positive charge, reducing the peptide's net charge.

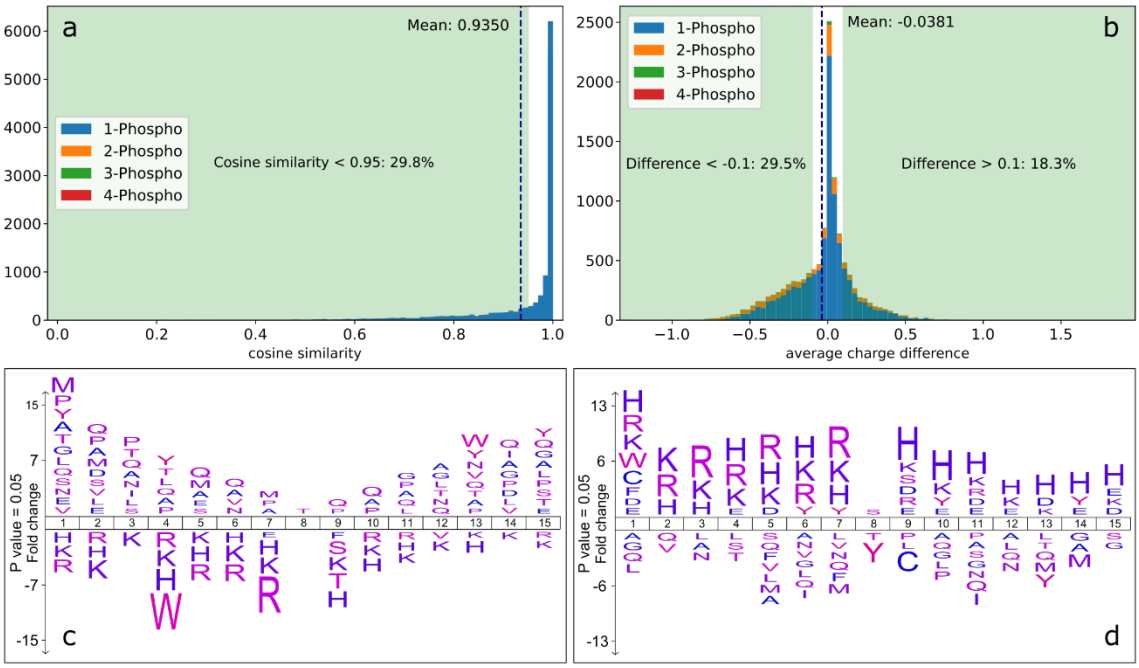

Figure 3: (a) cosine similarity and (b) average observed charge difference between pairs of non-phosphorylated and phosphorylated peptides with the same sequence. Parts of the distribution are colored according to the total number of phosphosites present. IceLogo (seven amino acids up- and downstream from phosphorylation site) for peptides demonstrating positive (c) and negative (d) shift in average observed charge upon phosphorylation. Only singly-phosphorylated peptides have been employed for the analysis.

We anticipate that the limitation of the peptide conformational space should also be evident as a CSD with a narrower spread. This is because such types of phosphopeptides would adopt fewer distinct conformations, thereby limiting the possibility of charging to only a few potential sites. In Figure 4, the cumulative frequency plot reveals that the shift in average charge state upon phosphorylation indeed produces contrasting effects in the two populations of phosphopeptides. For the group experiencing a negative shift in charge due to suggested salt-bridge formation (depicted in blue), there is a trend toward a more dominant single-dominating charge state. For instance, 30% of dephosphorylated peptides have a major single charge constituting 80% or more of the ion abundance, and this figure rises to 40% after phosphorylation. This suggests that phosphopeptides in this group have a narrower charge distribution. Conversely, for the group represented in orange (no salt bridge), the situation is reversed. In this group, the proportion of peptides with more than 80% single charge abundance is 55% for non-phosphorylated and decreases to 45% for phosphopeptides.

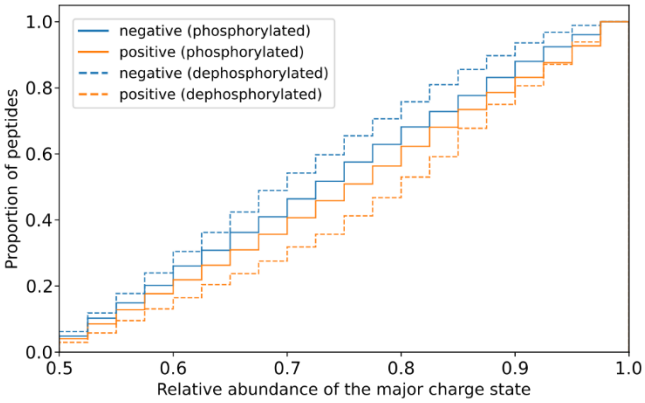

Figure 4. Cumulative histogram for major charge state relative abundance of phosphopeptides demonstrating positive and negative shifts in average observed charge state. Solid lines correspond to phosphopeptides, while dashed lines – to their dephosphorylated counterparts. Peptides observed only in a single charge state were excluded.

Regarding the average charge state increase of phosphorylated peptides, it is to the best of our knowledge a novel observation. Although some phosphopeptides can form salt bridges upon phosphorylation, leading to compaction, this is not the most common outcome. This is supported by one observation from larger-scale IMS studies indicating that for a significant portion of phosphorylated peptides (approximately 66%), the collision cross-section increases[18]. As with other chemical modifications to peptides, it is possible that the presence of a phosphate motif (for this group we assume no involvement in a salt bridge structure) can alter the conformational space of the peptide and thereby expose and possibly promote the accessibility of certain ionizable sites, influencing the distribution of charges in the gas phase. However, such a claim should be further substantiated by other complementary data, such as molecular dynamics simulations.

Our analysis also includes an examination of the number of basic sites, peptide mass, hydrophobicity, relative phosphosite position, and charge for both populations of peptides (Supplementary Figure S10). The peptides demonstrating an increase in the observed average charge tend to have fewer basic sites in general and more specifically lower basic site density (i.e. number of basic sites per unit of mass). Thus, it is expected that these peptides are less prone to form tightly bridged compact structures, due to the lack of positive charge donors. As mentioned earlier, an increase in the average charge can be associated with a larger CCS. Secondly, one can observe that peptides displaying average charge increase tend to be more hydrophobic (larger GRAVY score) – the latter is reported to promote larger charge states[16] partly due to an increased propensity to be present at the surface of

the electrospray droplets, where charging primarily occurs[19]. Next, peptides with the negative shift in average charge, demonstrate a slight preference for N-terminal phosphorylation. A similar observation was also made in large-scale IMS data[18]. Finally, peptides with the positive charge shift have an average charge larger than the number of basic sites in the sequence, thus, such "excess" of positive charge might promote mitigating the negative charge introduced during phosphorylation.

All the above observations made from the CSD are very intriguing as they present a new approach to classifying and characterizing molecular structures of phosphorylated peptides on a large scale. The consistency and robustness of the charge state feature justify such a methodology for differentiating, for instance, phosphopeptides with a predominant salt bridge structure from those that do not form one. Classification of phosphopeptides demonstrating extensive intramolecular interactions has, until now, only been demonstrated experimentally in the gas phase using ion collision cross-section measurements[11,14,18] and electron capture dissociation[20,21]. Moreover, it is reasonable to assume that the CSD should reflect the formation of all types of salt bridges, even if this does not lead to significant conformational changes, and, thus, potentially should be a more sensitive tool to probe for such structures. For instance, changes to the CCS can be expected to be only moderate in phosphopeptides where the salt bridge is formed by residues in proximity. Additionally, reports have documented that limited conformational changes are observed upon phosphorylation in cases where the structure of the non-phosphorylated peptide is already compact[11].

The potential of the CSD dimension in proteomics

For a more precise evaluation of the discriminative power of the CSDs, we categorized the data from the K562 dataset into bins based on peptide properties, including theoretical mass, retention time (apex of the elution profile), and average observed charge. We employed various mass intervals (2, 5, and 10 ppm), retention times (0.5, 1, and 2 min), and charge bins (0.1, 0.2, and 0.4). The width of the bins has been chosen based on a typical performance of the proteomics experiment, the expected width of the peptide elution profile (dataset uses 63 min gradient), and the reproducibility of charge-state features determined above. Figure 5 shows the distribution of the number of unique peptides falling into a single bin, depending on the subset of criteria used (distributions with all possible combinations of parameters are provided as supplementary material). Using the accurate mass alone with a narrow 2 ppm bin width allows having a single peptide in a bin in 28% of the cases, such a high mass accuracy is difficult to reach under routine LC-MS analysis while using more common 5 ppm and 10 ppm binning lowers the number of single peptide bins to 13% and 6%, respectively. Employing CSD information together with mass, i.e. a bin is determined in mass and CSD dimension, provides a significant increase in the number of single peptide bins (58% and 47% single peptide bins for 5 and 10 ppm, respectively) highlighting the orthogonality of this dimension. Using retention time and mass bins dramatically improves partitioning of the peptide search space – over 90% of peptides can be separated using 2 ppm mass bins with 0.5 min retention time bin. Such a level of orthogonality between retention time and mass is well-known and, thus, the retention time dimension is currently actively employed in the state-of-the-art proteomics data analysis pipelines[22–24]. While direct infusion was among the early techniques used in proteomics[25], it has gained renewed attention in recent research[26,27]. Our data shows that utilizing CSD information substantially aids in the discrimination of peptide candidates, although not to the same degree as the retention time dimension. While retention time and mass information allow superior partitioning of the peptide space, including the CSD dimension further adds small but consistent improvement in discrimination across all simulated conditions (red bars in Figure 5). The effect of charge information is (as expected) more pronounced in cases when retention time and mass cannot provide almost complete separation (for example in

the 10-ppm case). Our analysis suggests that CSD information is a useful addition and should allow for improved discrimination of peptides in LC-MS data analysis in general.

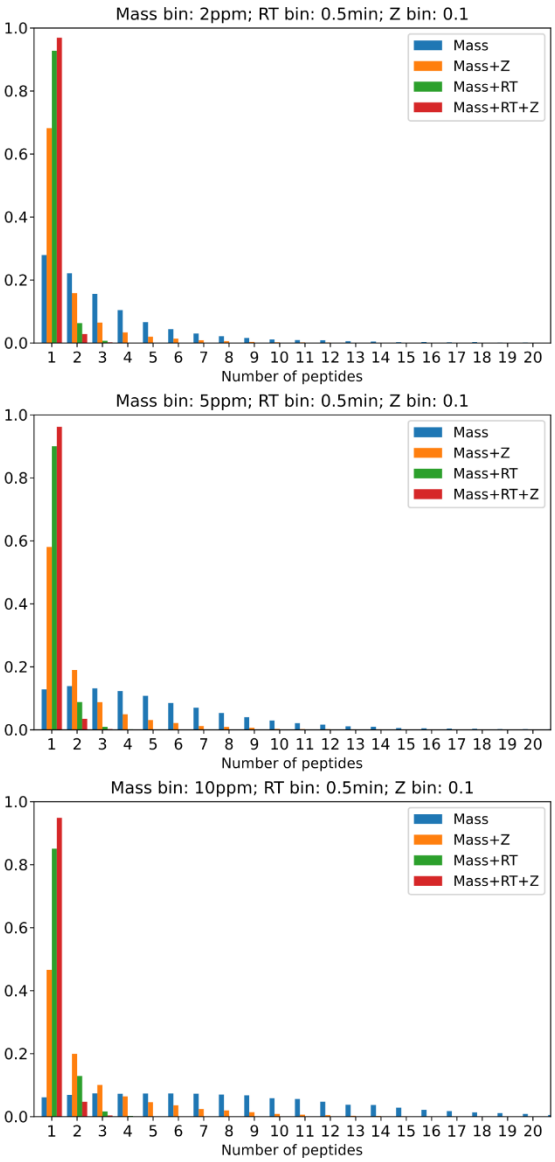

Figure 5. Distribution of the number of peptides in a single bin. Mass bin: top – 2ppm, middle – 5ppm, bottom – 10ppm; RT bin 0.5min; Z bin 0.1. Binning by mass (blue), mass and charge (orange), mass and retention time (green), mass, retention time, and charge (red).

Building CSD prediction model

The evident consistency in the reproducibility of peptide CSD permits the development of a predictive model for this attribute. The K562 dataset (non-phosphorylated peptides) and phosphopeptides dataset described above were combined and used to train a deep-learning prediction model. The architecture of the model has been published earlier[15], but the output dimension layer has been changed to six to accommodate the CSD vectors calculated earlier. Since the input layer of the model has a dimension of 40, peptides longer than 40 elements (after encoding) were removed from the dataset. A holdout test set consisting of 27740 peptide sequences has been randomly selected from the dataset, the remaining part (249655 peptides) has been used during training and validation. The prediction error (negative cosine similarity) calculated for the holdout set was –0.988921 and –0.985018 for all peptides and only phosphopeptides, respectively. Altogether, the model

demonstrated a level of prediction accuracy in line with the reproducibility of the previously observed CSD data. A slightly worse error for phosphopeptides could be due to the worse quality of the CSD data itself (such as erroneous phosphosite localization, see the discussion in the corresponding section) or just since phosphopeptides were a minor peptide class, representing about 15% of peptides in the dataset. Since we have observed that the CSD of sequence isomers were generally well correlated our initial predictor used simpler and more explainable models, using peptide amino acid composition as the input. Simpler models, however, demonstrated significantly lower prediction accuracy (summarized in Table 2 and Supplementary Figure S11). Thus, we can conclude that amino-acid interactions (that are not accounted for in the composition-based input) play an important role in the observed CSD. For our further investigation, we have used the deep learning model.

Table 2. Prediction error (mean cosine similarity) for different machine learning models

Model	Training set error	Test set error
Linear regression	0.8798	0.8790
K-nearest neighbors (K=7)	0.9126	0.8813
Decision tree	0.9997	0.9321
Generalized additive model (per charge state)	0.8933	0.8935
Deep learning (LSTM)	0.9902	0.9889

Investigation of the discrimination ability of the trained CSD predictor using complete human *in silico* digestion is provided in Supplementary Text 1

Using CSD for DirectMS1 and MS/MS approach

Inspired by the results obtained from experimental and theoretical datasets, we implemented the CSD dimension in various proteomics data analysis approaches. To begin with, we utilized a recently developed approach called DirectMS1. DirectMS1 is a rapid proteomics screening method relying solely on the survey scan data (no fragmentation) and allowing more than 2000 proteins to be identified from 200 ng of cell lysate in approximately eight minutes per sample (180 samples per day)[28]. In this iteration, we integrate our CSD model into the DirectMS1 workflow by extending the set of peptide features used in the last machine-learning-based assessment step with the predicted CSD features. The following three approaches were tested: 1) standard DirectMS1, i.e. no CSD features added; 2) reduced CSD, only the features denoting the number of observed charge states (i.e. number of non-zero elements in CSD vector) and the number of times a particular peptide has been observed in the experiment added; 3) full CSD, features denoting the correlation between predicted and experimental CSD added. The reduced CSD approach served as additional control, to assess the importance of CSD prediction features. We have used several LC-MS experiments acquired with the DirectMS1 approach (a detailed description can be found in the Supplementary Tables) focusing on the number of proteins identified with 1% FDR for every tested experiment (Figure 6a). All experiments except the ones utilizing FAIMS separation demonstrate an increase (although very moderate) in the number of identified proteins. The reduction in performance by the CSD dimension in FAIMS-enabled experiments was expected, since it is known that ion mobility separation will separate different charge states of the same peptide and, thus, distort CSD, making the predictions inaccurate. While the reduced CSD feature set offers gains in detecting more proteins, the complete CSD feature set typically excels even further in performance.

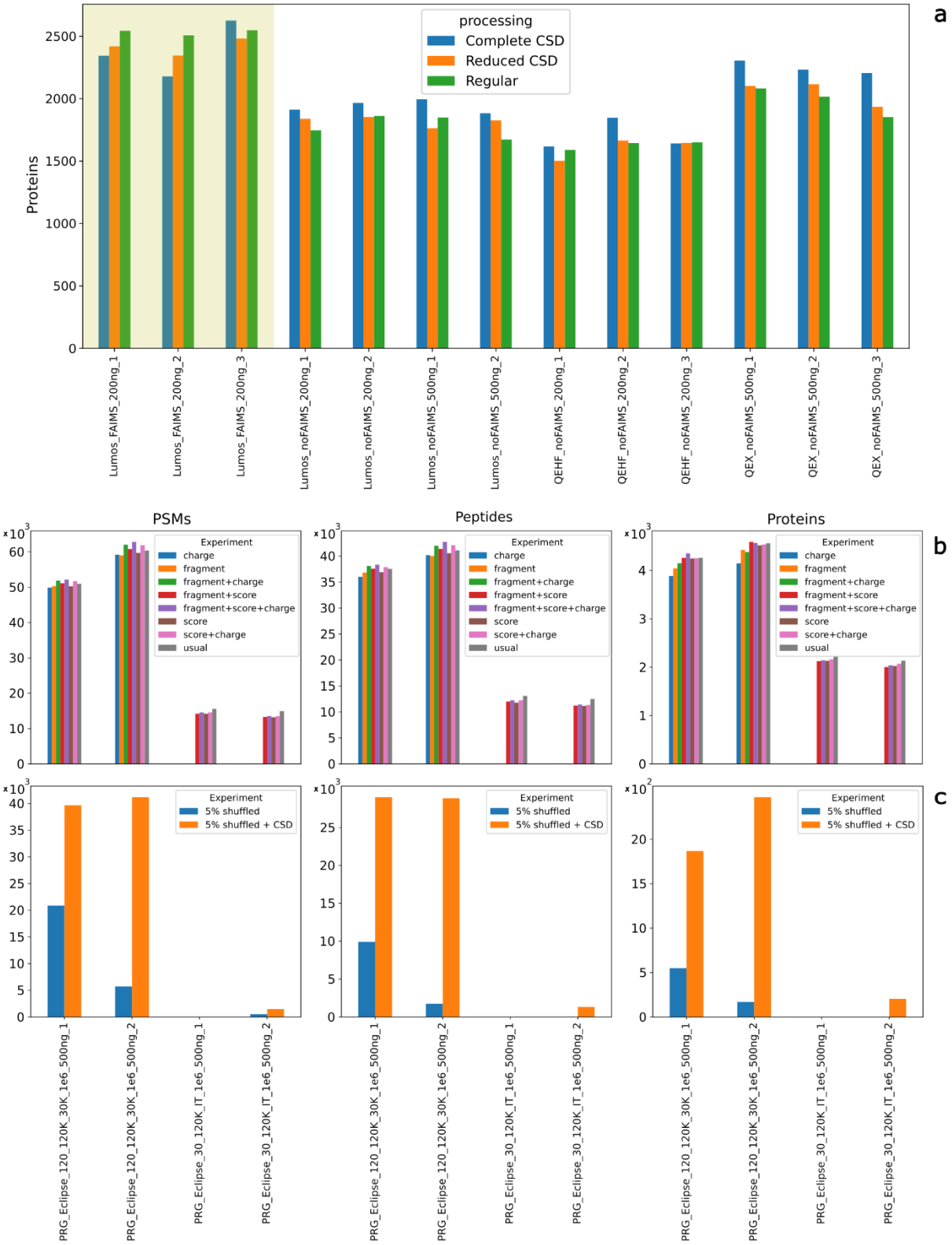

Figure 6. (a) The number of proteins identified at 1% FDR with the DirectMS1 approach with different feature sets used in the machine-learning step. Shaded region – LC-MS runs employing online FAIMS separation. Identified PSMs, peptides, and proteins, using different feature sets in Percolator (b), and randomly shuffling 5% of scores (c).

Finally, we integrated CSD features into the conventional MS/MS shotgun proteomics framework. Several LC-MS/MS experiments (details can be found in the supplementary tables) were analyzed

using MSGF+ and Percolator using either a standard feature set or a few other feature sets. The default features used by Percolator were split into several classes: score – features related to MSGF+ scores; fragment – features corresponding to fragmentation, and all other features. The charge feature set contains extra features calculated from the predicted CSDs (the complete list of the features and their classification can be found in Supplementary Table 1). Lastly, we always included a simple matching score (i.e. the number of matched theoretical fragment ions, and the number of match ions divided by the sequence length) for every peptide. Identification performance on PSM, peptide, and protein levels, when different sets of features were employed, are presented in Figure 6b. The observations differ for experiments employing ion trap (low resolution) and Orbitrap (high resolution) for fragmentation spectra acquisition. For low-resolution experiments using MSGF+ score features was essential for successful identification, a simplistic score even when supplemented with charge or fragmentation features doesn't allow sufficient separation of target and decoy populations. With the Percolator input including score features identification performance returns to a level slightly below the regular one. It should be noted that the inclusion of the charge dimension allows for a slight identification improvement. For the high-resolution experiments, on the contrary, simplistic scores in combination with charge or fragment features demonstrate a performance close to the regular feature set. Using charge features results in a more visible increase in identification while using score, fragmentation, and charge features in combination outperforms the regular approach. To further investigate the discrimination ability of charge features we randomly shuffled the score features for 5% of entries in the Percolator input and discarded all other features, the remaining 95% were preserved in the correct order, thus, diminishing their ability to serve as a good discriminator, the charge features, however, were included in the correct order (without shuffling). The results of this experiment are presented in Figure 6c. Shuffling of just 5% of the Percolator features compromises identification ability almost completely. Employing charge features allows reviving some of the performance, although even in the best cases it is significantly lower compared to regular feature set.

The combined results of the MS/MS identification experiments demonstrate the CSD features contain valid information allowing for the discrimination of correct and incorrect peptide candidates. The discriminating power is, however, incomparable with MS/MS-based information, and thus, alone, cannot provide efficient separation of target and decoy populations.

Conclusions

We have conducted a detailed investigation of the peptide charge state dimension to explore its analytical utility in proteome research. When comparing datasets from different laboratories, instruments, and time points, we found that CSDs are highly reproducible for both regular and phosphorylated peptides. This reproducibility suggests the possibility of using this characteristic to distinguish peptides. While the sensitivity of CSD to *sequence* and *PTM isomers* of a peptide is not particularly high, CSD is quite sensitive to *elemental isomers*. Clustering peptides by their average observed charge or by the similarity of their CSD offers strong discrimination. Using mass and observed charge alone, close to 70% of peptides from a large experimental dataset can be separated, while retention time and mass allow for over 95% of peptides to be distinguished. As a result, CSDs can be an attractive dimension to utilize in direct infusion proteomics data analysis. Notably, CSD and retention time provide somewhat complementary separation; using both consistently yields deeper proteome coverage. Although the gain is moderate, it does not require additional experiments or equipment. A deep-learning predictor for CSD has been developed, demonstrating excellent performance (cosine similarity on a separate test set: 0.989) based on approximately 250,000 experimental CSDs. The trained model was tested in both MS1-only (DirectMS1) and regular MS/MS (MSGF+ and Percolator) searches, proving to enhance the discriminative power in both contexts.

By studying CSDs of phosphopeptides and their unmodified counterparts we can confidently detect two peptide populations with opposite behavior upon phosphorylation, namely, the increase or decrease in the observed average charge. The difference in the phosphopeptide CSD behavior is believed to be due to the propensity to form a salt bridge structure. The peptides showing an increase in the average charge tend to have fewer basic amino acids in the vicinity of phosphosite and a lower number of basic sites per unit of mass in general and, thus, are less prone to form salt bridges between the phosphosite and basic amino acid. That previously unknown feature can assist in the investigation of gas-phase structures of phosphorylated peptides. The discovery of this unnoticed feature holds the potential for sensitive characterization of gas-phase structures of phosphorylated peptides.

In conclusion, we have demonstrated that the CSD dimension, which is currently underutilized in proteomics despite its readily available nature, holds the potential for improving proteomics analyses and efficiently characterizing different phosphopeptide structures.

Methods

Software tools

Unless stated otherwise, data manipulation and plotting were performed in Python (3.10.2) with the following modules installed: ***numpy*** (1.23.5)[29], ***scipy*** (1.10.0)[30], ***matplotlib*** (3.5.1)[31], ***pandas*** (1.4.0)[32], ***scikit-learn*** (1.0.2)[33], ***seaborn*** (0.11.2)[34], ***pyteomics*** (4.6)[35], ***keras*** (2.10.0)[36], ***statsmodels*** (0.13.1)[37], ***pygam*** (0.8.0)[38]. Relevant scripts are provided as supplementary material.

Database search (non-phosphorylated)

Raw files were searched using MSGF+ (2022.01.07)[39] against the UniProtKB human protein database supplemented with reversed decoy sequences. Trypsin was used as a digestion enzyme. Parent ion mass accuracy was 10 ppm and isotopic error of 0 and 1, precursor charge from 2 to 6 (both included), theoretical peptide length from 6 to 40 (both included), carbamidomethylation of cysteine as the fixed modification, oxidation of methionine and acetylation of protein N-terminus as variable ones. Instrument protocol and enzyme were set according to the LC-MS experiment. Search results were validated using Percolator (3.05)[40], q-values and PEPs calculated by Percolator were added to the resulting mzIdentML file, and PSMs were filtered to q-value < 0.01 for any further processing.

Database search (phosphorylated)

Deposited MaxQuant results were used for phosphoproteomics datasets when available. Raw files belonging to the PXD000138 dataset (synthetic phosphopeptides) were reanalyzed using MaxQuant (2.0.3.0)[41]. The search was performed against the deposited FASTA file. Trypsin has been selected as an enzyme and up to two missed cleavages were allowed. Oxidation of methionine and phosphorylation of serine, threonine, and tyrosine were used as variable modifications. The minimal peptide length was 7, and the maximal peptide mass was 4600. Results were filtered to 0.01 FDR for proteins, peptides, and sites. All other parameters were set to their default values.

Phosphoproteomics datasets analysis

MaxQuant results (msms.txt file) were read using ***psm-utils*** (0.2.3)[42]. PSMs with all phosphorylated sites having a localization score above 0.95 and Andromeda score above 50 were considered for further processing. For the synthetic phosphopeptide datasets, peptides matching the synthesis goal were considered valid and accepted for further processing. Since msms.txt files contain only calibrated mass values, corresponding evidence.txt files were used to calculate the observed (non-calibrated) mass values matching the features detected in the raw files.

Charge-state distribution calculation

Raw files were converted to mzML format using ***ThermoRawFileParser*** (1.4.2)[43] with default parameters and further processed by ***biosaur2*** (0.2.11)[44] to detect LC-MS features. The minimal length of a hill was set to 2, and the mass accuracy for hills and isotopes was set to 8. Peptide-spectrum matches (PSMs) were mapped to detected LC-MS features, by the following criteria: spectrum retention time has to be between the start and the end of LC-MS feature elution, the charge of the PSM has to match the charge of the LC-MS feature, and the absolute difference between peptide mass and the feature mass must be below 5 ppm. Next, unmatched features coeluting with the matched ones (i.e. overlapping retention time) and having similar mass (absolute difference below 5 ppm) were grouped. The averaged feature mass, the average elution apex, the lowest start, and the largest end time were extracted for every feature group and peptide sequence. If more than one peptide sequence was matched to a feature group, each of the peptide sequences had the same result associated with it. The masses were used to calculate theoretical monoisotopic m/z values for all charge states from 1 to 6 (including both ends), m/z values falling outside the m/z range of the LC-MS file were discarded. ThermoRawFileParser was used to produce extracted ion chromatograms (XICs) for every m/z value, using a 5-ppm mass window, and retention time borders determined earlier for every peptide. The area under the curve (AUC) was calculated for every XIC and, finally, XICs belonging to the same peptide were assembled into a vector representing abundances of the corresponding charge states (raw CSDs). Raw CSDs were calculated for every LC-MS run in the dataset and the results merged for the complete dataset. Next, all raw CSDs were normalized (i.e. divided by the sum of all XIC areas), and complete duplicates (i.e. the same peptide sequence, and normalized charge state abundance) were discarded. CSDs belonging to the same peptide sequence were grouped and the consensus CSD was calculated: cosine similarity was calculated for all members of the group and only the subgroup having cosine similarity above 0.9 for all its members was used for the consensus CSD calculation by averaging all individual CSDs; groups with no correlated samples were discarded; groups with single member were used without any processing. Consensus CSDs for every unique peptide sequence were stored for every dataset.

CSD prediction model

Peptide sequences in a dataset were encoded into single-letter codes, with special codes used for N-terminal acetylation, oxidated methionine, and phosphorylated serine, threonine, and tyrosine. Cysteine was always considered carbamidomethylated and, thus, was encoded as cysteine. Peptides longer than 40 after encoding were discarded. For the deep learning model, peptide sequences were one-hot encoded, i.e. the input matrix for every peptide was [40, 25], while for all other models, the input was a vector [1, 25] representing the number of amino acids of the corresponding type. The complete dataset was split into training and test parts (10% of the complete dataset) (*train_test_split* from ***scikit-learn***). The following model classes were used for the prediction: *LinearRegression*, *KNeighborsRegressor* (with K = 7), *DecisionTreeRegressor*, from ***scikit-learn***, *GAM* (with lambda = 0.01) from ***pygam***, in the latter case 6 regressors were joined using *MultiOutputRegressor* from ***scikit-leam***. All other parameters were set to their default values. Deep learning model architecture was described earlier[15], briefly, the network consisted of a sequentially connected masking layer, two bidirectional LSTM layers, and two dense layers, the latest one had an output dimension of 6. The model was compiled with an Adam optimizer and negative cosine similarity loss function. The training was performed on the computer equipped with NVIDIA A100 GPU running ***tensorflow***/***keras*** 2.12.0 under Python 3.8.10. The training was performed with a batch size of 512, for a maximum of 100 epochs and early stopping monitoring loss on a validation set (validation split 0.2) and patience of 5. Three models using 10%, 30%, and 100% of the data selected at random were trained. Since a monotonous decrease

in the prediction error (calculated on the holdout set) has been observed, the final model (trained on the complete dataset) was used.

DirectMS1 analysis

Raw files were converted to mzML using ***ThermoRawFileParser*** (1.4.2) and, next, ***biosaur2*** (0.2.11) has been used for feature detection. The minimal length of a hill was set to 1, and the mass accuracy for hills and isotopes was set to 8. The search was performed by ***ms1searchpy***[28] against the SwissProt human protein database, concatenated with decoy sequences generated by pseudo-shuffling[45]. The following parameters were used: minimum number of scans was 1, minimum number of isotopes was 2, no missed cleavages, minimum precursor charge was 1, initial precursor tolerance was 8 ppm, protein FDR was 1%, two staged RT training, employing additive model first and DeepLC model next, minimal candidate peptide length was 6. CSD-specific features were included in the feature table used in the machine-learning-based PFM classification. For the reduced CSD experiment, the following features were included: the number of times a peptide sequence is observed in the PFM list; the number of unique charge states a peptide sequence is observed in; the number of unique ion mobility values a peptide sequence is observed in. For the complete CSD experiment, a CSD prediction step has been implemented in ***ms1searchpy*** (both reduced and complete options are included in version 2.6.3). To this end all PFMs were collected and m/z values corresponding to charge states from 1 to 4 (both included) were calculated; m/z values outside the m/z range of the instrument were set to zero; ThermoRawFileParser (1.4.2) was used to produce XICs using the calculated m/z values, 5 ppm tolerance window, and retention time borders of each LC-MS feature; AUC has been calculated for each XIC and these values were assembled to experimental CSD vectors; the model trained earlier was employed to predict CSDs for every unique peptide sequence; charge states corresponding to m/z values outside instrument m/z range were set to zero; both experimental and predicted CSD vectors were normalized; average observed charge state was calculated for both experimental and predicted CSDs as discussed earlier. The following properties were calculated for every PFM and added to the table for the machine-learning-assisted classification: spectral angle between the experimental and predicted CSD, the difference between experiment and prediction of each charge state, and average observed charge all expressed as z-score, i.e. number of standard deviations from the mean error of each property error. Note, that in the current implementation, the LC-MS file in vendor RAW format and ThermoRawFileParser (version 1.4.2 or later) are necessary to perform the XIC extraction.

MS/MS analysis with CSD

Raw files were searched using MSGF+ and Percolator and LC-MS features were calculated as described earlier. The Percolator input file was modified to include CSD-related features for each entry in it. Each entry in a Percolator input represents PSM (both target and decoy and without any quality filtering) and, thus, experimental CSD calculation followed the same basic steps as described earlier. Briefly, PSMs were mapped to LC-MS features, matched features were grouped by co-elution and mass, m/z values for charges from 1 to 6 were calculated, and, lastly, corresponding XICs were extracted using ThermoRawFileParser and AUCs assembled to experimental CSDs. The deep learning model was employed to predict CSDs for each unique peptide sequence, and later entries belonging to m/z values outside the instrument m/z range were zeroed and CSDs normalized. Errors between the experimental and predicted values of each charge state relative abundance and observed average charge represented as z-score were used as features for Percolator (feature names: *1*, *2*, *3*, *4*, *5*, *6*, *z*). For PSMs lacking experimental CSD, an average value of the corresponding features was used (for target and decoy populations separately). Finally, the simple matching score was calculated as two extra features: *nMatch* – the number of theoretical fragments matched in the fragmentation spectrum, and *cover* – *nMatch* divided by peptide length. A separate input file was generated for every combination

of employed features and submitted to Percolator. Another set of input files was created by shuffling randomly score-related features with and without charge features.

Acknowledgment

This study was supported by the Lundbeck Foundation (Grant R346-2020-1215 to F.K.). Proteomics and mass spectrometry infrastructure at the University of Southern Denmark SDU were supported by generous grants to the VILLUM Center for Bioanalytical Sciences (VILLUM Foundation grant no. 7292), PRO-MS: Danish National Mass Spectrometry Platform for Functional Proteomics (grant no. 5072-00007B), and the Novo Nordisk Foundation (INTEGRA, NNF20OC0061575). The authors would like to thank Dr. Mark Ivanov for his valuable help in implementing and testing the DirectMS1-CSD approach.

Supporting Information

The following supporting information is available along with the manuscript

- Supporting figures and tables, additional experimental data (PDF)
- Description of all datasets used in the manuscript (XLSX)
- Scripts (Python) used to analyze the data and generate results and intermediate figures (ZIP)

References

(1) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science* **1989**, *246* (4926), 64–71. https://doi.org/10.1126/science.2675315.

(2) Kaltashov, I. A.; Fenselau, C. C. A Direct Comparison of "First" and "Second" Gas Phase Basicities of the Octapeptide RPPGFSPF. *J. Am. Chem. Soc.* **1995**, *117* (39), 9906–9910. https://doi.org/10.1021/ja00144a017.

(3) Kjeldsen, F.; Savitski, M. M.; Adams, C. M.; Zubarev, R. A. Determination of the Location of Positive Charges in Gas-Phase Polypeptide Polycations by Tandem Mass Spectrometry. *Int. J. Mass Spectrom.* **2006**, *252* (3), 204–212. https://doi.org/10.1016/j.ijms.2005.10.009.

(4) Johnson, R. S.; Martin, S. A.; Biemann, K. Collision-Induced Fragmentation of (M+H)+ Ions of Peptides. Side Chain Specific Sequence Ions. *Int. J. Mass Spectrom. Ion Process.* **1988**, *86*, 137–154. https://doi.org/10.1016/0168-1176(88)80060-0.

(5) Dongré, A. R.; Jones, J. L.; Somogyi, Á.; Wysocki, V. H. Influence of Peptide Composition, Gas-Phase Basicity, and Chemical Modification on Fragmentation Efficiency: Evidence for the Mobile Proton Model. *J. Am. Chem. Soc.* **1996**, *118* (35), 8365–8374. https://doi.org/10.1021/ja9542193.

(6) Baldwin, R. L. Energetics of Protein Folding. *J. Mol. Biol.* **2007**, *371* (2), 283–301. https://doi.org/10.1016/j.jmb.2007.05.078.

(7) Konermann, L.; Aliyari, E.; Lee, J. H. Mobile Protons Limit the Stability of Salt Bridges in the Gas Phase: Implications for the Structures of Electrosprayed Protein Ions. *J. Phys. Chem. B* **2021**, *125* (15), 3803–3814. https://doi.org/10.1021/acs.jpcb.1c00944.

(8) Bakhtiari, M.; Konermann, L. Protein Ions Generated by Native Electrospray Ionization: Comparison of Gas Phase, Solution, and Crystal Structures. *J. Phys. Chem. B* **2019**, *123* (8), 1784–1796. https://doi.org/10.1021/acs.jpcb.8b12173.

(9) Loo, R. R. O.; Loo, J. A. Salt Bridge Rearrangement (SaBRe) Explains the Dissociation Behavior of Noncovalent Complexes. *J. Am. Soc. Mass Spectrom.* **2016**, *27* (6), 975–990. https://doi.org/10.1007/s13361-016-1375-3.

(10) Jackson, S. N.; Wang, H.-Y. J.; Yergey, A.; Woods, A. S. Phosphate Stabilization of Intermolecular Interactions. *J. Proteome Res.* **2006**, *5* (1), 122–126. https://doi.org/10.1021/pr0503578.

(11) Ruotolo, B. T.; Verbeck; Thomson, L. M.; Woods, A. S.; Gillig, K. J.; Russell, D. H. Distinguishing between Phosphorylated and Nonphosphorylated Peptides with Ion Mobility–Mass Spectrometry. *J. Proteome Res.* **2002**, *1* (4), 303–306. https://doi.org/10.1021/pr025516r.
(12) Ruotolo, B. T.; Gillig, K. J.; Woods, A. S.; Egan, T. F.; Ugarov, M. V.; Schultz, J. A.; Russell, D. H. Analysis of Phosphorylated Peptides by Ion Mobility-Mass Spectrometry. *Anal. Chem.* **2004**, *76* (22), 6727–6733. https://doi.org/10.1021/ac0498009.
(13) Thalassinos, K.; Grabenauer, M.; Slade, S. E.; Hilton, G. R.; Bowers, M. T.; Scrivens, J. H. Characterization of Phosphorylated Peptides Using Traveling Wave-Based and Drift Cell Ion Mobility Mass Spectrometry. *Anal. Chem.* **2009**, *81* (1), 248–254. https://doi.org/10.1021/ac801916h.
(14) Glover, M. S.; Dilger, J. M.; Acton, M. D.; Arnold, R. J.; Radivojac, P.; Clemmer, D. E. Examining the Influence of Phosphorylation on Peptide Ion Structure by Ion Mobility Spectrometry-Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2016**, *27* (5), 786–794. https://doi.org/10.1007/s13361-016-1343-y.
(15) Guan, S.; Moran, M. F.; Ma, B. Prediction of LC-MS/MS Properties of Peptides from Sequence by Deep Learning. *Mol. Cell. Proteomics* **2019**, *18* (10), 2099–2107. https://doi.org/10.1074/mcp.TIR119.001412.
(16) Xu, A. M.; Tang, L. C.; Jovanovic, M.; Regev, O. *A High-Throughput Approach Reveals Distinct Peptide Charging Behaviors in Electrospray Ionization Mass Spectrometry*; preprint; Systems Biology, 2023. https://doi.org/10.1101/2023.03.31.535171.
(17) Sinitcyn, P.; Richards, A. L.; Weatheritt, R. J.; Brademan, D. R.; Marx, H.; Shishkova, E.; Meyer, J. G.; Hebert, A. S.; Westphall, M. S.; Blencowe, B. J.; Cox, J.; Coon, J. J. Global Detection of Human Variants and Isoforms by Deep Proteome Sequencing. *Nat. Biotechnol.* **2023**. https://doi.org/10.1038/s41587-023-01714-x.
(18) Ogata, K.; Chang, C.-H.; Ishihama, Y. Effect of Phosphorylation on the Collision Cross Sections of Peptide Ions in Ion Mobility Spectrometry. *Mass Spectrom. Tokyo Jpn.* **2021**, *10*, A0093. https://doi.org/10.5702/massspectrometry.A0093.
(19) Cech, N. B.; Enke, C. G. Practical Implications of Some Recent Studies in Electrospray Ionization Fundamentals. *Mass Spectrom. Rev.* **2001**, *20* (6), 362–387. https://doi.org/10.1002/mas.10008.
(20) Kim, D.; Pai, P.-J.; Creese, A. J.; Jones, A. W.; Russell, D. H.; Cooper, H. J. Probing the Electron Capture Dissociation Mass Spectrometry of Phosphopeptides with Traveling Wave Ion Mobility Spectrometry and Molecular Dynamics Simulations. *J. Am. Soc. Mass Spectrom.* **2015**, *26* (6), 1004–1013. https://doi.org/10.1007/s13361-015-1094-1.
(21) Simmonds, A. L.; Lopez-Clavijo, A. F.; Winn, P. J.; Russell, D. H.; Styles, I. B.; Cooper, H. J. Structural Analysis of 14-3-3-ζ-Derived Phosphopeptides Using Electron Capture Dissociation Mass Spectrometry, Traveling Wave Ion Mobility Spectrometry, and Molecular Modeling. *J. Phys. Chem. B* **2020**, *124* (3), 461–469. https://doi.org/10.1021/acs.jpcb.9b08506.
(22) Declercq, A.; Bouwmeester, R.; Hirschler, A.; Carapito, C.; Degroeve, S.; Martens, L.; Gabriels, R. MS2Rescore: Data-Driven Rescoring Dramatically Boosts Immunopeptide Identification Rates. *Mol. Cell. Proteomics* **2022**, *21* (8), 100266. https://doi.org/10.1016/j.mcpro.2022.100266.
(23) Degroeve, S.; Gabriels, R.; Velghe, K.; Bouwmeester, R.; Tichshenko, N.; Martens, L. *Ionbot: A Novel, Innovative and Sensitive Machine Learning Approach to LC-MS/MS Peptide Identification*; preprint; Bioinformatics, 2021. https://doi.org/10.1101/2021.07.02.450686.
(24) Wilhelm, M.; Zolg, D. P.; Graber, M.; Gessulat, S.; Schmidt, T.; Schnatbaum, K.; Schwencke-Westphal, C.; Seifert, P.; De Andrade Krätzig, N.; Zerweck, J.; Knaute, T.; Bräunlein, E.; Samaras, P.; Lautenbacher, L.; Klaeger, S.; Wenschuh, H.; Rad, R.; Delanghe, B.; Huhmer, A.; Carr, S. A.; Clauser, K. R.; Krackhardt, A. M.; Reimer, U.; Kuster, B. Deep Learning Boosts Sensitivity of Mass Spectrometry-Based Immunopeptidomics. *Nat. Commun.* **2021**, *12* (1), 3346. https://doi.org/10.1038/s41467-021-23713-9.

(25) Wilm, M.; Shevchenko, A.; Houthaeve, T.; Breit, S.; Schweigerer, L.; Fotsis, T.; Mann, M. Femtomole Sequencing of Proteins from Polyacrylamide Gels by Nano-Electrospray Mass Spectrometry. *Nature* **1996**, *379* (6564), 466–469. https://doi.org/10.1038/379466a0.
(26) Meyer, J. G.; Niemi, N. M.; Pagliarini, D. J.; Coon, J. J. Quantitative Shotgun Proteome Analysis by Direct Infusion. *Nat. Methods* **2020**, *17* (12), 1222–1228. https://doi.org/10.1038/s41592-020-00999-z.
(27) Jiang, Y.; Hutton, A.; Cranney, C. W.; Meyer, J. G. Label-Free Quantification from Direct Infusion Shotgun Proteome Analysis (DISPA-LFQ) with CsoDIAq Software. *Anal. Chem.* **2022**, acs.analchem.2c02249. https://doi.org/10.1021/acs.analchem.2c02249.
(28) Ivanov, M. V.; Bubis, J. A.; Gorshkov, V.; Abdrakhimov, D. A.; Kjeldsen, F.; Gorshkov, M. V. Boosting MS1-Only Proteomics with Machine Learning Allows 2000 Protein Identifications in Single-Shot Human Proteome Analysis Using 5 Min HPLC Gradient. *J. Proteome Res.* **2021**, *20* (4), 1864–1873. https://doi.org/10.1021/acs.jproteome.0c00863.
(29) Harris, C. R.; Millman, K. J.; Van Der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; Van Kerkwijk, M. H.; Brett, M.; Haldane, A.; Del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E. Array Programming with NumPy. *Nature* **2020**, *585* (7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2.
(30) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; Van Der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; Van Mulbregt, P.; SciPy 1.0 Contributors; Vijaykumar, A.; Bardelli, A. P.; Rothberg, A.; Hilboll, A.; Kloeckner, A.; Scopatz, A.; Lee, A.; Rokem, A.; Woods, C. N.; Fulton, C.; Masson, C.; Häggström, C.; Fitzgerald, C.; Nicholson, D. A.; Hagen, D. R.; Pasechnik, D. V.; Olivetti, E.; Martin, E.; Wieser, E.; Silva, F.; Lenders, F.; Wilhelm, F.; Young, G.; Price, G. A.; Ingold, G.-L.; Allen, G. E.; Lee, G. R.; Audren, H.; Probst, I.; Dietrich, J. P.; Silterra, J.; Webber, J. T.; Slavič, J.; Nothman, J.; Buchner, J.; Kulick, J.; Schönberger, J. L.; De Miranda Cardoso, J. V.; Reimer, J.; Harrington, J.; Rodríguez, J. L. C.; Nunez-Iglesias, J.; Kuczynski, J.; Tritz, K.; Thoma, M.; Newville, M.; Kümmerer, M.; Bolingbroke, M.; Tartre, M.; Pak, M.; Smith, N. J.; Nowaczyk, N.; Shebanov, N.; Pavlyk, O.; Brodtkorb, P. A.; Lee, P.; McGibbon, R. T.; Feldbauer, R.; Lewis, S.; Tygier, S.; Sievert, S.; Vigna, S.; Peterson, S.; More, S.; Pudlik, T.; Oshima, T.; Pingel, T. J.; Robitaille, T. P.; Spura, T.; Jones, T. R.; Cera, T.; Leslie, T.; Zito, T.; Krauss, T.; Upadhyay, U.; Halchenko, Y. O.; Vázquez-Baeza, Y. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17* (3), 261–272. https://doi.org/10.1038/s41592-019-0686-2.
(31) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9* (3), 90–95. https://doi.org/10.1109/MCSE.2007.55.
(32) The pandas development team. Pandas-Dev/Pandas: Pandas 1.4.0, 2022. https://doi.org/10.5281/ZENODO.5893288.
(33) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12* (85), 2825–2830.
(34) Waskom, M. Seaborn: Statistical Data Visualization. *J. Open Source Softw.* **2021**, *6* (60), 3021. https://doi.org/10.21105/joss.03021.
(35) Levitsky, L. I.; Klein, J. A.; Ivanov, M. V.; Gorshkov, M. V. Pyteomics 4.0: Five Years of Development of a Python Proteomics Framework. *J. Proteome Res.* **2019**, *18* (2), 709–714. https://doi.org/10.1021/acs.jproteome.8b00717.
(36) Chollet, F.; others. Keras, 2015. https://keras.io.
(37) Seabold, S.; Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. In *9th Python in Science Conference*; 2010.

(38) Servén, D.; Brummitt, C.; Abedi, H.; Hlink. Dswah/pyGAM: V0.8.0, 2018. https://doi.org/10.5281/ZENODO.1208723.
(39) Kim, S.; Pevzner, P. A. MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics. *Nat. Commun.* **2014**, *5* (1), 5277. https://doi.org/10.1038/ncomms6277.
(40) The, M.; MacCoss, M. J.; Noble, W. S.; Käll, L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J. Am. Soc. Mass Spectrom.* **2016**, *27* (11), 1719–1727. https://doi.org/10.1007/s13361-016-1460-7.
(41) Tyanova, S.; Temu, T.; Cox, J. The MaxQuant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics. *Nat. Protoc.* **2016**, *11* (12), 2301–2319. https://doi.org/10.1038/nprot.2016.136.
(42) Gabriels, R.; Declercq, A.; Bouwmeester, R.; Degroeve, S.; Martens, L. Psm_utils: A High-Level Python API for Parsing and Handling Peptide-Spectrum Matches and Proteomics Search Results. *J. Proteome Res.* **2023**, *22* (2), 557–560. https://doi.org/10.1021/acs.jproteome.2c00609.
(43) Hulstaert, N.; Shofstahl, J.; Sachsenberg, T.; Walzer, M.; Barsnes, H.; Martens, L.; Perez-Riverol, Y. ThermoRawFileParser: Modular, Scalable, and Cross-Platform RAW File Conversion. *J. Proteome Res.* **2020**, *19* (1), 537–542. https://doi.org/10.1021/acs.jproteome.9b00328.
(44) Abdrakhimov, D. A.; Bubis, J. A.; Gorshkov, V.; Kjeldsen, F.; Gorshkov, M. V.; Ivanov, M. V. Biosaur: An Open-source Python Software for LC/MS1 Peptide Feature Detection with Ion-mobility Support. *Rapid Commun. Mass Spectrom.* **2021**. https://doi.org/10.1002/rcm.9045.
(45) Lee, S.; Park, H.; Kim, H. Comparison of False-Discovery Rates of Various Decoy Databases. *Proteome Sci.* **2021**, *19* (1), 11. https://doi.org/10.1186/s12953-021-00179-7.