

The structural context of PTMs at a proteome wide scale

Isabell Bludau¹, Sander Willems¹, Wen-Feng Zeng¹, Maximilian T. Strauss², Fynn M. Hansen¹, Maria C. Tanzer¹, Ozge Karayel¹, Brenda A. Schulman³, Matthias Mann^{1,2}

¹ Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany

² Department of Clinical Proteomics, NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, 2200 Copenhagen, Denmark

³ Department of Molecular Machines and Signaling, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany

Abstract

The recent revolution in computational protein structure prediction provides folding models for entire proteomes, which can now be integrated with large-scale experimental data. Mass spectrometry (MS)-based proteomics has identified and quantified tens of thousands of post-translational modifications (PTMs), most of them of uncertain functional relevance. In this study, we determine the structural context of these PTMs and investigate how this information can be leveraged to pinpoint potential regulatory sites. Our analysis uncovers global patterns of PTM occurrence across folded and intrinsically disordered regions. We found that this information can help to distinguish regulatory PTMs from those marking improperly folded proteins. Interestingly, the human proteome contains thousands of proteins that have large folded domains linked by short, unstructured regions that are strongly enriched in regulatory phosphosites. These include well-known kinase activation loops that induce protein conformational changes upon phosphorylation. This regulatory mechanism appears to be widespread in kinases but also occurs in other protein families such as solute carriers. It is not limited to phosphorylation but includes ubiquitination and acetylation sites as well. Furthermore, we performed three-dimensional proximity analysis which revealed examples of spatial co-regulation of different PTM types and potential PTM crosstalk. To enable the community to build upon these first analyses, we provide tools for 3D visualization of proteomics data and PTMs as well as python libraries for data accession and processing.

Introduction

Post translational modifications (PTMs) are an important mechanism to regulate the activity and function of proteins. Mass spectrometry (MS)-based proteomics has become the method of choice to not only identify and quantify proteomes (Aebersold and Mann, 2016, 2003), but also to investigate PTMs on a proteome-wide scale (Bekker-Jensen et al., 2020; Hansen et al., 2021; Sharma et al., 2014). Despite impressive technological progress, a key challenge in the PTM and signaling fields remains to distinguish PTMs that are of direct functional relevance from the tens of thousands that are routinely measured. This is necessary to match the proteomics data to dedicated, low throughput biochemical follow-up studies that characterize the biological functions of candidate PTMs.

To assess functional relevance of PTMs on a more global scale, Beltrao and co-workers recently presented a machine learning model that uses information from different features indicative of proteomic, structural, regulatory or evolutionary relevance to derive a functional score for a large catalog of phosphosites (Ochoa et al., 2019). Another recent study directly evaluated the functional relevance of phosphorylations purely based on available structural information (Kamacioglu et al., 2021). Based on these and many previous studies, we know that phosphorylations are predominantly observed on spatially accessible amino acids and particularly in intrinsically disordered regions (IDRs) (Iakoucheva et al., 2004; Kamacioglu et al., 2021; Tyanova et al., 2013). Furthermore, it stands to reason that phosphorylations in flexible regions within folded domains and on binding interfaces are more likely to be functional compared to those that are buried or less accessible in rigidly folded regions (Kamacioglu et al., 2021). Although these studies impressively highlight the value of integrating structural information into the analysis of PTMs, they have been limited to phosphorylation and to the set of available experimentally derived structures deposited in PDB, which furthermore inherently favor stable regions of proteins (Burley et al., 2021).

Recently, there has been a key breakthrough in computational protein structure prediction from just the amino acid sequence of a protein. The novel deep learning models in AlphaFold2 (henceforth referred to as AlphaFold) (Jumper et al., 2021), rapidly followed by RoseTTAFold (Baek et al., 2021), were shown to regularly achieve high accuracy in predicting protein structures that are largely comparable to those determined by experimental methods. By providing structural information for almost the complete human proteome as well as the proteomes of over 20 model organisms, the AlphaFold protein structure database (AlphaFold DB, <https://alphafold.ebi.ac.uk>) now enables structural investigations on a proteome-wide scale, thus promising to accelerate our understanding

of the structure-to-function relationship of proteins (Tunyasuvunakool et al., 2021; Varadi et al., 2021).

Here we set out to combine this wealth of structural information with proteomics data, especially large-scale PTM information, with the goal of shedding new light on the long-standing question of functional relevance of PTMs. We present a first systematic assessment of how PTM data can be integrated with deep-learning predicted structures on a proteome-wide scale. We then explore key features in the structure function domain by combining predictions of functional relevance with domain features and discover a multitude of sites with potential regulatory roles. To enable the community to further explore the numerous related biological questions, we provide a Python package called StructureMap, that allows to easily and quickly access and integrate structural data from AlphaFold DB with proteomics data and information on PTMs. Finally, we provide an extended version of our previously published AlphaMap tool for sequence visualization (Voytik et al., 2021), which now enables the mapping of peptides and PTMs to three-dimensional protein structures.

Results

Estimation of side chain exposure and intrinsically disordered regions from predicted protein structures

To make the information provided by predicted structures accessible for systematic analyses, we first wanted to extract it into tractable metrics such as amino acid side chain exposure or the categorization of amino acids into structured and intrinsically disordered regions (IDRs). At this point, it is important to make a clear distinction between predicted and experimentally derived structures. Experimental structures are often incomplete and may only cover a specific sequence region. In contrast, the predicted structures in AlphaFold DB in principle cover the entire protein sequence from N- to C-terminus. Importantly, each amino acid in the predicted structure is associated with a specific prediction confidence (pLDDT) derived from the deep learning models (Jumper et al., 2021). Additionally, the relative position of amino acids to each other is annotated with an expected distance error in Ångströms (predicted aligned error, PAE). Although experimental structures are available for a large set of proteins, we decided to base all our analyses on predicted structures only. This prevents ambiguities in the integration of multiple conflicting structures and allows us to leverage the complete sequence information, confidence metrics and PAE estimates.

We found that it is crucial to take AlphaFold's confidence and PAE metrics into account for best accuracy, in accordance with others (Akdel et al., 2021). Therefore, we developed a prediction-aware metric. To evaluate amino acid side chain exposure, we built upon the previously introduced half-sphere exposure (HSE) (Hamelryck, 2005; Heffernan et al., 2016). This method essentially calculates the half-sphere of a given amino acid in the direction of its side chain at a defined radius, and counts the number of alpha carbon atoms from other residues within it, with a larger number reflecting less exposure and vice-versa (Supplementary Figure 1A). We adjusted the HSE to take prediction uncertainties into account, meaning that an alpha carbon atom is only considered as a neighbor if it still lies within the defined radius after addition of its PAE for this alpha carbon pair. We further introduce an angle parameter that determines whether to consider the full-sphere, half-sphere, or any other angle in direction of the amino acid side chain. Accordingly, we termed our metric prediction-aware part-sphere exposure (pPSE). To illustrate, Figure 1A shows the AlphaFold predicted structure of Mitogen-activated protein kinase 3 (MAPK3) colored by pLDDT and by the pPSE using a radius of 12 Å and an angle of 70°.

The higher the pPSE, the more other amino acids are in close proximity to the amino acid being evaluated, and hence the more structured its environment. In this respect pPSE offers a similar metric as the commonly used solvent accessible surface area (SASA), or relative SASA (RSA). However, the pPSE directly considers side chain orientation and, more importantly, it takes the prediction error of AlphaFold into account. Estimating the pPSEs for all amino acids in the 20,053 predicted human protein structures on AlphaFold only takes minutes on a laptop computer with our implementation. This makes the tool especially useful for system-wide studies where tens of thousands of proteins are evaluated for particular properties. In a community effort, it was recently shown that a smoothed AlphaFold confidence metric (pLDDT) or RSA metric based on the predicted structures can confidently determine IDRs (Akdel et al., 2021), improving on IUPred2, a state-of-the-art tool for IDR prediction (Mészáros et al., 2018). We found that our smoothed pPSE metric even obtains slightly better results when using a radius of 24 Å and a full sphere (the true positive rate improves from 83% for RSA to 86% for pPSE, see Figure 1A, 1B, 1C and Supplementary Figure 1C and 1D). Importantly, considering the positional uncertainty during pPSE estimation considerably improves IDR prediction compared to neglecting it (TPR increase from 79% to 86%).

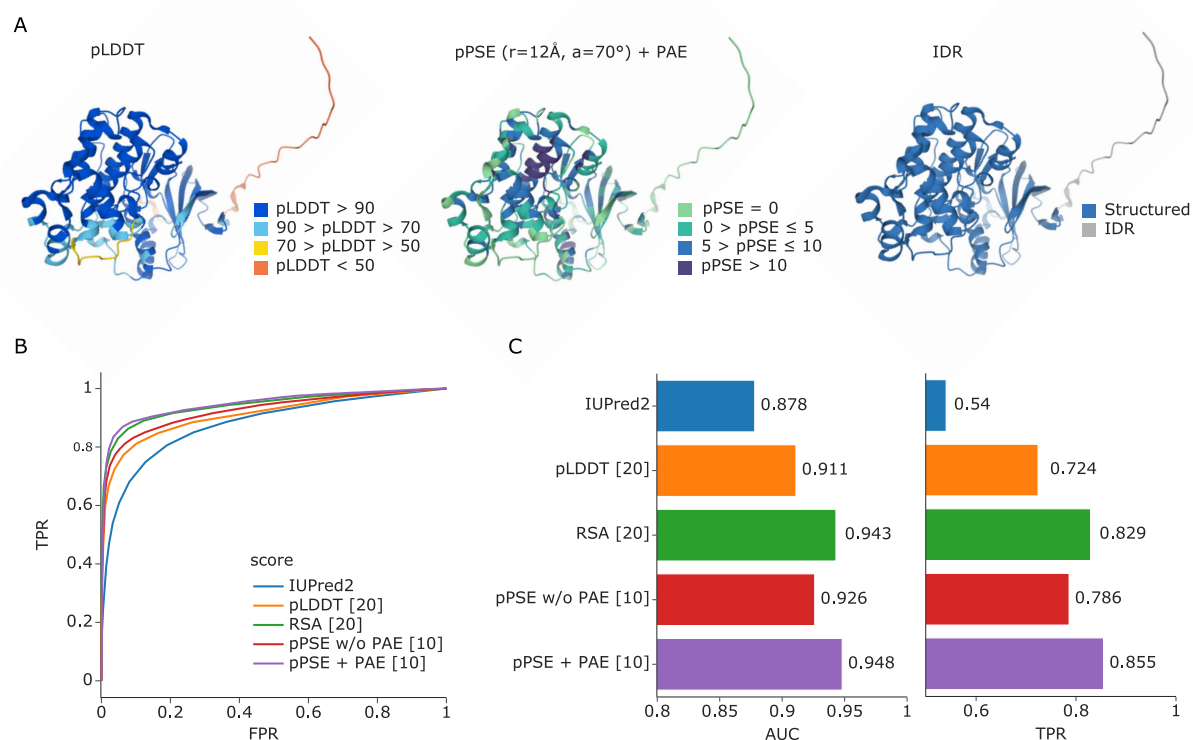


Figure 1. Estimation of amino acid side chain exposure and intrinsically disordered regions (IDRs). (A) AlphaFold predicted structure of Mitogen-activated protein kinase 3 (MAPK3) colored by prediction confidence (pLDDT, left), colored by our prediction-aware part-sphere exposure (pPSE) metric using a radius of 12 Å and an angle of 70° (center), and colored by our prediction of structured regions and IDRs (right). (B) Receiver operating characteristic (ROC) curve for predicting IDRs based on IUPred2 in comparison to the smoothed pLDDT confidence scores from AlphaFold, the smoothed relative solvent accessible surface areas (RSA), and the pPSE with (+) and without (w/o) considering the predicted aligned error (PAE) (radius = 24 Å, angle = 70°). (C) Corresponding area under the curve (AUC) values and the true positive rates (TPRs) at a 5% false positive rate (FPR). The numbers in square brackets behind each metric indicate the smoothing windows that were used, see Supplementary Figure 1 for a comprehensive parameter screen.

Most PTMs are enriched in intrinsically disordered regions, whereas ubiquitinations accumulate in structured domains

Having the proteome-wide information on IDRs at hand, we next performed an enrichment analysis of different PTMs located within those regions across the entire human proteome. We first used all PTMs annotated in the PhosphoSitePlus database that overlapped with structural data, comprising a total of 334,529 sites, including phosphorylations (p), ubiquitinations (ub), sumoylations

(sm), acetylations (ac), methylations (m) and the glycosylations O-GalNAc (gl) and O-GlcNAc (ga) (Hornbeck et al., 2015). In agreement with previous observations of phosphorylations (Iakoucheva et al., 2004; Tyanova et al., 2013), most PTMs were indeed significantly enriched in IDRs (Figure 2A). In contrast, our analysis revealed that ubiquitinations and, to a lesser extent, acetylations were significantly underrepresented in IDRs (ubiquitination: odds-ratio = 0.6, adj. p-value ≈ 0 , number of sites = 91,388; acetylation: odds-ratio 0.9, adj. p-value = $6e-08$, number of sites = 21,202). However, if only PTM sites with a known regulatory function are considered, this effect disappeared for ubiquitination and was even reversed for acetylation (ubiquitination: odds-ratio = 1.0, adj. p-value = 0.7, number of sites = 451; acetylation: odds-ratio 2.0, adj. p-value = $1e-18$, number of sites = 631).

A possible explanation for non-regulatory ubiquitination sites in structured regions is the tagging of misfolded proteins for degradation by the proteasome. Importantly, most datasets that contribute ubiquitination sites to PhosphoSitePlus are from samples treated with proteasome inhibitors (Figure 2B). This leads to the accumulation of misfolded proteins that presumably expose normally inaccessible lysine residues. Furthermore, ubiquitin might also be specifically attached to structured regions to destabilize the protein fold, creating new short IDRs that are required for proteasome binding and subsequent degradation (Carroll et al., 2020).

To directly test our hypothesis, we contrasted ubiquitination sites from proteasome inhibitor-treated and untreated samples in the same experiment (Hansen et al., 2021). Interestingly, ubiquitination sites unique to the proteasome inhibition condition confirm the overrepresentation of ubiquitination in structured regions (Figure 2A right, odds-ratio = 0.4, adj. p-value = 0, number of sites = 19,517). The same effect can still be observed for the sites shared between both datasets (odds-ratio = 0.6, adj. p-value = $1e-193$, number of sites = 11,741), whereas the ubiquitination sites unique to the untreated condition are enriched in IDRs (odds-ratio = 1.4, adj. p-value = $1e-32$, number of sites = 6321), similar to most other PTMs. Overall, 78 % of ubiquitinated lysines unique to the inhibitor treatment condition and 71 % of shared ubiquitin sites were in structured regions. In notable contrast, in the uninhibited condition, only 50 % of observed ubiquitinations are in structured regions.

To further pursue these findings, we disregarded all amino acids in predicted IDRs and asked if PTMs are enriched in amino acids with side chains of high versus low exposure within structured regions. To this end, we calculated the pPSE for each amino acid at a radius of 12 Å and an angle of 70°. We considered amino acids with a pPSE ≤ 5 to have a high exposure and those with a pPSE > 5 as low exposure (see Supplementary Figure 1B and the methods section for details on the cutoff selection). Due to much lower numbers of annotated PTMs in structured regions, statistical

significance decreases, but phosphorylations were still enriched in amino acids with high side chain exposure, whereas ubiquitinations were enriched in those with low side chain exposure (Supplementary Figure 2). This indicates that ubiquitinations are located on lysines that are buried within the structure of a properly folded protein rather than on outwards facing amino acids of a helix or beta-sheet at the protein's surface.

In addition to these global analyses, we further explored modified proteins individually to test if PTMs were specifically enriched in certain structural elements. For ubiquitination, this revealed 71 proteins with a significant enrichment in structured regions (odds-ratio < 1, adj. p-value ≤ 0.05). Interestingly 80% of them were DNA or RNA binding proteins. A striking example is Ras GTPase-activating protein-binding protein 2 (G3BP2), an RNA-binding protein that plays an essential role in cytoplasmic stress granule formation (Matsuki et al., 2013). Here all 9 ubiquitination sites are in structured regions (Figure 2C).

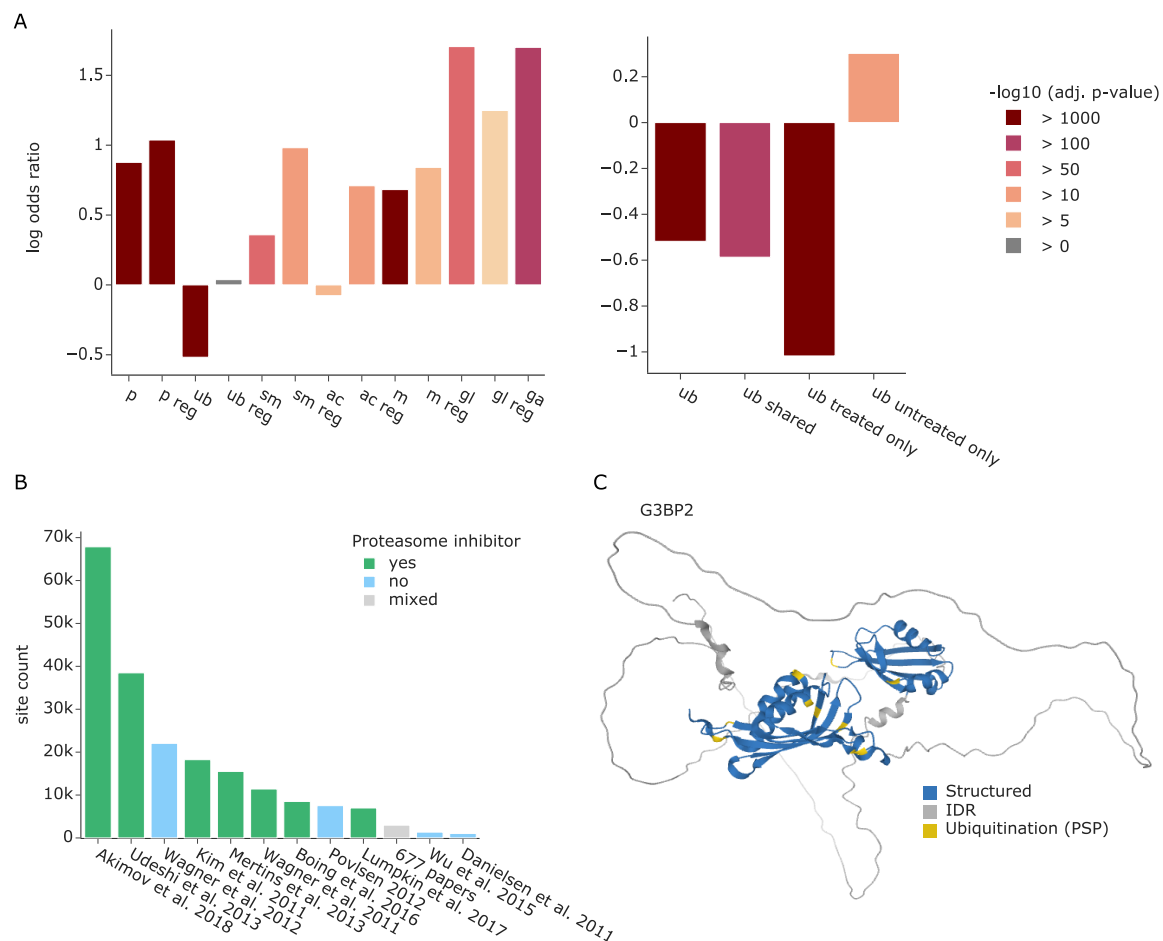


Figure 2. Enrichment analysis of PTMs in IDRs. (A) Enrichment of different PTMs annotated in the PhosphoSitePlus database in IDRs (left). Enrichment of ubiquitinated lysines annotated in PhosphoSitePlus versus ubiquitinations detected in a dataset treated with proteasome inhibitor or untreated (right). PTMs are abbreviated as follows: phosphorylations (p), ubiquitinations (ub), sumoylations (sm), acetylations (ac), methylations (m) and the glycosylations O-GalNAc (gl) and O-GlcNAc (ga). (B) Overview of datasets that contribute ubiquitination sites to PhosphoSitePlus and their use of proteasome inhibitors. (C) AlphaFold predicted structure of Ras GTPase-activating protein-binding protein 2 (G3BP2) colored by structured regions (blue) and predicted IDRs (grey) as well as ubiquitination sites annotated in PhosphoSitePlus (yellow).

Improving sequence motif analysis through structural context

Post-translational modifications are commonly introduced by dedicated enzymes such as kinases for phosphorylation, E3-ligases for ubiquitination and proteases for proteolytic cleavage, which generally recognize specific sequence motifs. Given that most PTMs have a preference for exposed amino acids, we reasoned that sequence motifs in accessible protein regions should be preferred compared to inaccessible ones, adding another layer of selectivity. To explore this hypothesis, we first selected a curated list of kinase phosphorylation motifs available in Perseus (Tyanova et al., 2016). Based on phosphosites in both PhosphoSitePlus and a recent in-depth, COVID-related phospho study (Stukalov et al., 2021), we first confirmed that phosphorylations are generally enriched in kinase phosphorylation motifs compared to all possible STY sites. This effect is even more pronounced for regulatory sites and sites from the Stukalov study (Figure 3A). Confirming our hypothesis, motifs in IDRs and motifs harboring exposed STY sites were indeed preferentially modified (Figure 3B and Supplementary Figure 3A).

These results highlight that proteome-wide structural information can provide valuable insights for motif analysis and help interpret experiments determining enzyme-substrate relationships. We illustrate this with a large in vitro kinase substrate screen by Ishihama and colleagues (Sugiyama et al., 2019). These authors dephosphorylated HeLa cell lysates with phosphatases, which they then deactivated by heat. The resulting – partially denatured – sample then served as a substrate pool to which 385 different recombinant human protein kinases were individually added to investigate which kinases phosphorylate which specific amino acid sites. Their study resulted in an unprecedented set of 20,669 phosphosites, 175,574 proposed kinase-substrate relationships and 1,427 kinase phosphorylation motifs.

Based on the denaturation of the proteome, and as already indicated by the authors, we speculated that many of the previously inaccessible sites would now be amenable to phosphorylation,

providing an ideal test case for structure-based interpretation. Indeed, our analysis revealed that the identified phosphosites were not enriched in IDRs, in contrast to the above examples. However, the sites that overlapped with other studies did show the expected enrichment in IDRs (Figure 3C). These results suggest that structural information can be used to refine the list of reported phosphosites to a set that better represents the sites expected to occur on endogenously folded proteins.

To further test the effect of 3D exposure filtering on sequence motif analysis, we selected kinases and performed a motif analysis separately for sites of high surface exposure ($pPSE \leq 5$) and sites of low exposure ($pPSE > 5$) using the PSSMSearch tool (Krystkowiak et al., 2018; O'Shea et al., 2013). As can clearly be seen in Figure 3D, kinases showed striking differences in sequence motifs between sites of high and low exposure. While motifs for sites of high exposure are mostly in agreement with the reference set provided by Perseus (Tyanova et al., 2016), this was not the case for the sites with low exposure. To account for the fact that there are fewer sites in structured regions, we also selected random, equally sized subsets of the high-exposure sites and repeated the motif analysis. This resulted in similar patterns as for the full set of sites, but with lower enrichment scores (Supplementary Figure 3B).

In the case of the RAC-alpha serine/threonine-protein kinase (ACT1), the phosphosites of high exposure clearly display the R-x-R-x-x-pS-F motif (Figure 3D, top left panel). In contrast, phosphosites of low exposure only provide a noisy motif (Figure 3D, top right panel). For other kinases, such as Stress-activated protein kinase JNK1, a serine/threonine-specific protein kinase, the phosphosites of low exposure even have an unexpected enrichment for a phosphorylated central tyrosine residue and the proline at the +1 position is hardly enriched (Figure 3D, right panel).

Together, these results establish that the structural information from AlphaFold and the tools presented herein can guide determining potential regulatory PTM sites found by in vitro screens, increasing the confidence of measured kinase-substrate pairs by filtering out a subset that are less likely to be true in-vivo substrates. As we have shown, this can improve kinase phosphorylation motif predictions and help to interpret individual sites of interest. Here we focused on phosphorylations but we expect similar benefits for the analysis of any other types of motifs, including enzyme recognition or general protein binding.

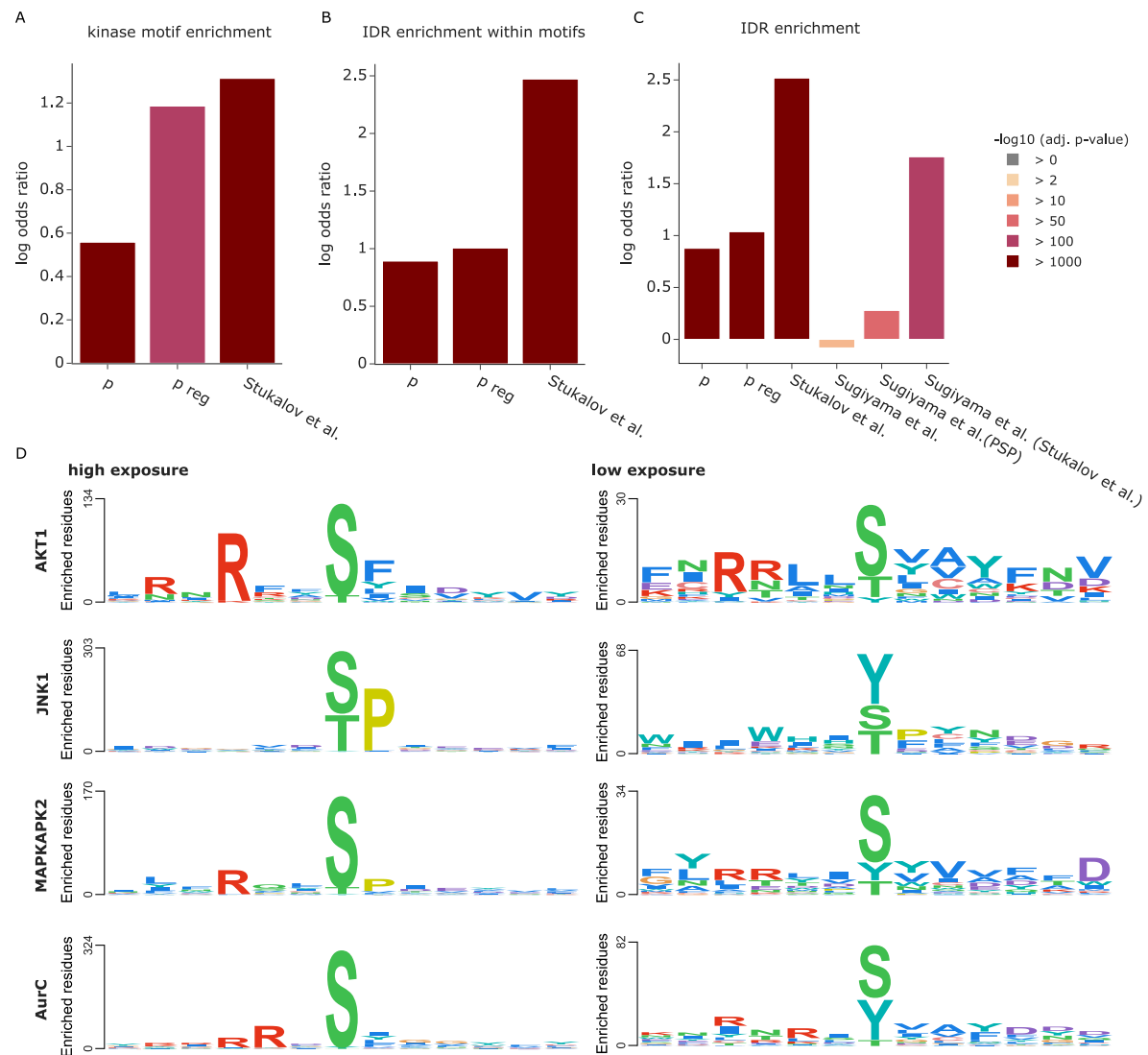


Figure 3. Exploiting the 3D context of kinase phosphorylation motifs. (A) Enrichment of phosphorylation events in kinase motifs compared to all possible STY sites. **(B)** Enrichment of phosphorylations in kinase motifs within IDRs compared to all possible kinase motif occurrences. **(C)** Enrichment of phosphorylations in IDRs compared to all possible STY sites. The phosphosites reported by Sugiyama et al. (Sugiyama et al., 2019) were filtered for sites also reported in PhosphoSitePlus and by Stukalov et al. (Stukalov et al., 2021). **(D)** Sequence logos for different kinases. The PSSMSearch tool (Krystkowiak et al., 2018) was used with a log odds scoring method (O'Shea et al., 2013).

Functionally relevant PTMs are enriched in short IDRs within large structured domains exemplified by kinase activation loops

Having established the enrichment of most PTMs in unstructured regions in the dataset, we further explored the structural context of this effect. We found that these PTMs are often located in short IDRs that are embedded within larger structured domains. To investigate if this was a random effect or whether these short IDRs could be of biological relevance, we extracted all proteins with short IDRs of maximally 20 amino acids length between two flanking structured regions of at least 80 amino acids. Among the 20,053 human proteins in AlphaFold DB, 2,454 have such a pattern. Notably, enrichment analysis of these proteins revealed a significant overrepresentation of GO molecular functions related to ATP binding, protein kinase activity, ATPase activity, transmembrane transporter activity and motor activity (Figure 5A).

To further evaluate the relevance of the short IDRs, we analyzed their occupancy with functional phosphosites. To this end, we first considered regulatory phosphosites in PhosphoSitePlus (Hornbeck et al., 2015). Our analysis revealed that these regulatory sites, as compared to phosphosites in general, are significantly enriched in short IDRs versus all IDRs (odds-ratio = 1.57, adj. p-value = 0.001, Figure 2B). We also extracted a second set of phosphosites from the above-mentioned study of Beltrao and co-workers, where sites were given a functional score between zero and one, with scores equal to or above 0.5 considered functional (Ochoa et al., 2019). We indeed observed that the higher the score, the stronger the enrichment of functional phosphosites in short IDRs, ranging from an odds-ratio of 1.43 at a cut-off of 0.5 to an odds-ratio of 7.7 at 0.9 (Figure 4B). This raises the exciting possibility that at least a subset of these phosphorylation sites may play important roles in structural and functional rearrangements of their neighboring domains.

Due to the strong enrichment of kinases among proteins with short IDRs, we next investigated whether they overlap with any known kinase substructures annotated in KinaseMD (Hu et al., 2021). That database contains substructure annotations for 388 kinases, 365 of which also have predicted structures in AlphaFold DB. We found a large overlap of 72 short IDRs with the 309 annotated activation loops, but none with the 171 G-loops or 230 C α -helix positions in these kinases. If the five amino acids flanking a short IDR are also considered, this number increases to 79 (also see methods section). These results are particularly interesting, because the activation loops of many kinases undergo structural rearrangements upon phosphorylation (Nolen et al., 2004). Strikingly, 55 of the 79 kinases (70 %) with an overlap of the extended short IDR and the annotated activation loop have an annotated regulatory phosphosite or a functional score higher than 0.5, for a total of 99 different

phosphosites. To illustrate, receptor-interacting serine/threonine-protein kinase 2 (RIPK2) and Mitogen-activated protein kinase kinase kinase kinase 1 (MAP4K1) both show an overlap between our predicted short IDR and the annotated activation loop with known regulatory phosphosites (Figure 4C, D and E).

Next, we evaluated short IDRs outside of annotated kinase activation loops. An interesting example of these is the serine/threonine-protein kinase CHK2 (CHEK2). Although the annotated activation loop (amino acids 367 to 389) was not detected as a short IDR, our data contained an alternative short IDR region (amino acids 262 and 263). Directly flanking this IDR is a phosphorylation site with a functional score of 0.44 (S260). Notably, CHEK2 was reported to be autophosphorylated at residue S260, which is important for triggering a conformational change in CHEK2 that favors a dissociation of dimers into fully active monomers (Zannini et al., 2014).

We also observed short IDRs in many proteins apart from kinases. One example is Band 3 anion transport protein (SLC4A1) where we found that the short IDR from residue 354 to 369 contains a known regulatory phosphosite on Y359 (Flatt et al., 2020). In addition to phosphorylations, other PTMs might also be biologically relevant in the short IDRs. Indeed, one of three regulatory ubiquitination sites in another solute carrier protein - SLC22A6 – is located directly in a short IDR, whereas the other two are in close proximity (Figure 4F). These three ubiquitination sites have previously been shown to play an important role for the internalization of this protein (Li et al., 2013).

Compared to phosphorylation, the percentage of other PTM sites with known functions is even smaller. Our findings suggest that selecting candidates from PTM sites within or in close proximity to short IDRs is a promising strategy to discover functional relevance. We found that our predicted short IDRs extended by five amino acids contain a wealth of PTM sites that are not yet annotated as regulatory in PhosphoSitePlus (1,437 phosphosites, 898 ubiquitination sites, 118 acetylations, 43 sumoylations, 53 methylations, 33 GalNAc and 1 GlcNAc) (Figure 4G, Supplementary dataset 1). We further provide a list of all human short IDRs for researchers to explore their favorite proteins, enabling the integration of own experimental data from PTM studies or other types of studies, such as mutational screens (Supplementary dataset 2).



Figure 4. Regulatory PTMs accumulate in short IDRs. (A) Enrichment analysis of proteins with short IDRs. (B) Enrichment of regulatory phosphosites from PhosphoSitePlus in short IDRs compared to all other IDRs. (C) Sequence plot showing the N- to C-terminus of different proteins colored by whether the amino acid is part of a structured region (blue) or an IDR (grey). All phosphosites annotated in PhosphoSitePlus are indicated by circles. Regulatory sites are colored in dark red and stand out higher than non-regulatory sites (salmon). Regions of short IDRs including a five amino acid extension are indicated in light green below the sequence. Annotated kinase activation loops (A-loops) from KinaseMD are indicated in dark green above the sequence. The predicted structures of RIPK2 (D) and MAP4K1 (E) are colored by structured regions (blue) and predicted IDRs (grey) as well as known regulatory phosphosites annotated in PhosphoSitePlus (dark red). Specific regions of interest are highlighted by an orange circle. (F) The predicted structure of SLC22A6 is colored by structured regions (blue) and predicted IDRs (grey) as well as known regulatory ubiquitination sites annotated in PhosphoSitePlus

(yellow). **(G)** Overview of phosphorylations (left) and other PTMs (right) that lie within short unstructured regions or their flanking five amino acids (AA).

PTMs on proteins preferentially occur in three-dimensional clusters

It is well known that many proteins have hotspots of modifications. For example, multisite phosphorylation in specific sequence regions is critical in regulating the activity of many enzymes and their binding properties (Cohen, 2000). Furthermore, in the context of circadian biology, we recently showed that regulated ubiquitinations frequently occur in sequence proximity (Hansen et al., 2021). Beyond co-localization of the same PTM types, phospho-acceptors near PTM-modified lysines were shown to be preferentially phosphorylated in comparison to more distant residues (Beltrao et al., 2012). Those prior findings were obtained on the basis of linear sequence analysis. Now, with the spatial coordinates of each PTM acceptor residue provided by AlphaFold, we set out to evaluate PTM proximity in three-dimensional space.

First, we investigated whether PTM acceptors near a modified amino acid residue are more frequently observed to also be modified compared to more distant residues or to random expectation. For this, we extended the strategy of Krogan and co-workers (Beltrao et al., 2012) to evaluate distance in 3D space and to assess both individual PTM types and PTM co-occurrence (see methods section). Importantly, our metric considers the predicted positional uncertainty between any two PTM sites as a factor in the analysis. Furthermore, we only take structured regions and short IDRs into account for the proximity analysis (that is, we removed all IDRs of more than 20 amino acids). This ensures that proximity results are not influenced by regions of high structural uncertainty, as is the case for IDRs. It further avoids any biases that arise from the fact that many PTMs are enriched in IDRs and tend to cluster there in linear sequence space.

Our analysis revealed that the observed sites of PTM types annotated in PhosphoSitePlus indeed form 3D modification hotspots. Phosphorylations, ubiquitinations, sumoylations, acetylations and methylations each form tight clusters in 3D space, where proximal amino acids are preferentially modified compared to more distant residues or an equivalent random selection (Figure 5A). Due to an overall lower number of O-GlcNAc and O-GalNAc modified sites, the results for these modifications are less conclusive but also show a similar trend (Supplementary Figure 4). In addition to evaluating PTM types by themselves, we further investigated co-localization of different PTM types. This confirmed that phospho-acceptors near modified lysines (including ubiquitination, sumoylation and acetylation) are more frequently phosphorylated compared to random expectation (Beltrao et al.,

2012) (Figure 5B). This is also true for methylated lysines and arginines. Conversely, investigating ubiquitination sites near other PTMs revealed that they also preferentially occur close to phosphosites (Figure 5C). Other lysine modifications, however, often compete for the same or directly neighboring residues, but they do not generally favor proximity. Overall, our analysis reveals that many proteins have specific 3D regions and folds that are particularly prone to being modified by the same or different PTMs. This structurally supports the notion of PTM cross-talk.

Following these global analyses, we next explored 3D PTM clusters of all individual proteins. For this, we calculated all pairwise distances between modified amino acids and compared their average against a distribution of random PTM sites (see methods section). We again only considered structured protein regions and included the positional uncertainty between any two PTM sites in the distance calculation. Clustering analysis of phosphorylation and ubiquitination sites in PhosphoSitePlus revealed many proteins with significant PTM clusters, showing a strong enrichment for transmembrane proteins (Figure 5D). On those proteins, we detected these PTMs on the cytosolic domains and in 3D proximity, nicely confirming that the proximity analysis worked as intended.

To enable a more fine-grained inspection of PTM clusters, we explored an in-house phospho-dataset (Stukalov et al., 2021). Of 47 phosphoproteins with three or more sites in structured regions, 3D proximity analysis yielded six significant ones (adj. p-value ≤ 0.05 and ≥ 3 phosphosites, Supplementary Dataset 3). As an example, the mitochondrial Pyruvate dehydrogenase E1 component subunit alpha (PDHA1) had an adjusted proximity p-value of 0.005 and all six detected phosphosites are located within one protein pocket (Figure 5E). Three of these (S232, S293 and S300) have previously been reported to be substrates of PDK family kinases. Any single one of these is sufficient to inactivate PDHA1 and dephosphorylation of all sites is required for reactivation (Kato et al., 2008). Our 3D analysis suggests that the other three phosphorylation sites would have the same effect. Another interesting example is Aldo-keto reductase family 1 member B1 (AKR1B1) (adj. p-value = 0.03). In linear sequence space, S211 and S215 are close together, whereas S23 and Y49 are far apart. However, in the folded protein structure, all four are in the same pocket (Figure 5F). Interestingly, it also contains the annotated NADP binding site of AKR1B1, which consists of two distinct sequence stretches from amino acids 10 to 19 and 211 to 273 (based on UniProt annotation).

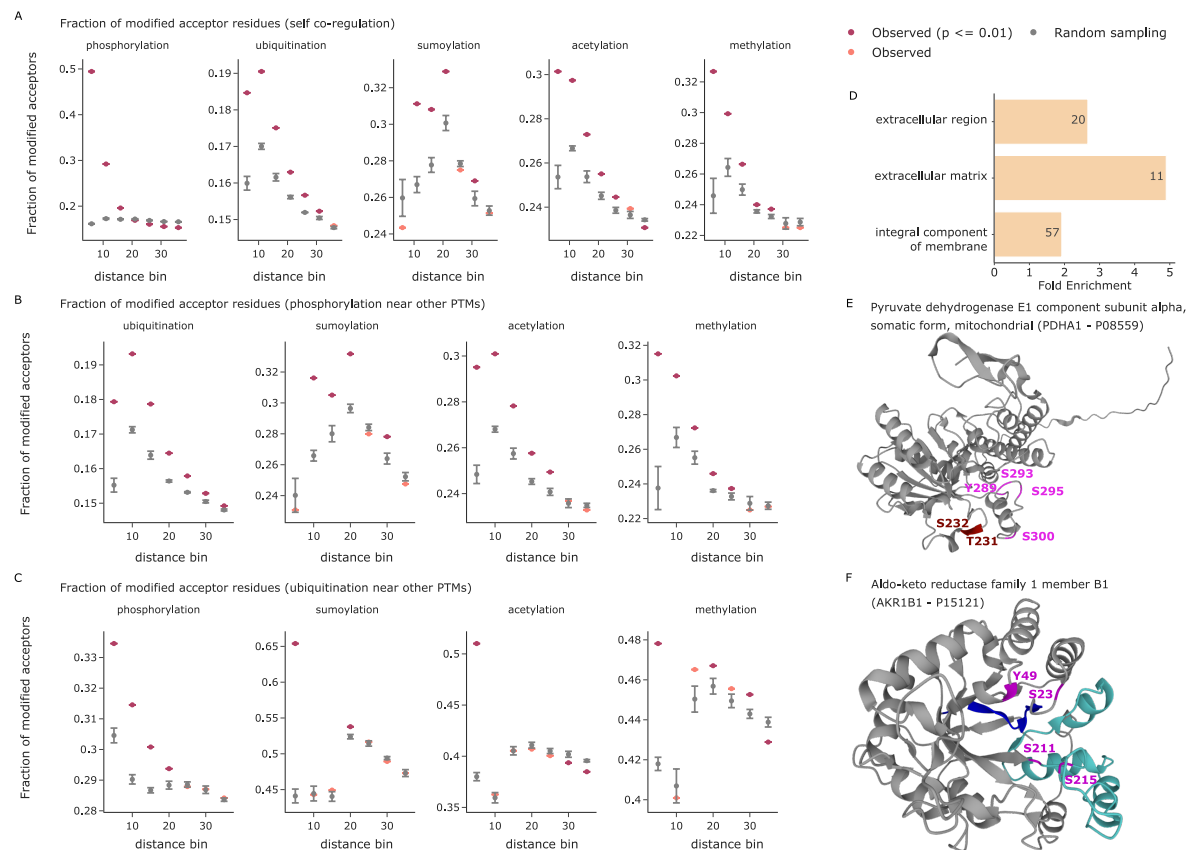


Figure 5. PTM proximity analysis in 3D. (A) The fraction of modified PTM acceptor residues is shown as a function of the 3D distance to a given modified amino acid in Å. Observed values (indicated in red when statistically significant and colored in salmon otherwise) are compared to the mean of five random samples including the same number of modified PTM sites (grey). Error bars indicate one standard deviation. The x-axes are divided in distance bins ranging from each previous bin to the indicated cutoff in Å. (B) The fraction of modified phospho-acceptor residues is shown as function of the 3D distance to a given modified amino acid in Å. (C) The fraction of ubiquitinated lysines is shown as function of the 3D distance to a given modified amino acid in Å. The smallest bin shows competition for the same central lysine residue. (D) Enrichment analysis of proteins with 3D phospho- and/or ubiquitination clusters ($FDR \leq 0.01$). (E) The predicted structure of Pyruvate dehydrogenase E1 component subunit alpha (PDHA1). Phosphorylations on the phospho-loop A are indicated in dark red (T231 and S232). The phosphorylations on phospho-loop B are indicated in magenta (Y289, S293, S295 and S300) (Kato et al., 2008). (F) The predicted structure of Aldo-keto reductase family 1 member B1 (AKR1B1). Residues annotated as NADP binding sites are highlighted in blue (amino acids 10 to 19) and turquoise (amino acids 211 to 273). Phosphorylations are indicated in magenta.

Community resources for enabling the systematic integration of PTM data with structure predictions from AlphaFold

This study only scratches the surface of biological insights that can be gained from combining PTM data with structural information. To enable the community to further investigate various research questions of interest to them in their own data or in public repositories, we created a toolset that facilitates systematic data exploration and integration. First, we provide StructureMap, an open-source Python package for processing predicted structures from AlphaFold DB and for integrating the data with PTM information. Its functionalities include (1) accession of predicted structures from AlphaFold DB and extraction of essential information into an internal data format, (2) calculation of the pPSEs of individual amino acids as well as estimation of IDRs, (3) extraction of short IDRs for PTM site prioritization, (4) import and formatting of PTM datasets, (5) enrichment analyses of PTMs in different structural regions and IDRs, (6) motif analysis in 3D context and filtering based on side chain exposure, and (7) multi-dimensional proximity analysis of PTMs. To further enable easy visualization of PTMs on the three-dimensional structure of proteins, we also extended our previously published AlphaMap tool, which is available as Python library as well as a stand-alone application with graphical user interface (Voytik et al., 2021). The source code of both tools is openly available with an Apache license on the MannLabs GitHub page and includes extensive documentation to readily enable researchers to understand and further adopt code to any specific needs (<https://github.com/MannLabs/structuremap> and <https://github.com/MannLabs/alphamap>).

Discussion and Outlook

PTMs provide essential mechanisms to regulate the activity and function of proteins. Although mass-spectrometry based proteomics routinely enables the identification and quantification of thousands of PTMs, systematic assessment of their functional relevance remains a persisting challenge. While previous work already demonstrated the merits of structural information for PTM analyses (Beltrao et al., 2012; Iakoucheva et al., 2004; Kamacioglu et al., 2021; Ochoa et al., 2019; Tyanova et al., 2013), the recent revolution in computational protein structure prediction (Baek et al., 2021; Jumper et al., 2021) only now enables the proteome-wide integration of structural information with PTM data. In this study, we provide a first overview of how the comprehensive structural context of all detected PTMs can provide global and protein-specific insights into biological mechanisms, to filter in vitro datasets for physiological relevance and to identify promising candidates for biochemical follow-up studies.

Key to most of our analyses was that we used whole proteome structural information to determine the exposure of each individual amino acid side chain, thus providing a measure of how amenable that residue is for harboring a modification (Figure 1A). In contrast to experimentally derived structures, the *in silico* structures come with prediction errors and positional uncertainties, which turned out to be crucial for assessing amino acid exposure (Akdel et al., 2021) (Figure 1B/C and Supplementary Figure 1). Our analyses confirmed that most PTMs are strongly enriched on exposed amino acids compared to residues that are buried within the protein fold (Supplementary Figure 2). This effect is even more evident when comparing IDRs and structured regions (Figure 2A) and for specific kinase phosphorylation motifs (Figure 3B and Supplementary Figure 3A).

In contrast to the other analyzed PTMs, ubiquitinations were strikingly enriched on structured regions and on amino acids that are expected to be inaccessible (Figure 2A and Supplementary Figure 2). We showed that this effect is triggered by proteasome inhibition, supporting the idea that the ubiquitinations on structured regions are predominantly placed on misfolded proteins, tagged for degradation. Thus, our toolkit can help distinguish ubiquitination associated with protein quality control from that mediating site-specific regulation. Interestingly, a majority of the proteins with ubiquitin modifications enriched in structured regions are DNA or RNA binders, many of which are known to be ubiquitinated in cellular stress response (Maxwell et al., 2021). As an example, the RNA-binding protein G3BP2, which is essential for cytoplasmic stress granule formation, has all of its nine ubiquitination sites in structured regions (Matsuki et al., 2013). SET8 is ubiquitinated during DNA damage response, causing its degradation followed by chromatin rearrangements (Jørgensen et al., 2011), and our analysis placed 9 of 10 ubiquitination sites in normally structured regions (adj. p-value = 0.01). Together, these results raise the possibility that ubiquitination of DNA and RNA binders and their direct regulators provides an effective regulatory mechanism for cellular stress responses. Further investigations could combine the above analyses with information about different ubiquitin side chain architectures to elucidate possible 3D topologies and associated functionalities.

As PTMs on properly folded proteins are expected to reside on exposed amino acid side chains, we reasoned that our 3D analysis could help to prioritize sites from experiments performed under less than physiological conditions. Here, we exemplified such a case by an *in vitro* kinase substrate screen (Sugiyama et al., 2019). While the screen allowed defining sequence preferences for kinase phosphorylation motifs, determining which sites could mediate bona fide regulation remained a challenge. Modified sites that are observed on inaccessible amino acid residues can be filtered out to reduce the target list to more physiologically relevant sites. Moreover, in our analysis, the motifs of best retained sites were strikingly better defined than those based on the discarded ones (Figure 3D).

We strongly suggest to employ such analyses in future in vitro PTM studies. Apart from phosphorylations, we expect similar benefits for the analysis of other types of PTMs or even for linear motifs involved in specific molecular binding events.

More generally, 3D analysis can highlight sites with a high potential to be functionally relevant. Our unbiased, global analysis revealed that regulatory phosphosites are strongly enriched in short unstructured regions between large folded domains (Figure 4B). Many of these short IDRs correspond to annotated kinase activation loops, which are known to undergo structural rearrangements upon phosphorylation (Nolen et al., 2004). Our analysis systematically reveals such functionally highly relevant sequence regions on a proteome-wide scale, opening up multiple interesting routes for further investigations: First, short IDRs in regions without known functional relevance could specifically be investigated. Second, PTMs lying in or adjacent to such regions could be prioritized in the selection of candidate PTMs for biochemical follow-up studies, given their potential to cause structural rearrangements with functional consequences. We provide the community with a resource of such short IDRs in the human proteome and also with a set of promising PTM candidates from PhosphoSitePlus, which lie within or directly adjacent to these short IDRs. Beyond phosphorylation, these candidate PTMs contain hundreds of ubiquitinations as well as tens of sumoylations, acetylations and methylations (Figure 4G). It would be exciting if future work in protein structure prediction enables the direct prediction of protein folds including PTMs.

Stepping back, the structural context of PTMs can not only be evaluated by means of spatial metrics, such as the exposure and folding state of individual amino acids, but also by using the three-dimensional fold directly. To this end, we evaluated PTM proximity in 3D space, revealing that not only phosphorylations, but also most other PTM types tend to cluster in protein folds (Figure 5A). Beyond this co-localization of PTMs of the same type, we further showed that phosphorylations and lysine modifications preferentially occur in 3D proximity (Figure 5B and 5C), extending previous findings in linear sequence space (Beltrao et al., 2012). Interestingly, lysine modifications, including ubiquitination, sumoylation, acetylation and methylation often compete for the same or directly neighboring residues, but they do not generally favor proximity (Figure 5C). A possible explanation for this effect could be the high reactivity of specific lysine residues, which will be an interesting follow-up to our study (Abbasov et al., 2021). We further investigated PTM proximity on individual proteins in a proteome-wide fashion. This systematic analysis resulted in the identification of interesting proteins with 3D clusters of phosphorylations in specific protein pockets. For instance, in PDHA1, the relevant phosphorylation sites may all be important for enzyme activation and inactivation (Kato et al., 2008). In AKR1B1, the relevant phosphorylations surround the known NADP binding site. We only

observed functionally relevant PTM hotspots in protein pockets for phosphorylations here, but we expect that future improvements in the coverage of other PTM types will also enable the identification of 3D clusters for them. Recent work extends the prediction of structures of individual proteins to the prediction of structures for entire protein complexes (Evans et al., 2021; Humphreys et al., 2021). The integration of proteome-wide estimates of binding interfaces may make studies of PTM co-regulation across proteins, such as the synergistic sumoylation of multiple proteins during DNA repair (Psakhye and Jentsch, 2012), feasible on a global scale.

In summary, our study is the first systematic assessment of the structural context of PTMs and how to leverage this information to gain novel biological insights and to augment proteomic studies. By providing the open-source Python packages StructureMap and AlphaMap, we equip the scientific community with a toolset to easily integrate information from AlphaFold predicted structures into any kind of proteomics study and to visualize proteomics data on the predicted protein fold. We envision that this will enable scientists to directly assess observed PTMs in their structural context and to evaluate their physiological feasibility, to prioritize promising candidates for functional follow-up studies, and to find three-dimensional hotspots of PTM regulation. Future work could explore 3D motifs for enzyme substrate recognition as well as the integration of protein binding information. Beyond our own work presented here, we are convinced that the systematic integration of structural information with proteomics data will open up new avenues to (re-)evaluate both old and new research questions.

Contributions

IB and MM conceived the study, IB performed all analyses; IB and SW implemented the StructureMap package; IB, WFZ and MTS integrated the 3D visualization into AlphaMap; IB, FMH, MT, MM and OK formulated different biological hypotheses and interpreted the results; BAS provided critical feedback to the paper; MM coordinated and supervised the study; IB and MM wrote the paper with input from all authors.

Acknowledgements

This study was supported by The Max-Planck Society for Advancement of Science and by the Bavarian State Ministry of Health and Care through the research project DigiMed Bayern (www.digimed-bayern.de). IB acknowledges funding support from her Postdoc.Mobility fellowship granted by the Swiss National Science Foundation [P400PB_191046].

We thank Julia P. Schessner for valuable discussions about the implementation of the pPSE exposure metric. We also thank J. Rajan Prabu for sharing his knowledge in structural bioinformatics.

Conflict of interest

The authors declare that they have no conflict of interest.

Methods

Prediction of intrinsically disordered regions (IDRs) and benchmarking against IUPred2

To benchmark IDR prediction, we first downloaded ground truth datasets for both disordered and structured protein regions as defined by Beltrao and colleagues (Akdel et al., 2021). To avoid inconsistencies, we removed any amino acids that were labeled as both disordered and structured. We downloaded and formatted the AlphaFold structure predictions for 3062 of the remaining 3080 proteins. The other 18 proteins did not have a complete structure deposited in AlphaFold DB (Tunyasuvunakool et al., 2021; Varadi et al., 2021). As a reference dataset for our benchmark, we obtained direct IDR estimates for all proteins from IUPred2A (<https://iupred2a.elte.hu/>) (Mészáros et al., 2018).

RSA values for all amino acids of the benchmarking proteins were calculated using DSSP from Bio.PDB (version 1.79) using default parameters (Hamelryck and Manderick, 2003). Similar to Beltrao and colleagues, we applied a local smoothing of both the pLDDT and RSA values by averaging along the sequence with a window size of 5, 10, 15, 20 and 25 amino acids (Akdel et al., 2021).

To evaluate the ability of our pPSE metric to predict IDRs, we calculated the pPSE values of all amino acids with different parameter settings. Given the goal of finding IDRs, we chose a constant angle term of 180°. This corresponds to the full sphere exposure, which is independent of side chain orientation. For the distance, we screened over multiple parameters including 12 Å, 16 Å, 20 Å, 24 Å and 28 Å. pPSE values were calculated both with and without considering the PAE. Similar to the pLDDT and RSA values, pPSE values were also smoothed by averaging along the sequence with a window size of 5, 10, 15, 20 or 25 amino acids.

The performance of IUPred2, pLDDT, RSA and pPSE in predicting IDRs was evaluated based on the AUC and TPR at 5% FPR, when screening over different scoring thresholds. The best results in this analysis were obtained by the pPSE metric when considering PAEs and a distance threshold of 24 Å.

Here, a smoothed $pPSE \leq 34.27$ obtained a TPR of 85% at a FPR of 5%. IDR prediction for our proteome wide analysis was therefore performed based on these parameters for pPSE calculation.

For evaluating the benchmark analysis, it is important to keep in mind that the annotations of structured regions and IDRs in the benchmark dataset were unfortunately not associated with a specified amino acid sequence, but only with sequence positions. While this should be correct for most proteins, there might be some differences between the original sequences for which the IDRs were predicted and the sequences used to predict structures by AlphaFold, thus potentially leading to slight inconsistencies.

All data and code for the IDR benchmark analysis will be available as a Jupyter notebook.

Definition of short IDRs

The goal of our ‘short IDR’ detection strategy was to find flexible regions that are embedded into large folded domains. We defined short IDRs as amino acid stretches that we predicted to be IDRs and that comprise maximally 20 residues sandwiched between two flanking structured regions of at least 80 amino acids. Based on our strategy to identify IDRs by a sliding average of the pPSE exposure, some short IDRs only contain very few amino acids that reach the threshold to be classified as IDRs. To account for this effect, we introduced the extended regions of five amino acids on either side of short IDRs. We further expect the extension of short IDRs to be very beneficial in identifying potentially interesting PTMs, given that modifications on residues directly next to a flexible region could also affect its 3D conformation.

pPSE parameters for estimating amino acid side chain exposure

The pPSEs for side chain expose estimation were calculated based on a distance threshold of 12 Å and an angle of 70°. These values were chosen based on considerations about the average size of amino acids of approximately 3.5 Å and side chain flexibility around the direction of the beta carbon. Visualization of the resulting pPSE values showed expected patterns, as can be seen in Figure 1A (middle):

- Amino acids at the core of the protein fold have the highest pPSE values.
- Beta sheets have an alternating pattern of high and low exposure sites.
- Outwards-facing amino acids of alpha-helices have a higher exposure than inwards-facing sites.

- Amino acids in badly predicted sequence regions (e.g. the tailing sequence) have a pPSE value close to zero.

To categorize amino acids into low versus high exposure, we selected a pPSE threshold of 5. This means that amino acids with a $pPSE \leq 5$ were considered to have a high exposure. Supplementary Figure 1B illustrates the distribution of pPSE values across all amino acids in structured protein regions (non-IDRs).

PTM enrichment analyses

All enrichment analyses were performed using the two-sided Fischer's exact test available in `scipy.stats` version 1.7.1 (Virtanen et al., 2020). P-values were subsequently adjusted for multiple testing (Benjamini and Hochberg, 1995) using `statsmodels.stats.multitest.multipletests` version 0.13.0 (Seabold and Perktold, 2010).

Motif analysis with PSSMMatch

The kinase specific phosphosites were extracted from Supplementary Table S2 from Sugiyama et al. (Sugiyama et al., 2019). UniProt IDs were matched to UniProt accessions and the sequence windows corresponding to ± 6 amino acids around each phosphosite were extracted from the current human protein fasta file (version: 2021_03, downloaded on 02.08.2021). Any sites not matching the expected phosphoacceptor residue were removed from the analysis. For each selected kinase, targeted sequence windows were filtered as follows:

- 1) All reported sequence windows with a phosphosite $pPSE \leq 5$. The number of resulting sequence windows is defined as N_{exposed} .
- 2) All reported sequence windows with a phosphosite $pPSE > 5$. The number of resulting sequence windows is defined as $N_{\text{not-exposed}}$. Usually, $N_{\text{exposed}} > N_{\text{not-exposed}}$.
- 3) A randomly selected subset of size $N_{\text{not-exposed}}$ of the reported sequence windows with a phosphosite $pPSE \leq 5$.

Motif analysis was subsequently performed using the PSSMSearch tool for each of the three sets of sequence windows by selecting 'log odds' as scoring method (Krystkowiak et al., 2018; O'Shea et al., 2013).

Gene ontology enrichment analysis with DAVID

All gene ontology enrichment analyses presented herein were performed on the DAVID platform version 6.8 (Huang et al., 2009a, 2009b).

For the enrichment analysis of proteins with short IDRs, we compared the 2454 proteins with short IDRs against the background of all 20053 human proteins with available structural information. Direct Gene Ontology terms of the ‘Molecular function’ category were considered. Filtering was set to an FDR threshold of 0.01, a minimum fold enrichment of 2 and a minimum count of 10.

PTM import from PhosphoSitePlus

Lists of PTMs were downloaded from PhosphoSitePlus on 01.08.2021 (Hornbeck et al., 2015). PTM types include phosphorylations (p), ubiquitinations (ub), sumoylations (sm), acetylations (ac), methylations (m) and the glycosylations O-GalNAc (gl) and O-GlcNAc (ga). Additionally, the set of regulatory PTMs was downloaded and filtered for the modifications mentioned above. For each PTM type, sites were filtered for a selected set of the most common acceptor residues as follows:

- p: S, T, Y
- ub, sm and ac: K
- m: K, R (Note: all types of methylations were grouped together)
- ga, gl: S, T

In rare cases where regulatory sites were reported for a specific PTM type, but no matching entry was found in the according dedicated PTM list, these sites were added there.

The data processing of PhosphoSitePlus data will be available as a Jupyter notebook.

Global PTM proximity analysis

In our global PTM proximity analysis we investigated whether PTM acceptors near a modified amino acid residue are more frequently observed to also be modified compared to more distant residues or random expectation, a concept previously introduced by Beltrao et al. (Beltrao et al., 2012). Here we extended the concept to evaluate PTM proximity in 3D space and to assess both individual PTM types and PTM co-occurrence. Starting from a set of observed modifications, e.g. phosphorylations, we calculate the 3D distance to all observed modifications of either the same type, e.g. also phosphorylations, or of a different type, e.g. ubiquitinations. Importantly, we consider the positional uncertainty of AlphaFold predictions by adding the PAE to each distance. We also generate

5 random permutations, where the same number of modifications are distributed randomly across all possible acceptor residues. Finally, we select distance bins and count the number of modified residues in each of these bins for both the real observations and the randomized background. To ensure that unstructured regions do not bias the results, only structured regions and short IDRs were considered in this analysis (i.e. we removed all IDRs stretching more than 20 amino acids).

For PTM self-proximity we started distance bins at 1 Å ranging up to a maximum of 35 Å in step sizes of 5 Å. For PTM co-localization analysis, we started distance bins already at 0 Å to enable the evaluation of competition for the same site, especially between different lysine modifications such as ubiquitination, sumoylation and acetylation.

Per protein PTM cluster analysis

To find proteins with a significant co-localization of PTMs, we calculated all pairwise distances between the alpha carbons of modified acceptor residues. The matching PAE provided by AlphaFold was added to each calculated distance to account for positional uncertainties. The average distance was subsequently compared to the average distances of 10,000 random permutations among the modified acceptor residues, thus resulting in empirical p-values. These p-values were adjusted for multiple testing (Benjamini and Hochberg, 1995) using `statsmodels.stats.multitest.multipletests` version 0.13.0 (Seabold and Perktold, 2010).

Protein structure visualization with AlphaMap

The 3D structure visualization in AlphaMap was implemented by integrating a Mol* Viewer (Sehna et al., 2021). Code was adapted from <https://github.com/molstar/pdbe-molstar>.

References

Abbasov ME, Kavanagh ME, Ichu TA, Lazear MR, Tao Y, Crowley VM, am Ende CW, Hacker SM, Ho J, Dix MM, Suci R, Hayward MM, Kiessling LL, Cravatt BF. A proteome-wide atlas of lysine-reactive chemistry. *Nat Chem* 2021 1311 2021;13:1081–92. <https://doi.org/10.1038/s41557-021-00765-4>.

Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. *Nature* 2016;537:347.

Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature* 2003;422:198.

Akdel M, Pires DE V, Pardo EP, Jänes J, Zalevsky AO, Mészáros B, Bryant P, Good LL, Laskowski RA, Pozzati G, Shenoy A, Zhu W, Kundrotas P, Serra VR, Rodrigues CHM, Dunham AS, Burke D, Borkakoti N, Velankar S, Frost A, Lindorff-Larsen K, Valencia A, Ovchinnikov S, Durairaj J, Ascher DB, Thornton JM, Davey NE, Stein A, Elofsson A, Croll TI, Beltrao P. A structural biology community assessment of AlphaFold 2 applications. *BioRxiv* 2021:2021.09.26.461876.
<https://doi.org/10.1101/2021.09.26.461876>.

Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Dustin Schaeffer R, Millán C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues A V., Van Dijk AA, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Christopher Garcia K, Grishin N V., Adams PD, Read RJ, Baker D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* (80-) 2021;373:871–6.
https://doi.org/10.1126/SCIENCE.ABJ8754/SUPPL_FILE/ABJ8754_MДАР_REPRODUCIBILITY_CHECKLIST.PDF.

Bekker-Jensen DB, Bernhardt OM, Högberg A, Martinez-Val A, Verbeke L, Gandhi T, Kelstrup CD, Reiter L, Olsen J V. Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. *Nat Commun* 2020;11.
<https://doi.org/10.1038/S41467-020-14609-1>.

Beltrao P, Albanèse V, Kenner LR, Swaney DL, Burlingame A, Villén J, Lim WA, Fraser JS, Frydman J, Krogan NJ. Systematic functional prioritization of protein posttranslational modifications. *Cell* 2012;150:413–25. <https://doi.org/10.1016/J.CELL.2012.05.036/ATTACHMENT/C23EB872-F0EF-436E-8146-15572DBC02C8/MMC1.PDF>.

Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B* 1995;57:289–300.

Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichtlow G V., Christie CH, Dalenberg K, Di Costanzo L, Duarte JM, Dutta S, Feng Z, Ganesan S, Goodsell DS, Ghosh S, Green RK, Guranovic V, Guzenko D, Hudson BP, Lawson CL, Liang Y, Lowe R, Namkoong H, Peisach E, Persikova I, Randle C, Rose A, Rose Y, Sali A, Segura J, Sekharan M, Shao C, Tao YP, Voigt M, Westbrook JD, Young JY, Zardecki C, Zhuravleva M. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of

biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res* 2021;49:D437–51. <https://doi.org/10.1093/NAR/GKAA1038>.

Carroll EC, Greene ER, Martin A, Marqusee S. Site-specific ubiquitination affects protein energetics and proteasomal degradation. *Nat Chem Biol* 2020 168 2020;16:866–75. <https://doi.org/10.1038/s41589-020-0556-3>.

Cohen P. The regulation of protein function by multisite phosphorylation – a 25 year update. *Trends Biochem Sci* 2000;25:596–601. [https://doi.org/10.1016/S0968-0004\(00\)01712-6](https://doi.org/10.1016/S0968-0004(00)01712-6).

Evans R, O’neill M, Pritzel A, Antropova N, Senior A, Green T, Žídek A, Bates R, Blackwell S, Yim J, Ronneberger O, Bodenstein S, Zielinski M, Bridgland A, Potapenko A, Cowie A, Tunyasuvunakool K, Jain R, Clancy E, Kohli P, Jumper J, Hassabis D. Protein complex prediction with AlphaFold-Multimer 2021. <https://doi.org/10.1101/2021.10.04.463034>.

Flatt JF, Stevens-Hernandez CJ, Cogan NM, Eggleston DJ, Haines NM, Heesom KJ, Picard V, Thomas C, Bruce LJ. Expression of South East Asian Ovalocytic Band 3 Disrupts Erythroblast Cytokinesis and Reticulocyte Maturation. *Front Physiol* 2020;11:357. <https://doi.org/10.3389/FPHYS.2020.00357/BIBTEX>.

Hamelryck T. An amino acid has two sides: A new 2D measure provides a different view of solvent exposure. *Proteins Struct Funct Bioinforma* 2005;59:38–48. <https://doi.org/10.1002/PROT.20379>.

Hamelryck T, Manderick B. PDB file parser and structure class implemented in Python. *Bioinforma Appl NOTE* 2003;19:2308–10. <https://doi.org/10.1093/bioinformatics/btg299>.

Hansen FM, Tanzer MC, Brüning F, Bludau I, Stafford C, Schulman BA, Robles MS, Karayel O, Mann M. Data-independent acquisition method for ubiquitinome analysis reveals regulation of circadian biology. *Nat Commun* 2021;12. <https://doi.org/10.1038/s41467-020-20509-1>.

Heffernan R, Dehzangi A, Lyons J, Paliwal K, Sharma A, Wang J, Sattar A, Zhou Y, Yang Y. Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins. *Bioinformatics* 2016;32:843–9. <https://doi.org/10.1093/BIOINFORMATICS/BTV665>.

Hornbeck P V., Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014:

Mutations, PTMs and recalibrations. *Nucleic Acids Res* 2015;43:D512–20.

<https://doi.org/10.1093/nar/gku1267>.

Hu R, Xu H, Jia P, Zhao Z. KinaseMD: kinase mutations and drug response database. *Nucleic Acids Res* 2021;49:D552–61. <https://doi.org/10.1093/NAR/GKAA945>.

Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009a;4:44–57. <https://doi.org/10.1038/nprot.2008.211>.

Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009b;37:1–13. <https://doi.org/10.1093/nar/gkn923>.

Humphreys IR, Pei J, Baek M, Krishnakumar A, Anishchenko I, Ovchinnikov S, Zhang J, Ness TJ, Banjade S, Bagde S, Stancheva VG, Li X-H, Liu K, Zheng Z, Barrero DJ, Roy U, Fernández IS, Szakal B, Branzel D, Greene EC, Biggins S, Keeney S, Miller EA, Fromme JC, Hendrickson TL, Cong Q, Baker D. Structures of core eukaryotic protein complexes. *BioRxiv* 2021:2021.09.30.462231. <https://doi.org/10.1101/2021.09.30.462231>.

Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res* 2004;32:1037–49. <https://doi.org/10.1093/NAR/GKH253>.

Jørgensen S, Eskildsen M, Fugger K, Hansen L, Larsen MSY, Kousholt AN, Syljuåsen RG, Trelle MB, Jensen ON, Helin K, Sørensen CS. SET8 is degraded via PCNA-coupled CRL4(CDT2) ubiquitylation in S phase and after UV irradiation. *J Cell Biol* 2011;192:43–54. <https://doi.org/10.1083/JCB.201009076>.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Židek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. Highly accurate protein structure prediction with AlphaFold. *Nat* 2021 596:783–9. <https://doi.org/10.1038/s41586-021-03819-2>.

Kamacioglu A, Tuncbag N, Ozlu N. Structural analysis of mammalian protein phosphorylation at a proteome level. *Structure* 2021;29:1219–1229.e3. <https://doi.org/10.1016/J.STR.2021.06.008>.

Kato M, Wynn RM, Chuang JL, Tso SC, Machius M, Li J, Chuang DT. Structural Basis for Inactivation of the Human Pyruvate Dehydrogenase Complex by Phosphorylation: Role of Disordered Phosphorylation Loops. *Structure* 2008;16:1849–59.
<https://doi.org/10.1016/J.STR.2008.10.010/ATTACHMENT/6F54FB3C-2DD6-47C2-B4F3-BD4B2928EAA3/MMC1.PDF>.

Krystkowiak I, Manguy J, Davey NE. PSSMSearch: a server for modeling, visualization, proteome-wide discovery and annotation of protein motif specificity determinants. *Nucleic Acids Res* 2018;46:W235–41. <https://doi.org/10.1093/NAR/GKY426>.

Li S, Zhang Q, You G. Three Ubiquitination Sites of Organic Anion Transporter-1 Synergistically Mediate Protein Kinase C–Dependent Endocytosis of the Transporter. *Mol Pharmacol* 2013;84:139–46. <https://doi.org/10.1124/MOL.113.086769>.

Matsuki H, Takahashi M, Higuchi M, Makokha GN, Oie M, Fujii M. Both G3BP1 and G3BP2 contribute to stress granule formation. *Genes Cells* 2013;18:135–46.
<https://doi.org/10.1111/GTC.12023>.

Maxwell BA, Gwon Y, Mishra A, Peng J, Nakamura H, Zhang K, Kim HJ, Taylor JP. Ubiquitination is essential for recovery of cellular activities after heat shock. *Science* 2021;372.
<https://doi.org/10.1126/SCIENCE.ABC3593>.

Mészáros B, Erdős G, Dosztányi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res* 2018;46:W329–37.
<https://doi.org/10.1093/NAR/GKY384>.

Nolen B, Taylor S, Ghosh G. Regulation of Protein Kinases: Controlling Activity through Activation Segment Conformation. *Mol Cell* 2004;15:661–75. <https://doi.org/10.1016/J.MOLCEL.2004.08.024>.

O’Shea JP, Chou MF, Quader SA, Ryan JK, Church GM, Schwartz D. pLogo: a probabilistic approach to visualizing sequence motifs. *Nat Methods* 2013 1012 2013;10:1211–2.
<https://doi.org/10.1038/nmeth.2646>.

Ochoa D, Jarnuczak AF, Viéitez C, Gehre M, Soucheray M, Mateus A, Kleefeldt AA, Hill A, Garcia-Alonso L, Stein F, Krogan NJ, Savitski MM, Swaney DL, Vizcaíno JA, Noh KM, Beltrao P. The functional landscape of the human phosphoproteome. *Nat Biotechnol* 2019 383 2019;38:365–73.
<https://doi.org/10.1038/s41587-019-0344-3>.

Psakhye I, Jentsch S. Protein group modification and synergy in the SUMO pathway as exemplified in DNA repair. *Cell* 2012;151:807–20. <https://doi.org/10.1016/J.CELL.2012.10.021>.

Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. PROC 9th PYTHON Sci CONF 2010.

Sehna D, Bittrich S, Deshpande M, Svobodová R, Berka K, Bazgier V, Velankar S, Burley SK, Koča J, Rose AS. Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res* 2021;49:W431–7. <https://doi.org/10.1093/NAR/GKAB314>.

Sharma K, D'Souza RCJ, Tyanova S, Schaab C, Wiśniewski JR, Cox J, Mann M. Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep* 2014;8:1583–94. <https://doi.org/10.1016/J.CELREP.2014.07.036>.

Stukalov A, Girault V, Grass V, Karayel O, Bergant V, Urban C, Haas DA, Huang Y, Oubraham L, Wang A, Hamad MS, Piras A, Hansen FM, Tanzer MC, Paron I, Zinzula L, Engleitner T, Reinecke M, Lavacca TM, Ehmann R, Wölfel R, Jores J, Kuster B, Protzer U, Rad R, Ziebuhr J, Thiel V, Scaturro P, Mann M, Pichlmair A. Multilevel proteomics reveals host perturbations by SARS-CoV-2 and SARS-CoV. *Nature* 2021;594:246–52. <https://doi.org/10.1038/S41586-021-03493-4>.

Sugiyama N, Imamura H, Ishihama Y. Large-scale Discovery of Substrates of the Human Kinome. *Sci Reports* 2019 91 2019;9:1–12. <https://doi.org/10.1038/s41598-019-46385-4>.

Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, Bridgland A, Cowie A, Meyer C, Laydon A, Velankar S, Kleywegt GJ, Bateman A, Evans R, Pritzel A, Figurnov M, Ronneberger O, Bates R, Kohl SAA, Potapenko A, Ballard AJ, Romera-Paredes B, Nikolov S, Jain R, Clancy E, Reiman D, Petersen S, Senior AW, Kavukcuoglu K, Birney E, Kohli P, Jumper J, Hassabis D. Highly accurate protein structure prediction for the human proteome. *Nat* 2021 5967873 2021;596:590–6. <https://doi.org/10.1038/s41586-021-03828-1>.

Tyanova S, Cox J, Olsen J, Mann M, Frishman D. Phosphorylation Variation during the Cell Cycle Scales with Structural Propensities of Proteins. *PLOS Comput Biol* 2013;9:e1002842. <https://doi.org/10.1371/JOURNAL.PCBI.1002842>.

Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* 2016;13:731–40. <https://doi.org/10.1038/NMETH.3901>.

Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Židek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, Cowie A, Hobbs N, Kohli P, Kleywegt G, Birney E, Hassabis D, Velankar S. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2021. <https://doi.org/10.1093/NAR/GKAB1061>.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Vázquez-Baeza Y, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020 173 2020;17:261–72. <https://doi.org/10.1038/s41592-019-0686-2>.

Voytik E, Bludau I, Willems S, Hansen FM, Brunner A-D, Strauss MT, Mann M. AlphaMap: an open-source Python package for the visual annotation of proteomics data with sequence-specific knowledge. *Bioinformatics* 2021. <https://doi.org/10.1093/BIOINFORMATICS/BTAB674>.

Zannini L, Delia D, Buscemi G. CHK2 kinase in the DNA damage response and beyond. *J Mol Cell Biol* 2014;6:442. <https://doi.org/10.1093/JMCB/MJU045>.