# Structural Analysis of Phosphorylation Sites at a Proteome Level and Their Functional Relevance

by

Altuğ Kamacıoğlu

A Dissertation Submitted to the

Graduate School of Sciences and Engineering

in Partial Fulfillment of the Requirements for

the Degree of

Master of Science

in

Molecular Biological and Genetics

KOÇ ÜNİVERSİTESİ

May 2021

# Structural Analysis of Phosphorylation Sites at a Proteome Level and Their Functional Relevance

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

**Altuğ Kamacıoğlu**

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

---

Asst. Prof. Nurhan Özlü (Advisor)

---

Assoc. Prof. Nurcan Tunçbağ

---

Assist. Prof. Özge ŞENSOY

Date:

*To my family who supports me all the time*

# ABSTRACT

Phosphorylation is an essential post-translational modification for the regulation of almost all cellular processes. Several global phosphoproteomics analyses revealed proteome-wide phosphorylation events and changes in phosphorylation profiles under different conditions. Beyond phospho-site sequence positions identified by proteomic approaches, protein structures add another layer of information to assess the biological relevance of phosphorylation events. In this study, we systematically characterize phosphorylation sites based on their 3D locations in the protein and establish a location map for phospho-sites. More than 250,000 phospho-sites have been analyzed of which 8,686 sites match at least one structure and are stratified based on their respective 3D positions. Core phosphorylation sites possess two distinct groups based on their dynamicity. Dynamic core phosphorylations are significantly more functional compared to static ones. Dynamic core and the interface phospho-sites are the most functional among all 3D phosphorylation groups. Our analysis provides global characterization and stratification of phosphorylation sites from a structural perspective that can be utilized for predicting the functional relevance and filtering out false positives in phosphoproteomic studies.

# ÖZET

Fosforilasyon neredeyse tüm hücre mekanizmasında etkili olan kritik bir posttranslasyon modifikasyondur. Birçok geniş çaplı fosfoproteomik analizi, protein bazındaki fosforilasyonları ve bu fosforilasyonların farklı durumlar altındaki değişimlerini ortaya çıkarmıştır. Protein yapısı, protemik tekniklerle bulunan fosforilasyon bölgelerinin ötesinde, fosforilasyonların biyolojik etkilerinin değerlendirilmesinde ek bir bilgi katmanı sunmaktadır. Bu çalışmada, sistematik olarak fosforilasyon bölgelerinin protein üzerindeki 3 boyutlu konumlarına göre karakterize edip, fosforilasyon bölgerin yerlerine dair bir harita çıkardık. 250.000'den fazla fosforilasyon bölgesinin analiz edilmesi sonucu en az bir protein yapısı bulunan 8686 fosforilasyon elde ettik. Bu elde edilen fosforilasyon bölgeleri 3 boyutlu konumlarına göre katmanlara ayırdık. Merkez fosforilasyon bölgeleri dinamik yapılarına göre iki gruba ayrıldığını bulundu. Dinamik olan merkez fosforilasyonlar statik olanlara göre istatistiksel olarak daha işlevsel oldukları bulundu. Dinamik merkez ile arayüz fosforilasyon bölgelerinin bütün 3 boyutlu gruplar arasında en işlevsel fosforilasyon bölgeleri oldukları gözlemlendi. Bütün fosforilasyon bölgeleri yapısal olarak karakterize edip katmanlara ayıran analizimiz, fosforilasyon bölgelerinin işlevsellikle ilişkisini bulmakta ve fosfoproteomik analizlerde yanlış pozitiflerin ayıklanmasında kullanılabilir.

## ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my thesis advisor Dr. Nurhan Özlü for giving me an opportunity to work in her laboratory. Her utmost support in my master's studies, her guidance, and her knowledge let me achieve a lot during my master. I also would like to thank Dr. Nurcan Tunçbağ for her support in my master's project. I could improve myself in computational biology with her valuable guidance.

I would like to thank Dr. Özge Şensoy for accepting to be a member of my thesis defense jury. Also, I would like to thank all the jury members for their critical readings of my thesis and their comments.

I acknowledge the financial support of the Scientific and Technological Research Council of Turkey (TUBITAK) and Koç University during my master education and research.

I am deeply grateful to Ezgi Memiş for always being there for me. Without her, my master would be very dull. Even studying until night with her was enjoyable. She listened all my whining with patience. We enjoyed the same things even though they seem offensive or meaningless from the outside. She becomes one of my best friends whom I can trust, and I feel her support all the time. Our time together will stay very precious for me all the time.

I would like to specially thank Tuğçe and Beste for their great friendship and support during my studies. I also would like to thank Mehmet for all his support.

I would like to thank one of my best friends Mehmet Ali for always welcoming me to his house, always helping me to overcome the hard times during my master. I was happy to know that you were there for me all the time. I am very grateful to Alican, Altan, Çağatay, Kaan, and Serhan for their great friendship. I am lucky to have you as my best friends.

I would like to thank Güniz for all her love and for being there for me all the time. We could share every memoir let it be happy or sad, and I was confident thanks to her

presence. She was one of the main reasons for me to stay strong against every obstacle. I am grateful to her for the joy and happiness she brings to me.

Lastly, I would like to thank my family Şakir, Sabire, and Melis to whom this thesis is dedicated. I am very lucky to have a family who supports and loves me all the time. Their presence and understanding let me be the person who I am right now. I could not do it without their endless love, encouragement, and support. I owe all my achievements to them.

# TABLE OF CONTENTS

ABSTRACT ..... iv

ÖZET ..... v

ACKNOWLEDGMENTS ..... vi

TABLE OF CONTENTS ..... viii

LIST OF TABLES ..... x

LIST OF FIGURES ..... xi

NOMENCLATURE ..... xiii

Chapter 1 INTRODUCTION ..... 1

Chapter 2 LITERATURE REVIEW ..... 4

2.1 Phosphorylation ..... 4

2.1.1 Kinases and Phosphatases ..... 5

2.1.2 Phosphorylation during Cell Cycle ..... 6

2.1.3 Phosphoproteomics ..... 9

2.2 Protein Structure ..... 11

2.2.1 Biophysical Techniques in Protein Structure Determination ..... 12

2.2.2 Protein Structure Prediction ..... 15

Chapter 3 MATERIALS AND METHODS ..... 17

3.1 Data collection and pre-processing ..... 17

3.2 Data acquisition and location detection ..... 17

3.3 Multi-structural analysis ..... 18

3.4 Analysis of 3D phosphorylation groups ..... 19

3.5 Kinase prediction ..... 20

3.6 Feature Selection ..... 20

3.7 Visualization ..... 21

Chapter 4 RESULTS ..... 22

4.1 Structural stratification of human phosphorylation sites reveals their signature properties ..... 22

4.2 Core phosphorylation sites possess two distinct groups based on their dynamicity ..... 34

4.3 Dynamic core sites are regulated during cell cycle ..... 44

4.4 Conformational change during phosphorylation is associated with functionality 47

Chapter 5 DISCUSSION ..... 51

Bibliography ..... 55

# LIST OF TABLES

**Table 1. Description of all parameters with detailed explanation in table format.** 21

**Table 2. Enrichment of functional sites for each 3D structural group with odds ratio.** ..... 30

# LIST OF FIGURES

**Figure 1. Distribution of phosphorylated residues in dbPAF database.** ..... 5
**Figure 2. The human kinases and their known substrates** ..... 7
**Figure 3. Regulation of cell cycle by CDKs, Aurora and PLK kinases.** ..... 8
**Figure 4. Distribution of protein structure detection techniques in PDB RCSB database.** ..... 14
**Figure 5. Flowchart of structural stratification of phosphorylation sites** ..... 23
**Figure 6. Residue distribution of 3D phosphorylation groups (Surface, Intermediate, Core and Interface) and downstream subgrouping (static core and dynamic core)** ..... 24
**Figure 7. Phosphorylation 3D structure distribution of all detected sites (A) and phosphorylation 3D structure percentage of overlapped sites (B)** ..... 25
**Figure 8. The protein length comparison of phosphorylation 3D structure groups** ..... 26
**Figure 9. Reactome enrichment analysis of interface sites** ..... 26
**Figure 10. The distribution of temperature factors (B-factor) for each phosphorylation types** ..... 28
**Figure 11. Functional scores from Ochoa *et al.* for each phosphorylation groups** . 29
**Figure 12. Comparison of cellular compartmentalization of proteins from each 3D phosphorylation groups with their respective frequencies** ..... 30
**Figure 13. Amino acid type tendency of phosphorylation residues from each group** ..... 31
**Figure 14. Distribution of minimum mutation score (EVscore) across 3D phosphorylation groups** ..... 32
**Figure 15. Distribution of minimum mutation score (EVscore) across 3D phosphorylation groups (orange) and mean 3D nonphospho-sites groups (green)** 33
**Figure 16. Classification of core sites as dynamic or static based on their NMR models** ..... 35
**Figure 17. Classification of core sites as dynamic or static based on their multiple X-ray structures** ..... 36

**Figure 18. Comparison of RSA scores between static (purple) and dynamic (yellow) core sites in selected optimal structures** ..... 37
**Figure 19. Comparison of average normalized B-factors from different X-ray structures for each phosphorylation sites between dynamic and static groups** ..... 38
**Figure 20. The Pearson correlation between RSA score and normalized B-factor** 39
**Figure 21. Comparison of disorder scores in $log_2$ scale between dynamic and static groups** ..... 40
**Figure 22. Mapping phospho-sites from dynamic and static groups to the corresponding secondary structures** ..... 40
**Figure 23. Comparison of functional scores across all 3D phosphorylation groups including static and dynamic core sites** ..... 41
**Figure 24. Percentage of core phosphorylation sites which is entered to Phosphositeplus database** ..... 42
**Figure 25. Unique kinases that phosphorylate each core phosphorylation sites in kinome-tree** ..... 43
**Figure 26. Percentages of cell cycle dependent (blue) and independent (yellow) phosphorylation sites in 3D phosphorylation groups** ..... 45
**Figure 27. Plotting of RSA scores of cell cycle dependent (bottom) and independent (top) phosphorylation sites from dynamic (blue) and static (orange) groups** ..... 46
**Figure 28. Cell cycle dependent phosphorylation dynamics of interface (orange), dynamic (blue) and static (green) core groups based on their fold changes during interphase, mitosis, and cytokinesis** ..... 47
**Figure 29. Ranking of the features based on their weight in functionality prediction of core phosphorylation sites** ..... 48
**Figure 30. 3D illustration of PKM (Phos-S37) and NEK2 (Phos-S184) in different conformations** ..... 49
**Figure 31. 3D illustration of IGF1R (Phos-Y1161) before and after phosphorylation** ..... 50
**Figure 32. Distribution of posterior error probability (PEP) score and best localization score of all 3D phosphorylation groups** ..... 53

# NOMENCLATURE

| | |
|---|---|
| aPKs | Atypical Kinases |
| ANOVA | Analysis of Variance |
| ASA | Accessible Surface Area |
| CASP | Critical Assessment of protein Structure Prediction |
| CDKs | Cyclin Dependent Kinases |
| Cryo-EM | Cryo-electron microscopy |
| dbPAF | Database of Phospho-sites in Animals and Fungi |
| DDA | Data-dependent acquisition |
| DIA | Data-independent acquisition |
| EGFR | Epidermal growth factor |
| ePKs | Eukaryotic Protein Kinases |
| FDR | False Discovery Rate |
| GPS | Group-based Prediction System |
| (HDX)-MS | Hydrogen-deuterium Exchange |
| IGF1R | Insulin-like Growth Factor 1 Receptor |
| iGNM | Gaussian Network Model |
| IMAC | Immobilized metal affinity chromatography |
| INCENP | Inner centromere protein |
| LiP-MS | Limited Proteolysis |
| PDB | Protein Database |
| PKM | Pyruvate Kinase PKM |
| PLK1 | Serine/threonine-protein kinase PLK1 |

| | |
|---|---|
| PSM | Peptide Spectrum Matches |
| PSP | PhosphoSitePlus |
| PTM | Post-translational modification |
| RSA | Relative Solvent Accessibility |
| MAPK | Mitogen-activated Protein Kinase |
| MD | Molecular Dynamics |
| MS | Mass Spectrometry |
| MYT1 | Myelin transcription factor 1 |
| NEK2 | Serine/threonine-protein Kinase Nek2 |
| NMR | Nuclear Magnetic Resonance |
| VMD | Visual Molecular Dynamics |
| XL-MS | Cross-linking-mass Spectrometry |
| XRC | X-ray crystallography |
| WEE1 | Wee1-like protein kinase |

# Chapter 1

# INTRODUCTION

Phosphorylation is a post-translational modification (PTM) that regulates almost all cellular events critical for life: apoptosis, signal transduction, cell growth, and cell division(Cohen, 2000). Approximately 75-90% of the proteome is phosphorylated (Sharma et al., 2014, Ardito et al., 2017). The rapid development of mass spectrometry (MS)-based large-scale phosphoproteomic approaches with high-resolution data acquisition and improved computational analysis uncovered phosphorylation sites at a proteome level and allowed the identification of even low abundant phosphopeptides (Olsen et al., 2006, Cox and Mann, 2008). Combining phosphoproteomics analyses with advanced labeling techniques enabled monitoring of phosphorylation dynamics in cellular activities in a time-dependent manner(Olsen et al., 2006). Most phosphorylation sites identified through high throughput experiments are publicly available in various database repositories (Phosphositeplus, dbPAF)(Ullah et al., 2016, Hornbeck et al., 2015).

As the phospho-site information is largely available for mammalian cells and model systems, a big challenge is to determine the functional relevance of detected phosphorylation events. Rudolf et al approached this challenge in the network context and functionality scores are calculated based on the phosphorylation levels of interactors in the reconstructed signaling network (Rudolph et al., 2016). Toward this aim, Ochoa et al. re-analyzed the phosphoproteome data gathered from different human cell types and tissues by applying strict search criteria. For each quality-controlled phosphopeptide, a functional score was assigned by using a machine learning algorithm that uses multiple parameters including protein abundance, evolutionary relevance, regulatory signatures, and structural properties (i.e., protein length, disorder, accessible surface area (ASA) (Lee and Richards, 1971, Ochoa et al., 2020).

Phosphorylation may trigger solvent accessibility and conformational changes in proteins (Hughes et al., 2001, Birck et al., 1999). Understanding the structural properties of phosphorylation sites provides valuable insight into their function. Multiple studies suggest that phosphorylation events have regional preferences and predominant in intrinsically disordered regions (Iakoucheva et al., 2004). During the cell cycle, dynamic phosphorylation sites are confined to the intrinsically disordered regions in contrast to the regions with regular secondary structures (Tyanova et al., 2013). Recently there is an emerging effort for constructing databases that integrate the phosphorylation information with protein structure (Li et al., 2020, Ramasamy et al., 2020). The structural biology field progresses at a fast pace where the number of structures deposited in PDB still increases exponentially. In line with experimental progress, computational methods predict a set of structures with high accuracy (Senior et al., 2019, Pieper et al., 2014, Yang et al., 2015, Hopf et al., 2019, AlQuraishi, 2019, Senior et al., 2020). Strumillo et al mapped ~500K eukaryotic phospho-sites to protein structures and found that highly conserved phospho-regions are prone to be located in functional parts such as protein's interface or catalytic parts (Strumillo et al., 2019). Despite the limited number of studies mapping the high-throughput phosphoproteomic data with known protein structures and the structure-derived features (Ochoa et al., 2020, Strumillo et al., 2019), it is challenging to find structural information of the phosphorylated state of the protein which can be partially explained by the limitations in experimental techniques, such as crystallization problems in highly flexible solvent-exposed regions in X-ray crystallography. (Renaud et al., 2018, Srivastava et al., 2018) Ultimately, these restrictions hinder the detection of PTMs and lead to a low number of protein structures with PTMs.

In this study, we aim to structurally stratify and characterize phosphorylation events. Towards this aim, we map the phosphoproteome data retrieved from the dbPAF database to known protein structures. (Ullah et al., 2016). In total, 8,686 phospho-sites are mapped to known structures out of 116,258 phospho-sites. We divided the phospho-sites into four 3D phosphorylation groups based on the protein regions they are located in: surface, intermediate, interface, and core. Our analysis revealed that core phosphorylation sites are significantly the least functional among all 3D phosphorylation groups whereas interface ones are the most functional. We called the core phospho-sites, which change their group label from the core to surface or intermediate in different structures, as

"dynamic core sites". We found the dynamic core sites to be significantly more functional than the static ones. These results suggest that dynamicity is one of the main factors for assessing functionality. We confirmed these by using an independent dataset where the interface and dynamic core phosphorylation sites are mostly cell cycle-dependent which indicates their functional relevance. As expected, the effect of mutated phospho-sites in protein stability is the lowest in surface sites compared to the core region. Overall, our study systematically stratifies and characterizes phosphorylation sites based on their structural information while assessing their functional relevance in phosphoproteome.

# Chapter 2

# LITERATURE REVIEW

## 2.1 Phosphorylation

Proteins are responsible for very diverse functionality in the living cell such as regulatory, structural, catalytic, and signaling. However, this varied functionality could not be achieved only by sequence differences and some proteins are non-functional even after translation. Regulation of protein structure and function is controlled by post-translational modifications (PTMs) which bind to specific residue via covalent interaction. These protein modifications regulate protein activity, structure, interactions, and location which control several biological processes(Conibear, 2020, Jensen, 2006, Prabakaran et al., 2012). More than 200 PTM have been identified and the most frequently encountered ones are phosphorylation, acetylation, and methylation(Jensen, 2006).

Protein phosphorylation happens by the addition of a phosphate group to a hydroxyl group via covalent bonding and this procedure is a reversible modification. Phosphorylation is the most frequently Phosphoryl group is added to specific residues which are mostly serine, threonine, and tyrosine (Figure 1)(Ullah et al., 2016). However, it is known that rarely phosphorylation happens in hydroxyl-proline, moreover, phosphorylated histidine residue is detected to be an essential key element in cellular regulatory mechanisms(Fuhs and Hunter, 2017, Reinders and Sickmann, 2005). Tyrosine phosphorylation is known as key phosphorylation in cell metabolism and is typically the epidermal growth factor receptor which is known as tyrosine kinase(Ardito et al., 2017). 30% of all cellular proteins have been estimated to be phosphorylated on at least 1 residue(Pinna and Ruzzene, 1996, Cohen, 2000). More than 200,000 distinct phosphorylation have been detected in human cells(Needham et al., 2019). Kinexun

website predicts additional 760,000 phosphorylation sites to be phosphorylated(Ardito et al., 2017). However, 95% of identified phosphorylation sites have no known biological function which limits current cell signaling knowledge and future therapeutic developments(Needham et al., 2019). Phosphorylation regulates the functionality of proteins, such as, by changing their localization, activity, and interactions. Phosphorylation can change the activity of enzymes and phosphorylation in some proteins can be stimulatory or inhibitory. Several diseases are caused by dysregulation of phosphorylation events, for example, almost all cancer types, diabetes, and Alzheimer's disease(Needham et al., 2019).

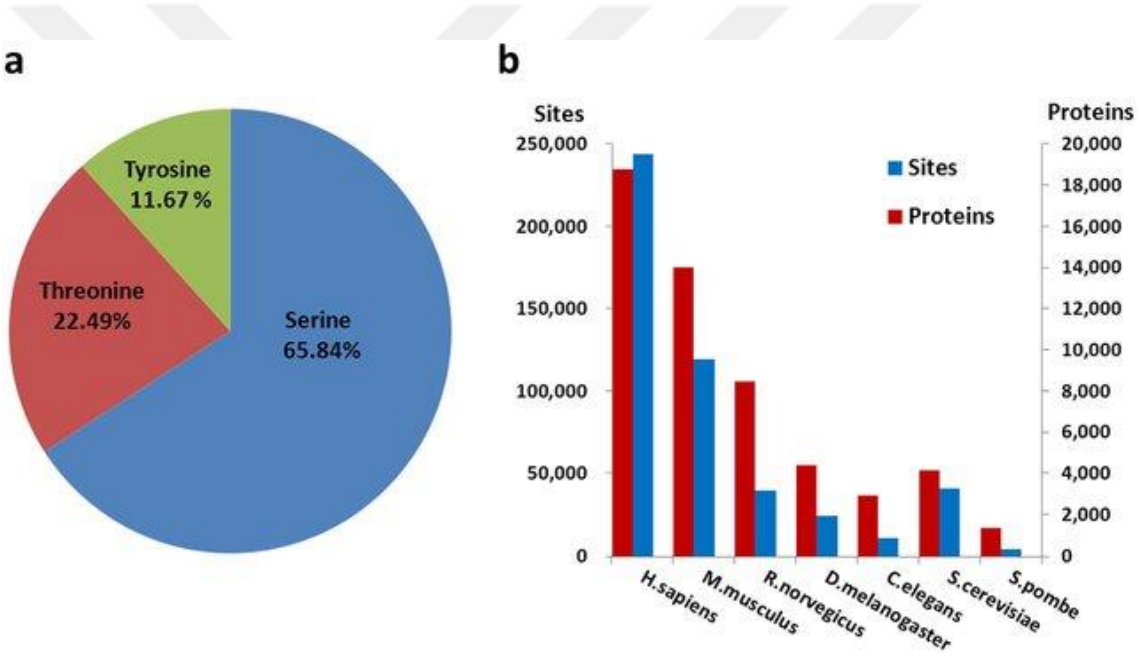



**Figure 1. Distribution of phosphorylated residues in dbPAF database.**

This figure is adopted from Ullah & Xue, 2016.

The removal of the phosphate group from protein is described as dephosphorylation. Phosphorylation and dephosphorylation events are mediated via specific enzymes, kinases, and phosphatases, respectively.

### 2.1.1 Kinases and Phosphatases

The relationship between kinases and phosphorylation sites defines the network architecture of cell signaling and determines its flow. Over 538 kinases which are responsible for serine, threonine, and tyrosine phosphorylation are encoded in the human

genome(Bhullar et al., 2018). Their role in phosphorylation makes them critical in several biological activities, such as transcription, translation, and cell signaling. The protein kinase family is the second-largest enzyme family and the fifth largest in humans. Human protein kinase families are grouped into two categories: eukaryotic protein kinases (ePKs) and atypical kinases (aPKs). Kinases without sequence similarity are categorized as atypical kinases which include lipid kinases(Bhullar et al., 2018). Besides these kinases, 106 kinases lack catalytic domain which is responsible for phosphoryl group addition to proteins called pseudokinase(Roskoski, 2015). 497 kinases are grouped into ePKs which is further divided into 9 groups: AGC, CAMK, CK1, CMGC, Other, RGC, STE, TK, and TKL(Manning et al., 2002). Targets of kinases are extraordinarily unequally distributed; 20% of kinases are associated with 87% of currently annotated substrates which might indicate biases toward well-studied molecules(Edwards et al., 2011) (Figure 2)(Needham et al., 2019). Regulatory problems in kinases through mutations primarily transform kinases to constitutively active forms which cause loss of negative regulators and ultimately several disorders(Johnson, 2009, Zhang et al., 2009).

### 2.1.2 Phosphorylation during Cell Cycle

The cell cycle is the cumulation of sequenced events for the production of two daughter cells. It is controlled by several mechanisms to guarantee correct cell division. The cell cycle in eukaryotic cells divides into two stages: interphase and mitotic which consist of mitosis and cytokinesis. These processes are controlled by cell cycle checkpoints which is a control mechanism to ensure check key elements of the cell cycle. These stages are regulated by cyclin-dependent kinases and cyclin kinases through cell cycle stages which are named $G_1$, S, $G_2$, and M phases. M phases are further divided into 5 phases which are prophase, metaphase, anaphase, and telophase(Schafer, 1998). The last phase is the cytokinesis stage where daughter cells are completely separated(Glotzer, 2005).

Phosphorylation is one of the most prepotent control mechanisms for the cell cycle. The critical role of phosphorylation throughout the cell cycle can be exemplified by cyclin-dependent kinases (CDKs). This evolutionarily conserved process controls cell cycle progression(Suryadinata et al., 2010). For instance, the beginning of cell division is started by the regulation of mitogenic factors in $G_1$ phase cells to activate CDKs(Pardee, 1989).

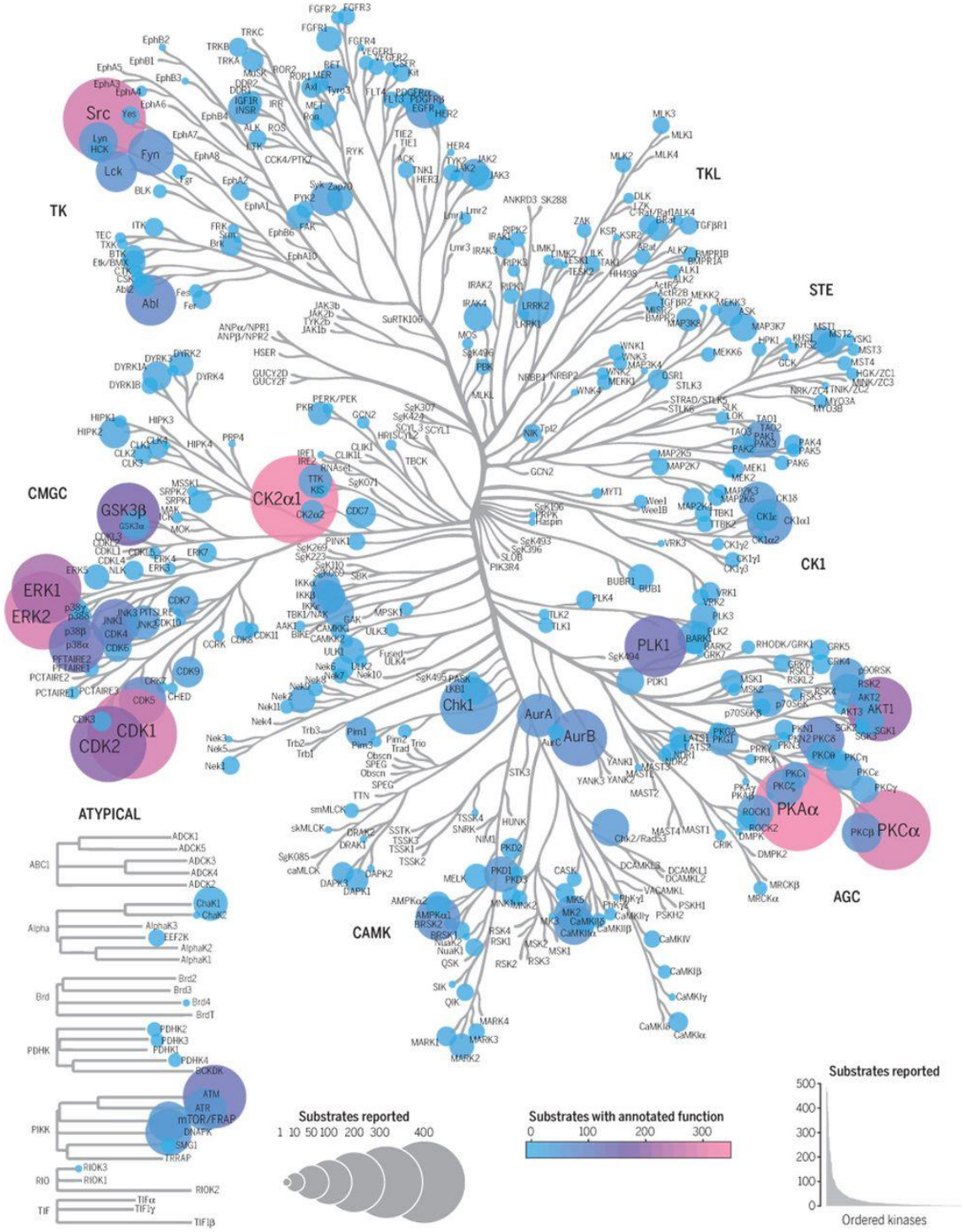



**Figure 2. The human kinases and their known substrates**

This figure is adopted from Needham & Humphrey, 2019.

Kinases play a critical role in regulating cell division dynamically, such as prementioned CDKs, Plks, and Aurora kinases(Lim and Kaldis, 2013, Petronczki et al., 2007, Kollareddy et al., 2008) **(Figure 3)**.

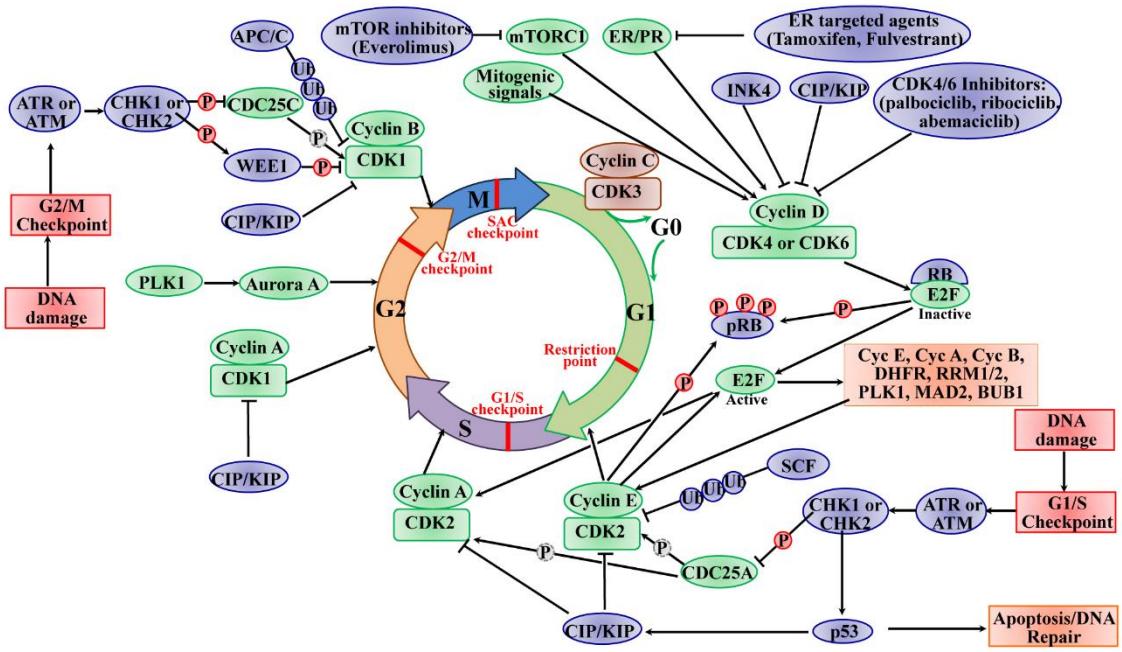



**Figure 3. Regulation of cell cycle by CDKs, Aurora, and PLK kinases.**

This figure is adopted from Ding & Cui, 2020.

More than 20 CDKs are identified to date. CDK2, CDK4, and CDK6 are activated by cyclin proteins during the G1 phase whereas CDK2 in the S phase, and CDK1 is during G2 and M phases (Lim and Kaldis, 2013). Changes in cyclin abundance during the cell cycle alter CDK activities, such as Cyclin B activates CDK1 during the M phase. During progression from G1 to S phase, Cyclin D activates CDK4 and CDK6 while Cyclin E activates CDK2. During G2 to M phase, CDK1 is activated by Cyclin B(Nabel, 2002). Activated CDKs are inactivated by APC (anaphase-promoting complex) during mitotic exit via degradation of cyclins which interacts with CDKs. CDC20 and CDH1 control activity of APC(Petronczki et al., 2007). Changes in cyclin abundance and CDKs activation during cell phases clearly show that cell cycle progression is strictly controlled by CDKs.

Plks (Polo like kinases) have roles in several pathways during cell division(Petronczki et al., 2007). CDK1 and CDC25 are activated by PLK1 via phosphorylation. Additionally, PLK1 phosphorylates WEE1 which causes inhibition and ultimately degradation of MYT1. These steps promote mitotic entry. Key processes for cell division like centrosome duplication, chromosome separation, and maturation are also regulated by PLKs. Moreover, during cytokinesis, PLK1 activates RhoGTPase which is essential for contractile ring contraction(Petronczki et al., 2007).

Aurora kinases are serine/threonine kinases that are essential for cell division. Members of Aurora kinases share similar protein structures, and they are highly evolutionary conserved among species. However, their cellular localization is highly diverse. Aurora A located in the centrosome and spindle microtubules while it functions in spindle formation. Aurora B is located in 3 localization in the cell: kinetochores (from prophase to metaphase), midzone, and midbody while regulating chromosome alignment. Lastly, Aurora C is located at centrosomes and binds to INCENP for surviving of the cell(Willems et al., 2018).

### 2.1.3 Phosphoproteomics

Advancements in MS-based phosphoproteomics enable quantitative, sensitive and, site-specific measurement of phosphorylation sites on a large scale. These developments make phosphoproteomics research a primary protein phosphorylation study method(Ficarro et al., 2002). A study from Olsen *et al.* combined multiple stable isotopes via modified amino acids in cell culture (SILAC) to measure the temporal change in phosphorylation after treatment with epidermal growth factor (EGFR). This study shows the widespread and dynamic nature of signaling pathways with a large proportion of phosphorylation sites (884 sites for only EGFR) were regulated within minutes. Large-scale phosphoproteomics researches have continued to broaden our view of phosphoproteome by revealing its complexity and huge scale. Kinases and phosphatases often affect a large portion of phosphoproteome instead of their direct substrates(Bodenmiller et al., 2010).

Ongoing developments for the deeper sampling of proteome make it possible to measure almost all proteins in a model organism(Mann et al., 2013, Richards et al., 2015). Phosphoproteomics is predicted to undergo the same transition which allows a shift from quantitative and categorizing phosphoproteome to completely understanding the nature

and functional implications. Innovations in MS technologies are steadily increasing the phosphoproteomics studies.

Phosphoproteomics analyses are performed in the following order: cell lysate is obtained in the presence of phosphatase inhibitor to block dephosphorylation. Next, cell lysate is fractioned to purify proteins and by using proteases, proteins are degraded to peptides. Trypsin is the most commonly used protease which cleaves C termini of lysine or arginine. Peptides are further fractioned and enriched for a higher detection rate.

To overcome the low abundance of phosphorylation, several phosphorylation enrichment methods are used which can be divided into peptide-level enrichment and protein level enrichment(Doll and Burlingame, 2015). Ionic interaction-based PTM enrichment strategies consist of IMAC (immobilized metal affinity chromatography) and TiO2 (titanium dioxide) which are based on binding of metal cation to negatively charged phosphopeptides(Pinkse et al., 2004, Gruhler et al., 2005). To increase enrichment, new methodologies are adopted by combining two methods which are termed as sequential elution from IMAC (SIMAC)(Thingholm et al., 2008). However, this enrichment can only detect serine and threonine. Tyrosine phosphorylation can be detected by antibody-based PTM enrichment methods. Protein level enrichment includes gel-eluted liquid fraction entrapment electrophoresis where proteins are separated based on their molecular weight, and LC techniques affinity chromatography, size-exclusion chromatography, ion-exchange chromatography, reverse phase chromatography, and online reversed-phase LC tandem MS approach(Eliuk et al., 2010, Zhang and Ge, 2011).

Even though improvements in sensitivity and acquisition speed, the dynamic range of phosphoproteomics remains a challenge for mass spectrometry. The wide dynamicity of phosphoproteome hinders the usage of popular data-dependent acquisition (DDA) which full scan all ions (MS1) and fragment top-N ions further (MS2). DDA causes the loss of several low-abundant peptides which could be phosphorylated. An increase in instrumental speed enables usage of data-independent acquisition (DIA) which combines the ease of DDA methods with the robustness of targeted MS technologies(de Graaf et al., 2015). In DIA methods, the MS cycles through a defined mass/charge section and acquires all precursors for each segment which reduces the problem of missing values. DIA would be a useful approach for future studies(Needham et al., 2019).

Data which is generated by large-scale phosphoproteomics are collected into databases which become a key resource for several studies. One of the largest and most active databases is PhosphoSitePlus which contains 233,295 phosphorylation sites that are annotated with kinase and functional information(Hornbeck et al., 2015). 29 % of these sites have been identified by more than one mass spectrometry-based study. However, several issues arise for phosphorylation sites in databases. Firstly, some phosphorylation sites are detected under extraordinary conditions which are normally not phosphorylated. Some sites can be hyperphosphorylated due to the inhibition of phosphatases. The last issue is the accumulation of false positives due to incorrectly localized phosphorylation sites in large-scale phosphoproteomics studies(Needham et al., 2019). These limitations advocate for the repository for raw data from MS-based phosphoproteomics studies, like ProteomeXchange which reanalyze data and re-estimate false localization and false discovery rates(Vizcaino et al., 2014). The phosphorylation of noncanonical residues remains under-investigated.

## 2.2 Protein Structure

Proteins are polymers which are constructed by three-dimensional formation of atoms in sequenced amino acids. 4 physical interaction decided the 3D structure of proteins: ionic interactions, hydrogen bonding, Van der Walls forces, and hydrophobic packing. The protein structure is divided into 4 formations. Primary structure refers to the amino acid sequence(Sanger, 1952). Secondary structure is highly regular sub-structures on polypeptide backbone chain. There are two main secondary structures: α-helix and β-sheets which are constructed via hydrogen bonds(Pauling et al., 1951). The tertiary structure of a protein refers to the three-dimensional structure of a single protein molecule which is driven by hydrogen bonds, hydrophobic interactions, and disulfate bonds. Lastly, quaternary structure refers structure which is consists of more than one subunit (multimer).

Proteins consist of several structural units, such as domains, motifs, and folds. Protein domains are conserved protein sequence and tertiary structures which can exist, evolve, and function independently from the remaining protein structure. These regions are stable and folded independently(Wetlaufer, 1973, Richardson, 1981, Bork, 1991). Motifs refer to a short segment of 3D structure or sequence found in several proteins. Protein fold is

general protein architecture such as helix bundle and Rossmann fold(Govindarajan et al., 1999).

Proteins are not static objects; they dynamically change their conformation states. Dynamicity and conformational change functionality as biological machineries, such as myosin. Even though proteins are thought of as relatively stable tertiary structures which undergo conformational changes, proteins have varying stability and some of them are intrinsically disordered proteins(Iakoucheva et al., 2004). These regions lack stable tertiary and secondary structures. Therefore, they are highly dynamic and this feature leads to the development of conformational ensembles to provide a more accurate and dynamic representation of proteins(Varadi et al., 2015). Molecular dynamics (MD) provides a routine to the understanding of dynamical information by stimulating protein motions(Adcock and McCammon, 2006). Discovery of NMR relaxation, fluorescence spectroscopy, and time-resolved X-ray crystallography yield information about the conformational flexibility of proteins (Kempf and Loria, 2003, Weiss, 1999, Schotte, 2003).

### 2.2.1 Biophysical Techniques in Protein Structure Determination

Structural biology techniques emerge with the incorporation of molecular biology principles, biochemistry, and biophysics to elucidate the molecular structure and their dynamics. The three-dimensional structure of proteins leads better understanding of biological laws, disease mechanisms, and ultimately diagnostic and therapeutic agents. There are 3 main techniques for structural biology: X-ray crystallography (XRC), nuclear magnetic resonance (NMR), and cryo-electron microscopy (Cryo-EM). All three of them have different advantages and limitations. All structures obtained by these techniques are stored in the PDB RCSB database(Rose et al., 2017). Most of the protein structures in the PDB database are detected by X-ray crystallography **(Figure 4)**.

X-ray crystallography determines the position and composition of atoms in a crystallized protein by X-ray. The most used method is single-crystal X-ray diffraction where X-ray beams are sent to crystallized atoms and scattered based on the position of atoms. Scattered beams land on the detector which produces a diffraction pattern. By rotating the crystal, the angle and intensity of diffracted beams are measured which creates a three-dimensional image of electron density. According to electron density, the average position of atoms, chemical bonds, and various other information is calculated. With high

purity, homogeneity, and regularity atomic positions can be detected within a few thousandths of an angstrom. Most of the macromolecule structures are detected by X-ray diffraction. The procedure divides into four steps. Firstly, a high-quality single crystal of the protein needs to be obtained. Protein aggregation and nucleation are obtained by supersaturation of solubilized protein. Qualified crystals should be larger than 50 μm in all dimensions and have regular size, no cracks, or twins(Newman, 2006). The crystallization step is the most limiting step for a high-quality structure. The second step is the diffraction experiment. Crystals are immobilized by highly intense X-ray beams which produce diffraction patterns and ultimately diffraction data. Crystal is rotated gradually to obtain diffraction intensity at each point. In the next step, the diffraction pattern is analyzed, and data is fitted to an electron density map. In the last step, the protein model is created from an electron density map. XRC may produce high atomic resolution and does not have any molecular weight restriction(Kuhlbrandt, 2014). It is also suitable for membrane proteins, soluble proteins, and macromolecules. In high-quality XRC results, the position and structure of active sites can be detected which is essential for the understanding of protein binding. However, there are some limitations. First of all, the sample needs to be crystallized where some proteins cannot be crystallized. Additionally, large molecules are hard to crystallize, particularly membrane proteins due to their large size and poor solubilization. Crystallization must allow appropriate diffraction. Lastly, 3D structure solely represents a static form of the molecule instead of a dynamic(Wang and Wang, 2017).

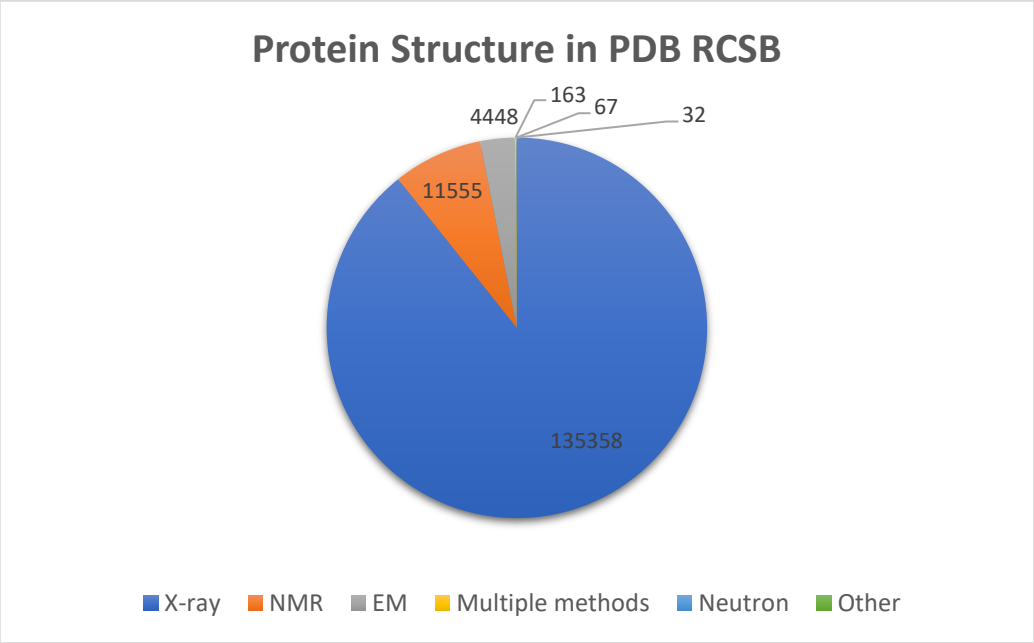



**Figure 4. Distribution of protein structure detection techniques in PDB RCSB database.**

Nuclear magnetic resonance (NMR) is the second most utilized method. It bases on nuclei instead of electrons and it exploits different resonance frequencies of the atomic nucleus. The movement of the nucleus depends on surrounding atoms where NMR uses this feature to detect the distance between atoms in other words structural information of the molecule. For example, each secondary structure reflects a different arrangement of atoms. Therefore, spacing between atoms varies based on the structural features of proteins, such as secondary structure, the interaction between nuclei, and the dynamicity of polypeptides. These nuclear features provide characteristic NMR signals and interpretation of these signals by computational methods provides three-dimensional structure. NMR experiment has 4 main steps. Firstly, high purity, high stability, and high concentration of protein are obtained in an aqueous sample. The usage of stable isotopes can increase signal intensity and resolution. Next, a multidimensional NMR experiment is used for data acquisition. Acquired spectral data is processed to ascertain the atoms of the proteins for each spectral peak from different NMR spectra. Lastly, structural analysis is performed to calculate the spatial structure by geometric and dynamic methods (Bothwell and Griffin, 2011). The most important advantage of the NMR method is that it allows the detection of 3D structure in the natural state from solution. This feature allows the acquisition of information about dynamics and intermolecular interactions. The resolution can be low, such as sub-nanometer. The biggest disadvantage is large

molecules are hard to interpret which limits the application of NMR analysis in large biomolecules. Moreover, NMR needs large amounts of pure samples for a low noise ratio which is challenging for low abundance proteins(Rankin et al., 2014).

Cryo-electron microscopy (Cryo-EM) technique includes 3 different methods: single particle analysis, electron tomography, and electron crystallography. Electron scattering is the underlying mechanism of Cryo-EM. Samples need to be prepared through cryopreservation for analysis and coherent electrons are used as a light source for detection. The electron beams pass through the sample and the lens system converts scattered signal to the magnified image which is recorded on the detector. After signal processing, a 3D structure is obtained. Negative staining EM can be exploited to screen homogeneous samples rapidly to pick particles. The Cryo-EM single particle analysis starts with sample vitrification where protein solution is instantly cooled to not crystallize water molecules. By this step, an amorphous solid is obtained. The frozen sample is screened, and data is collected. A series of two-dimensional images are taken. Based on these 2D images, particle alignment and classification are carried out. Lastly, data is processed by reconstruction software to generate a 3D structure(Carroni and Saibil, 2016). The rapid freeze step allows the sample to be closer to its native state. A small amount of sample is enough and allows flexibility in sample purity. Lastly, proteins do not have to be crystallized. However, this technique detects particles in unknown orientations. High levels of noise are another problem that limits the detection of orientations of small particles. These restrictions decrease the usage of Cryo-EM and it stays as a method for detection of large complexes with low-resolution (Wang and Wang, 2017).

### 2.2.2 Protein Structure Prediction

Since the number of protein structure is lower than the number of proteins itself, the prediction of 3D structure from amino acid sequence has been utilized. Even though the accuracy of the predictions remains modest, advancements in computing power and databases increase the accuracy rapidly. The development of accurate protein structure prediction is going to revolutionize engineer novel proteins. To observe the development of protein structure prediction, Critical Assessment of protein Structure Prediction (CASP) taking place every two years since 1994(Moult et al., 1995).

New machine learning algorithms are used to predict structurally interacting residues from sequence information by analyzing patterns of correlated mutations(Jones et al., 2015). Improvement of proteins' energy functions allowed approximate structure prediction and close the gap between experimentally determined structures(Huang et al., 2017). The rapid increase in CPU-based and GPU-based computational power enables the adaptation of new algorithms with high iterations. Moreover, deep-learning algorithms are employed which enhance development speed and accuracy drastically. Finally, AlphaFold predicted new protein structures with high accuracy which is compatible with experimental structures(Senior et al., 2020).

Predictions divide into two categories template-based modeling and template-free modeling. Template-based uses mutations, insertions, and deletions present in target-template alignment to decipher changes in protein structure, ultimately protein structure. Template-free modeling consists of 5 steps. The first step is the construction of multiple-sequence alignment to find correlated patterns of sequences. Secondly, local structures are predicted, such as psi, phi, and secondary structure. Local structures are assembled for model building, multiple sequence alignment is used to predict residue pairs based on observation of correlated mutations. Local structure and residue contact predictions are used as a guide for 3D modeling with techniques like distance geometry, gradient-based optimization, and fragment assembly. Lastly, these predictions lack representation of coarse-grained energy function. To increase the determination of structure as a near-native structure, these models are refined with all-atom energy function, then compared with each other to find low-energy conformations(Kuhlman and Bradley, 2019).

# Chapter 3

# MATERIALS AND METHODS

## 3.1 Data collection and pre-processing

In this study, we retrieved 244,034 phosphorylation sites of human proteins from dbPAF databases (version 1.0) (Ullah et al., 2016). However, due to lack of elaborative quality-control for phosphorylation sites in databases (Needham et al., 2019), we additionally used a dataset published by Ochoa et al. (Ochoa et al., 2020) where a list of 307 human datasets was reanalyzed and filtered accordingly which contain a total 116,259 phosphorylation site. 1181 residues in Ochoa et al. do not map to canonical protein sequence therefore they are eliminated from the dataset. For inspection of cell division specific dataset, we used Karayel et al. (Karayel et al., 2018) which shows the regulation of phosphorylation throughout cell division stages; interphase, mitosis, and cytokinesis. Karayel dataset which contains 5410 phosphorylation sites was categorized based on significant phosphorylation change during cell division as cell cycle dependent and independent. All data and statistical analyses were performed with Python 3.8.

## 3.2 Data acquisition and location detection

Cross-referencing from protein identifiers of phosphorylation sites to known protein structures in PDB were retrieved from Uniprot(UniProt, 2019) database. The structure resolution and coverage information were also retrieved. To obtain the exact location of a phosphorylation site in a protein complex with high precision, protein structures with the best resolution and the highest coverage were selected to be main structure and obtained from Protein Database (PDB)(Rose et al., 2017) to ascertain the location of phosphorylation sites. Missing protein structures were included by using ModBase(Pieper et al., 2014) homology models with higher than 1.1 ModPipe Protein Quality Score. The relative surface area (RSA) of each phosphorylation site was

calculated by freeSASA (Mitternacht, 2016) software with -RSA option. Phosphorylation sites were divided into four categories based on their RSA score: core, interface, intermediate and surface. If the RSA score of a phosphorylation site is higher than 50 in the monomer state, they were labelled as surface. If RSA score is less than 50 but higher than 5, those phosphorylation sites were labelled as intermediate. We checked the RSA score difference between monomer and complex state for each phospho-site. If a phospho-site has RSA greater than 5 in monomer state and its RSA is less than 5 in complex state, it is labeled as interface phospho-site. Therefore, they were excluded from the surface and intermediate phosphorylation sites. Remaining phosphorylation sites were labelled as core. Distribution of phosphorylation structure in the database and Ochoa et al. dataset was tested by using chi-square test.

Each model in NMR structures was extracted as separate structures. NMR structures cover a smaller part of the protein than X-ray structure and the difference between structural coverage hinders RSA score calculation and residues are calculated much closer to the surface due to lack of outer surface of protein (Yu, 1999). Absence of core location in some models was detected which is caused by protein size limitation of NMR technique. This limitation led to the identification of core sites as intermediate or surface in all models. Therefore, core phosphorylation sites without core structure models were excluded to detect the accurate dynamic change in these sites.

### 3.3 Multi-structural analysis

To accurately investigate dynamic change, we proceed with structures which have resolution less than 4 Å. We selected protein structures which cover more than 90% of the main structure with the highest coverage to prevent variation in the relative location of phosphorylation caused by the protein size difference. Core sites were filtered if there are less than three RSA scores to obtain more transition for core phosphorylation structure as well as to increase the accuracy of structure dynamicity. Core phosphorylation sites that have structures located other than core was categorized as dynamic core phosphorylation and remaining ones as static.

Secondary structure and disorder prediction scores were calculated from protein sequences for each core phosphorylation site by using NetSurfP-2.0(Klausen et al., 2019). Detailed secondary structure information (q8) was utilized to determine the secondary structure of core phosphorylation sites. Distribution of secondary structures between

static and dynamic core phosphorylation sites was tested by chi-square test. Significance of disorder score comparison between static and dynamic core phosphorylation was evaluated by Welch's t-test. Phosphositeplus database was used to control which phosphorylation site had been reviewed.

### 3.4 Analysis of 3D phosphorylation groups

We retrieved experimental B-factors from iGNM database (Li et al., 2016) which contains data in PDB in addition to their predicted values. Since B-factor values vary with each structure, we compared the relative B-factor value of phosphorylation sites acquired by following (B-factor/mean (total B-factor in a structure)). B-factors with higher than 3 and less than -3 Z-score were eliminated as outliers. Log2 values were taken for each B-factor to transform their log distribution to Gaussian distribution. Functional scores for phosphorylation sites were obtained from Ochoa *et al.* (Ochoa et al., 2020).0.5 functional score is determined as functionality cut-off. Both scores were tested using ANOVA and Tukey's honest significance test was applied, afterwards.

Main cellular location of proteins with phosphorylation sites was obtained from Human Protein Atlas (Uhlen et al., 2015). Top 10 sites for each type of phosphorylation sites were chosen for comparison. Chi-square test was used as hypothesis testing for occurrence in specific locations.

A score was defined to represent the relative serine, threonine, and tyrosine distribution of phosphorylation sites in different groups.

$relative_{i,j} = \frac{count_{i,j}}{count_j}$ where *i* is the residue type and j is the region (core, interface, surface or intermediate).

$score_{i,j} = relative_{i,j} - \mu_i$ where μ is the average value of relative scores of each residue in all regions. A negative score means the depletion of residue i in region j.

Effect of mutations in phosphorylation regions was retrieved by using the EVmutation web server (Hopf et al., 2017). In EVmutation, effects of residue change to other types of residues in proteins which are available in the server are provided. We selected the minimum value for obtaining the most damaging mutation. The significance of the mutation effect was evaluated by ANOVA and Tukey's honest significance test. For non phospho-sites, we selected the same proteins and structures with phospho-sites whose

structure and mutational effect is known. For mutational comparison, the same number of phospho-site was selected for 100 times and their mean value was compared with phospho-sites.

.

### 3.5 Kinase prediction

Kinases were predicted for each phosphorylation site by GPS 5.0(Wang et al., 2020) with the highest threshold for high reliable predictions. Predicted kinases were further filtered by scores which were generated by followed calculation:

$$1 < \left(\frac{GPS\ score - GPS\ cutoff}{GPS\ cutoff}\right)^2$$

Unique kinases to specific phosphorylation types were selected. Kinases were labelled in a kinome-tree by using CORAL (Metz et al., 2018).

### 3.6 Feature Selection

Categorical supervised learning was utilized with 2 classes: functional phosphorylation sites (>=0.5 functional score) as 1 and non-functional sites (<0.5 functional score) as 0. For each class, every phosphorylation site was annotated by 23 structure-related parameters (19 continuous variables and 3 discrete variables). Continuous variables were filled by fancyimpute package in Python to avoid loss of limited core phosphorylation sites. StandardScaler function in sklearn package was used to scale all continuous variables. 2 discrete variables were transformed into dummy variables to feed categorical parameters as numeric variables to the system.

10% of core phosphorylation sites were separated as a test dataset. The supervised machine learning model for classification was constructed with different machine learning methods in scikit-learn package(Pedregosa et al., 2011). Random forest classifier predicted test datasets with the highest $R^2$ score with 10 grid cross-validation which is integrated as KFold in GridSearchCV function of sci-kit learn, and all analysis was continued with random forest. Unequal distribution of classification limited prediction of functional core phosphorylation sites and the problem was overcome by oversampling the minor group by SMOTE and under-sampling the major group by RandomUnderSampler function(Blagus and Lusa, 2013). Description of all parameters are shown in **Table 1**.

| Feature | Description |
|---|---|
| RSA change amount | Maximum RSA difference between X-ray structures |
| X-ray B-factor (All structures) | Average normalized B-factor of all X-ray structures |
| P(Hydrogen Bonded Turn) | Probability of located in a hydrogen bonded turn structure |
| X-ray B-factor | Normalized B-factor of optimum X-ray structure |
| P(Turn) | Probability of located in a turn structure |
| Disorder Score | Probability of being in a disordered region |
| Core Status | Being a dynamic core site |
| Psi | Psi angle of core phospho-site |
| Y | Being a tyrosine residue |
| P(β-Sheet) | Probability of located in a beta sheet structure |
| P(α-helix) | Probability of located in a alpha helix structure |
| Pkinase | Being a Pkinase domain |
| P($3_{10}$ – helix) | Probability of located in a $3_{10}$ helix structure |
| RSA | RSA score of optimum X-ray structure |
| Maximum Mutation Score | Maximum EVmutation score (least damaging) from residue substitution |
| P(Bend) | Probability of located in a bend (non-hydrogend-bond based secondary structure) structure |
| P(Coil) | Probability of located in a coil structure |
| P(π – helix) | Probability of located in a pi helix structure |
| P(General state Coil) | Probability of located in a coil grouped structure |
| Minimum Mutation Score | Maximum EVmutation score (most damaging) from residue substitution |
| Phi | Phi angle of core phospho-site |
| T | Being a threonine residue |
| HSP90 | Being a HSP90 domain |
| Actin | Being a Actin domain |
| HSP70 | Being a HSP70domain |
| Carb_anhydrase | Being a anhydrase domain |

**Table 1. Description of all parameters with detailed explanation in table format.**

### 3.7 Visualization

Protein structures were visualized by VMD (Humphrey et al., 1996) Chain that contains the phosphorylation site was chosen to show the monomer location of the phosphorylation site. Proteins with phosphorylation sites were visualized with Gaussian surface and secondary structure visualization techniques. Secondary structure was colourized with temperature factors. Phosphorylation site was visualized according to Van der Waals.

# Chapter 4

# RESULTS

## 4.1 Structural stratification of human phosphorylation sites reveals their signature properties

We used 244,034 phospho-sites in dbPAF and 116,259 phospho-sites in Ochoa dataset where 89,318 of them overlap between two datasets. Despite the high overlap, Ochoa dataset is quality controlled that filters out potential false positives where PSM level FDR was set to 1% and phospho site localization probabilities below 75% were filtered out indicating high confidence localizations (Ochoa et al., 2020).

We divided the phospho-sites in dbPAF and Ochoa dataset into four 3D phosphorylation groups based on the region they are located obtained by the calculation of their relative solvent accessibility: surface (highly accessible), intermediate (partially accessible), core (buried) and interface (binding site) regions. Our structural data contains 28,301 phospho-sites (4989 proteins) in dbPAF database and 8686 phospho-sites (1020 proteins) in Ochoa et al., dataset that map at least to one known protein structure, 5624 phospho-sites are common in both datasets. Since Ochoa et al. applies quality control filters to minimize the noise caused by the false positive phosphopeptides and false phospho-site localization, we performed the rest of the analysis using Ochoa et al. dataset. For proteins having at least one phospho-site we additionally collected their different structural conformations either in multiple X-ray data or in different models in NMR data, if available. The core phosphorylation sites are separately labelled as ‘dynamic’ if they change their location in different structural conformations, otherwise as ‘static’ **(Figure**

**5)**.

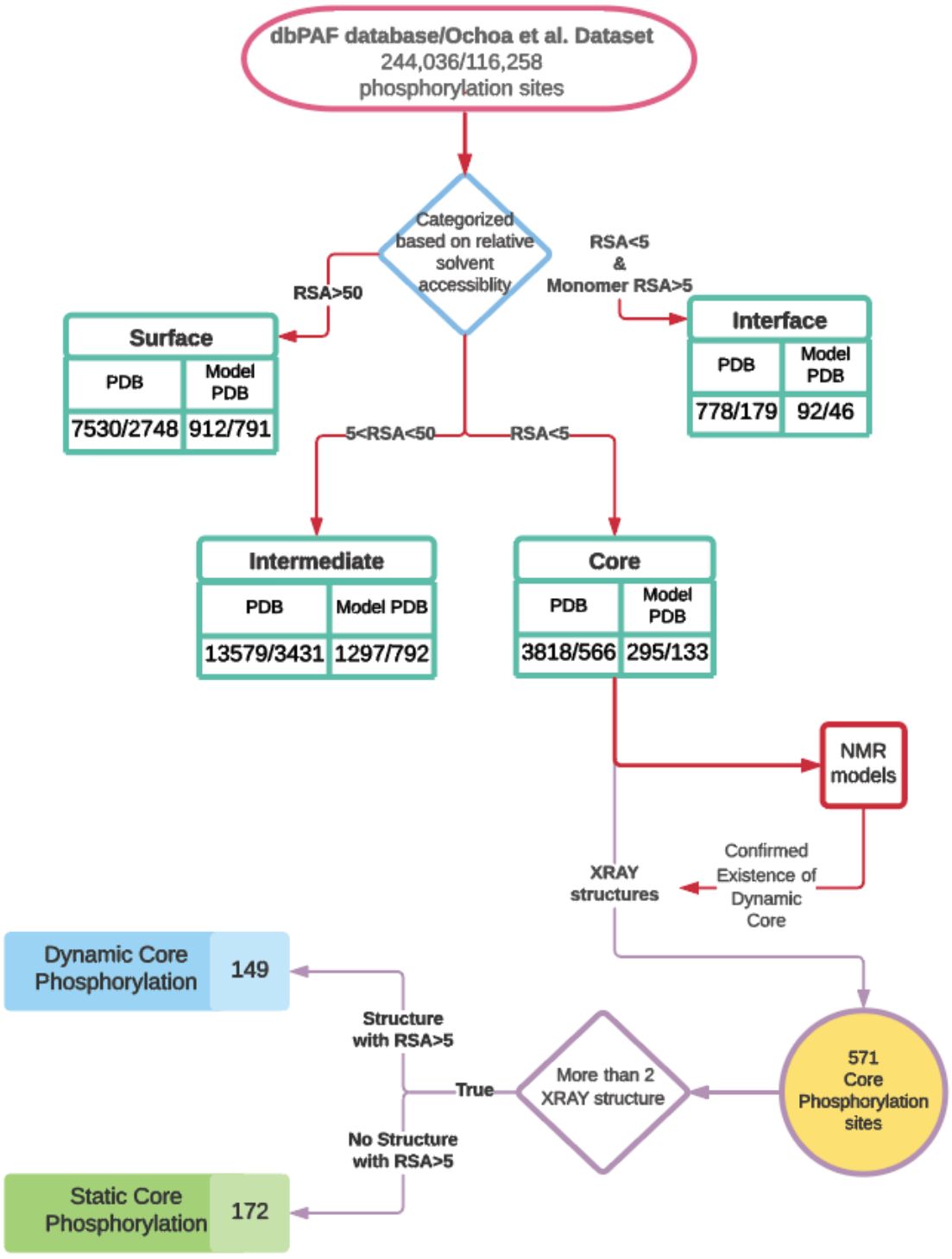



**Figure 5. Flowchart of structural stratification of phosphorylation sites**

Detection of 3D phosphorylation groups (Surface, Intermediate, Core and Interface) based on residue relative solvent accessibility (RSA) and downstream subgrouping (static core and dynamic core) are presented as a flowchart. The number of identified

phosphorylation sites from dbPAF/Ochoa et al., datasets are indicated for each group with known (PDB) and predicted (Modbase) structural information.

We triaged to phospho-sites to examine how the different phospho-residue types classified in 3D groups **(Figure 6)**. We observed that tyrosine residues are more frequent in dynamic core regions than in static ones albeit this is not statistically significant.

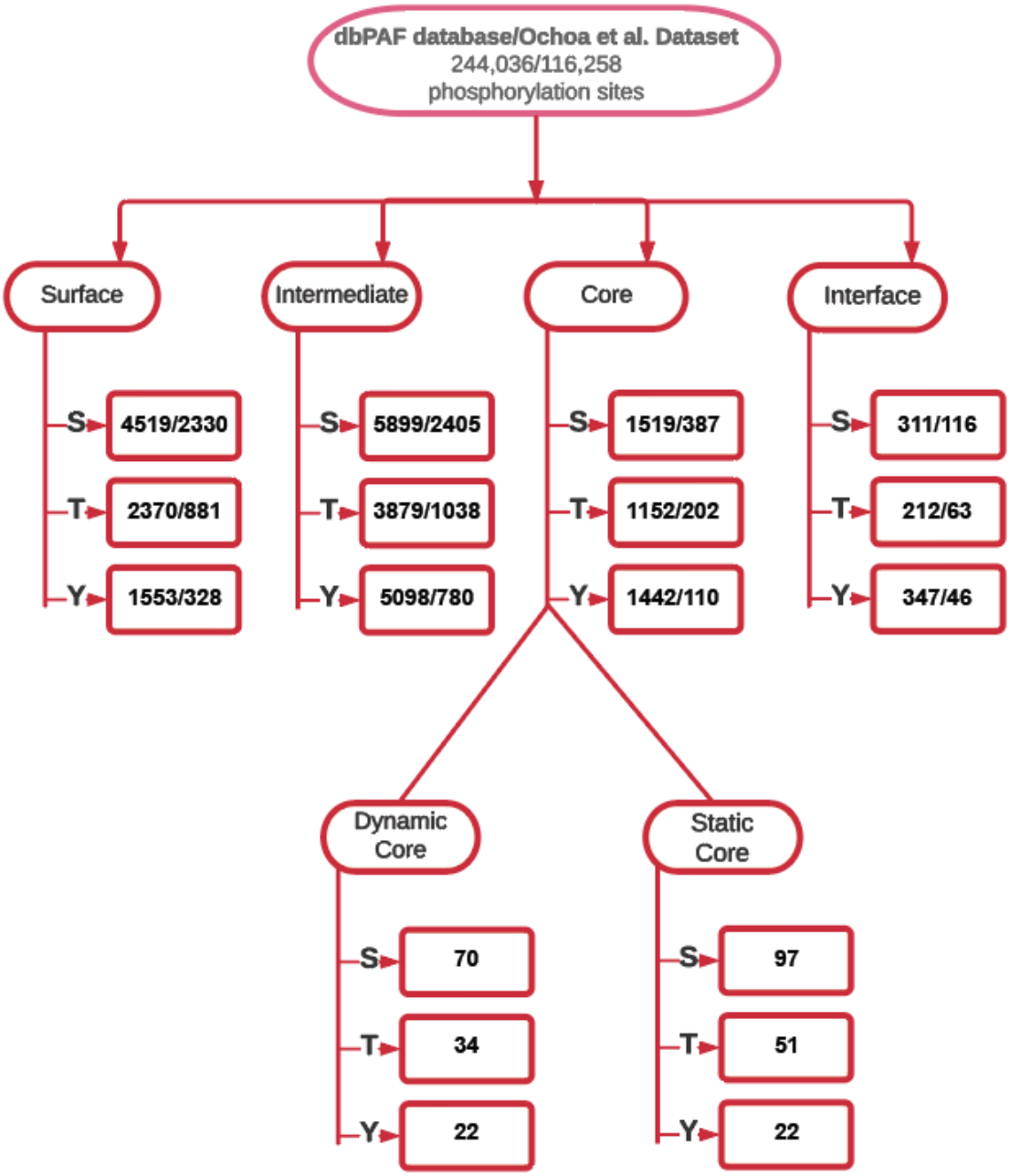
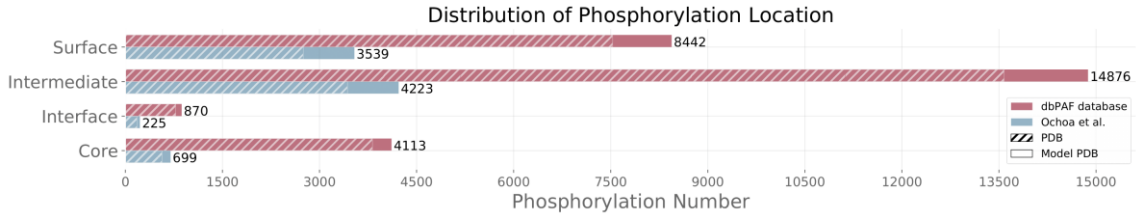



**Figure 6. Residue distribution of 3D phosphorylation groups (Surface, Intermediate, Core and Interface) and downstream subgrouping (static core and dynamic core)**

The numbers of residues are presented with dbPAF/Ochoa et al. order.

Interestingly, 91% of core sites did not pass quality control criteria of Ochoa *et al.* whereas, 72% of surface sites were able to pass the quality control (Ochoa et al., 2020) **(Figure 7A&7B)**.

A)




B)

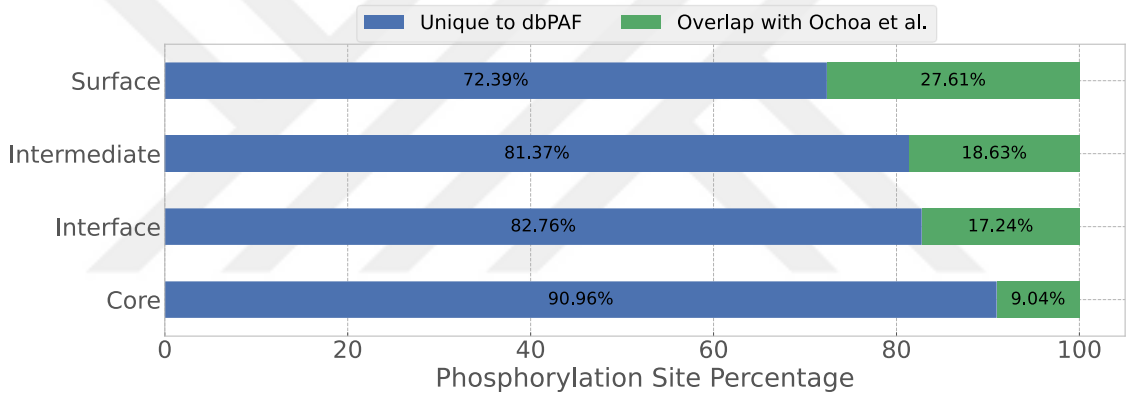



**Figure 7. Phosphorylation 3D structure distribution of all detected sites (A) and phosphorylation 3D structure percentage of overlapped sites (B)**

Distribution for Ochoa et al. and dbPAF database with separation of model PDB and PDB datasets are shown (A). Overlapped sites of dbPAF database with Ochoa *et al*. (B)

This difference is statistically significant based on the chi-square test (Chi square test | p-value<8.E-94). A phospho-site, that is located in the core region and labelled as static, is possibly inaccessible for kinase recognition. To check if these sites are potentially non-functional or biased to be located in proteins very large in size, we examined the length of the corresponding protein. We did not detect any significant bias about core phospho-sites to be located in large proteins. On the other hand, interface phospho-sites are located in relatively small-sized proteins (p-value<0.05, ANOVA and Tukey HSD, **Figure 8**) which is expected because the signaling pathways are composed of transient interactions which have relatively smaller interface size compared to permanent interactions (Perkins

et al., 2010). Additionally, the matching proteins are enriched in signaling pathways, namely MAPK signaling cascade, WNT signaling and NOTCH signaling pathways **(Figure 9)**.

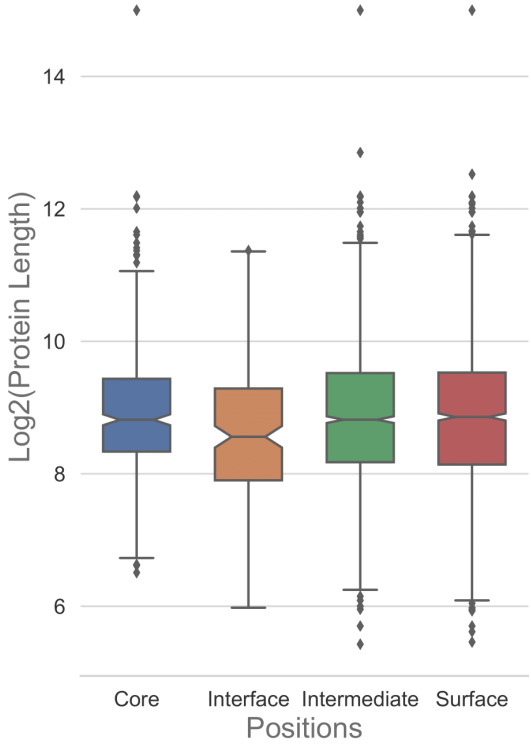



**Figure 8. The protein length comparison of phosphorylation 3D structure groups**

Amino acid lengths of the proteins which contains 3D phosphorylation groups are shown in log2 format.

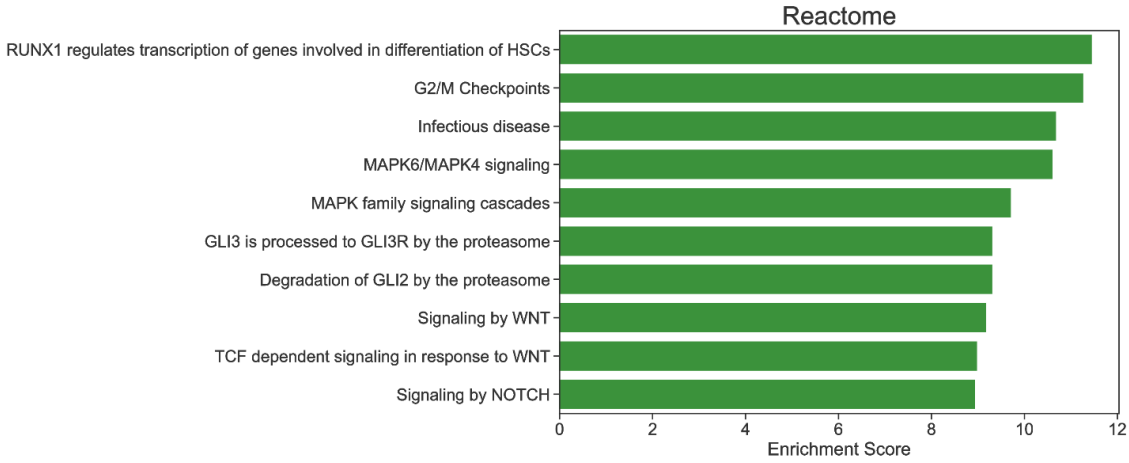



**Figure 9. Reactome enrichment analysis of interface sites**

Enrichment score represents -log10(p.value) and top 10 term with highest enrichment score is selected.

Stratification of phosphorylation sites based on their structural location provides us with a good quality dataset to further investigate distinct features of each 3D category. For this purpose, we examined the flexibility and mobility of the sites in each category by comparing their temperature factor (B-factor) distributions. B-factor represents the average of small motions of atoms within each protein. Higher B-factor implicitly indicates more flexible residues where B-factors greater than 50 indicate the atoms that are hard to detect during crystallization. The smaller B-factors values represent more rigid atoms. therefore, B-factor can be used for assessing the small motions of the residues. Since the scale of B-factor values is different across the protein structures, we calculated relative B-factor values within each protein structure to enable a fair comparison between multiple structures **(Figure 10)**. We observed that phospho-sites located in the surface region are significantly the more mobile, as expected (ANOVA and Tukey HSD | p-value=0.001). On the other hand, core and interface phosphorylation sites are in more rigid regions. In general, core and interface region characteristics are more similar to each other compared to the surface region. Therefore, relative B-factor values of phosphorylation sites are significantly differential based on their structural location.

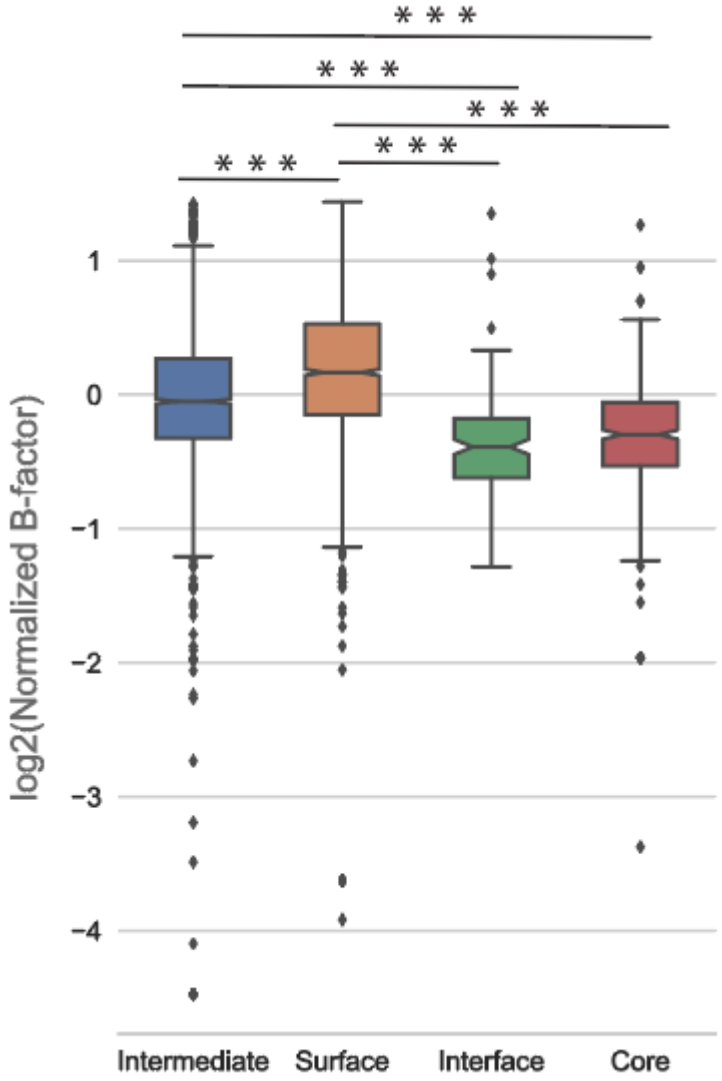



**Figure 10. The distribution of temperature factors (B-factor) for each phosphorylation types**

B-factor of each phosphorylation sites are normalized against average B-factor in each protein and presented in log2 scale (ANOVA and Tukey HSD | * ≤ 0.05, **≤ 0.01, ***≤ 0.001) (# of core: 472, interface:152, intermediate:2745, surface:2055).

The functionality of each phospho-site in Ochoa dataset (Ochoa et al., 2020) has been predicted with a learning-based model which incorporates fifty-nine features from four main categories: proteomic such as protein abundance; structural information such as solvent accessibility; evolution such as residue conservation, and regulatory information such as distinct motif features. We used their functionality score to assess the functional

relevance of the phospho-site in each 3D category and their statistical difference. A phospho-site having a functional score greater than 0.5 is counted as highly functional and some of them were experimentally validated (Ochoa et al., 2020). We analyzed the positional preference of highly functional phosphorylation sites **(Figure 11)**.

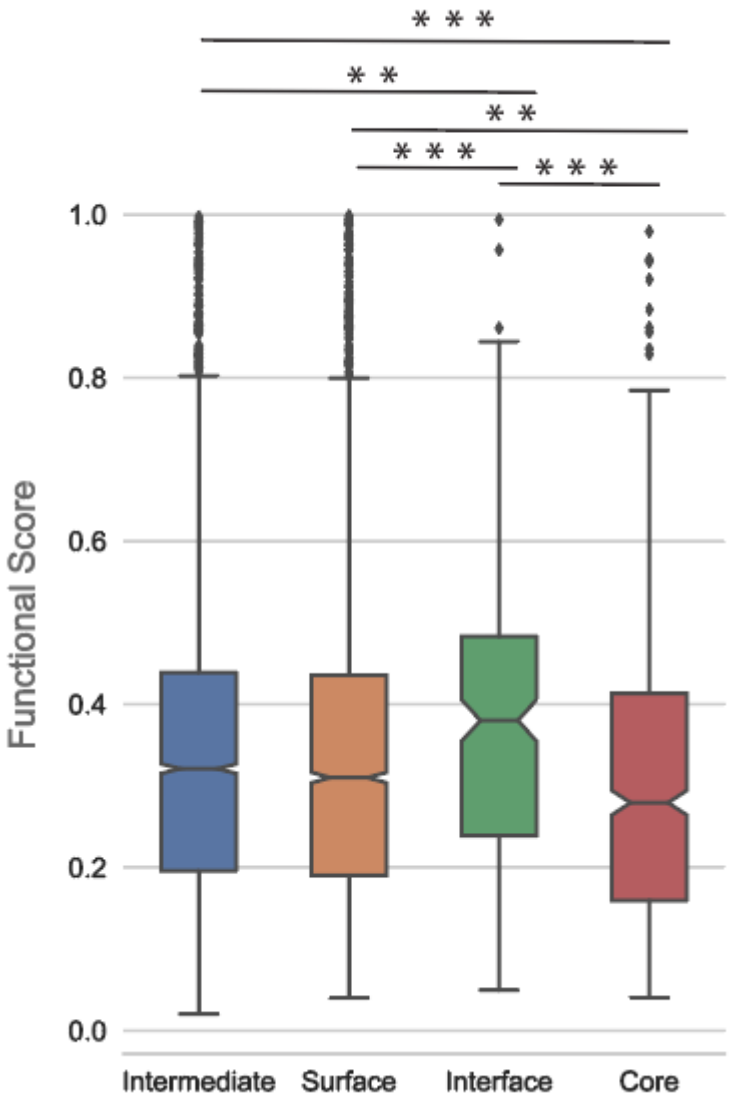



**Figure 11. Functional scores from Ochoa *et al.* for each phosphorylation groups**

ANOVA and Tukey HSD | * ≤ 0.05, **≤ 0.01, ***≤ 0.001 (# of core: 699, interface:225, intermediate:4223, surface:3539).

Interface phosphorylation sites are significantly more functional than other phosphorylation types (ANOVA and Tukey HSD | p-value=0.001). Whereas core

phosphorylation sites have significantly lower functionality (ANOVA and Tukey HSD | p-value=0.001) (Interface odds ratio: 1.428, core odds ratio 0.896) **(Table 2).**

| Odds Ratio | | | |
|---|---|---|---|
| Interface | Core | Intermediate | Surface |
| 1.428 | 0.796 | 0.998 | 1.026 |

**Table 2. Enrichment of functional sites for each 3D structural group with odds ratio.**

To test whether different 3D phosphorylation types have the differential function at different cellular compartments, we investigated their cellular location by using Human Protein Atlas (Uhlen et al., 2015). Enriched subcellular compartments are presented with their respective percentages **(Figure 12)**.

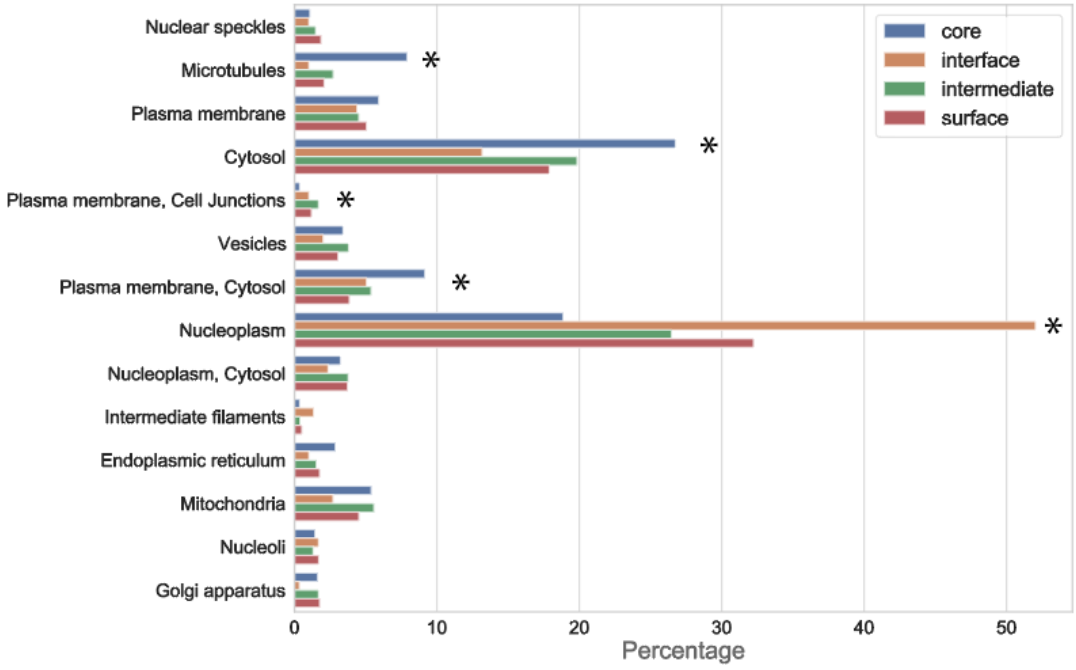



**Figure 12. Comparison of cellular compartmentalization of proteins from each 3D phosphorylation groups with their respective frequencies**

Significantly enriched places are illustrated with * (Chi square test | p-value<1.E-13).

Core phosphorylation sites are significantly enriched in microtubules (Chi square test | p-value<1.E-13) and interface sites are significantly enriched in the nucleoplasm in comparison to other 3D phosphorylation groups (Chi square test | p-value<6.8.E-28). Next, we calculated a score that represents how a residue type is frequently present or depleted in a specific region. More negative values imply the depletion of that residue in

the region and the more positive values imply more tendency to be located in that region **(Figure 13)**.

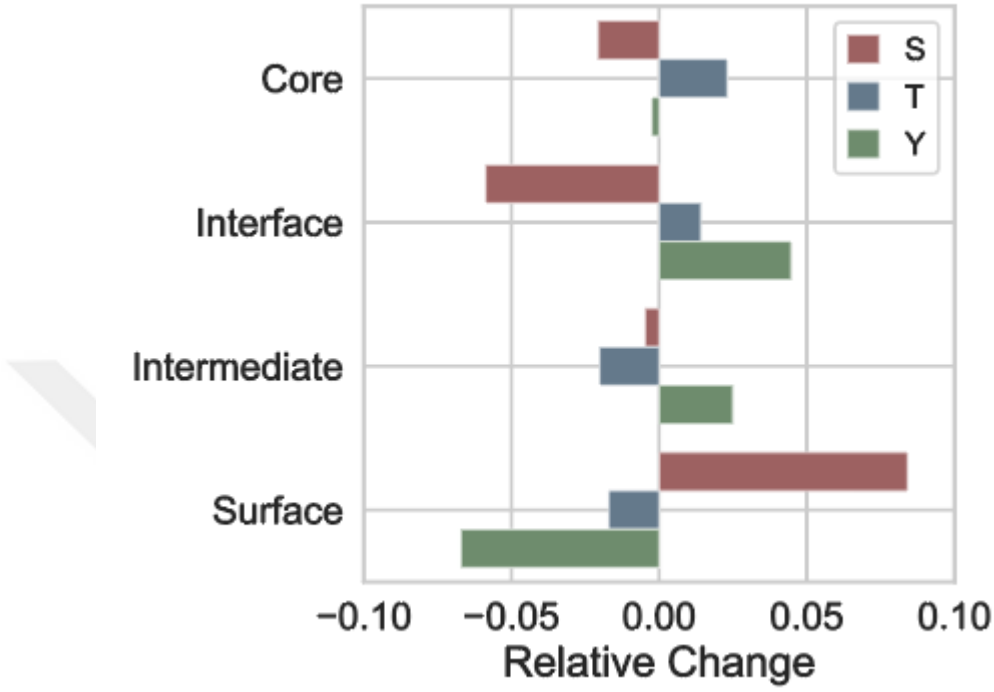



**Figure 13. Amino acid type tendency of phosphorylation residues from each group** (Chi square test | p-value<6.E-31).

All three-types of phospho-sites, tyrosine, serine, and threonine are distributed across different regions in the protein structure. However, some residue types tend to be located in specific regions. For example, tyrosine residues are rare in general and regulate multiple signalling pathways (Ardito et al., 2017). We found that tyrosine phosphorylation is significantly more frequent in the interface. Serine has the highest frequency in the phosphoproteome (Nishi et al., 2014, Schwartz and Murray, 2011) and these sites are notably enriched in the surface. Threonine phosphorylation is more likely to be in the core region (Chi square test | p-value<6.E-31).

Next, we compared the impact of mutations on those phosphorylation types by using EVmutation (Hopf et al., 2017) which is a statistical method that predicts the damaging effect of mutations. EVmutation relies on an unsupervised approach to predict the effect of mutations in a protein. In the prediction, it considers position dependencies of residues

and calculates the epistatic score representing the total statistical energy difference of mutated and wild type protein sequences. The more negative score represents the more damaging mutation. Dincer et *al*. (Dincer et al., 2019) have previously shown that the mutations in the interface and core regions have more damaging impact than the surface. In line with that, the same applies to phospho-sites as well. EVmutation scoring reveal that core and interface phosphorylation have significantly more damaging mutations based on the lowest EV mutation scores (ANOVA and Tukey HSD | p-value=0.001) **(Figure 14)**.

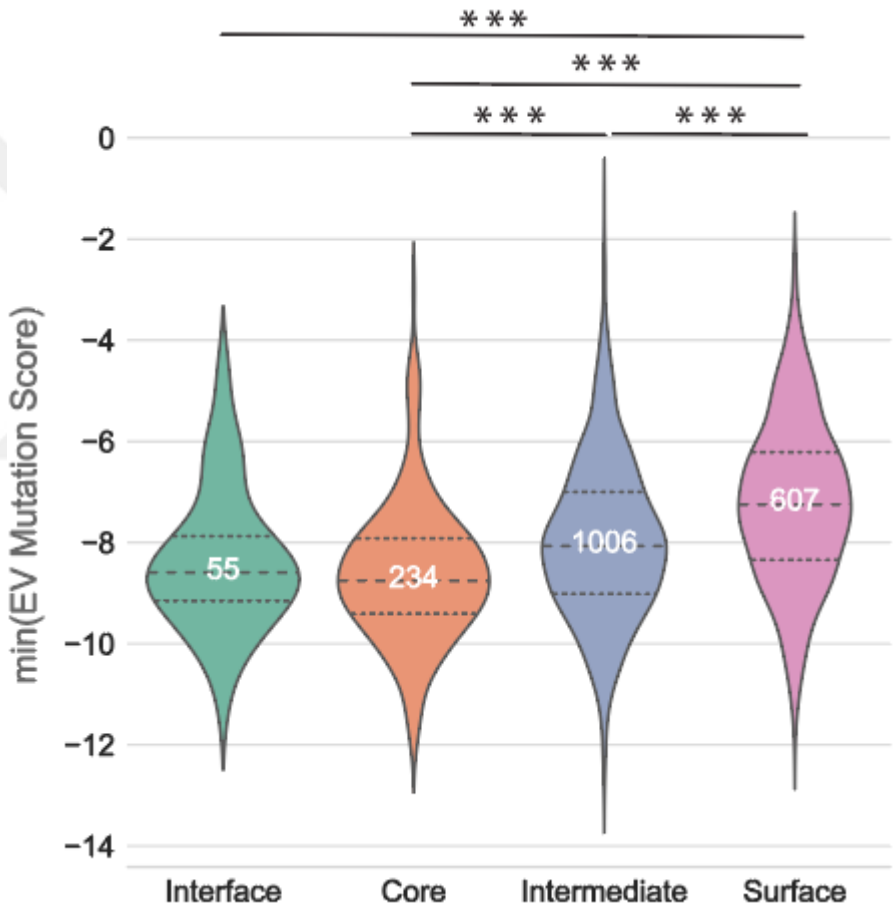



**Figure 14. Distribution of minimum mutation score (EVscore) across 3D phosphorylation groups**

ANOVA and Tukey HSD | * ≤ 0.05, **≤ 0.01, ***≤ 0.001.

Mutation in core residues of a protein may cause protein instabilities and structural changes stochastically which make them more damaging than other residues. We detect an increase in mutational effect while phosphorylation site accessibility decreases. To

further examine mutational effect differences, we compare phospho-sites with non phospho-sites in similar structural regions. We observed that phospho-sites in interface and core sites are more damaging than non phospho-sites within these regions. Whereas in surface and intermediate, phospho-sites did not have more damaging effect. Thus, in core and interface regions any damage in phospho-sites results in less stable proteins than a damage in non phospho-sites **(Figure 15)**.

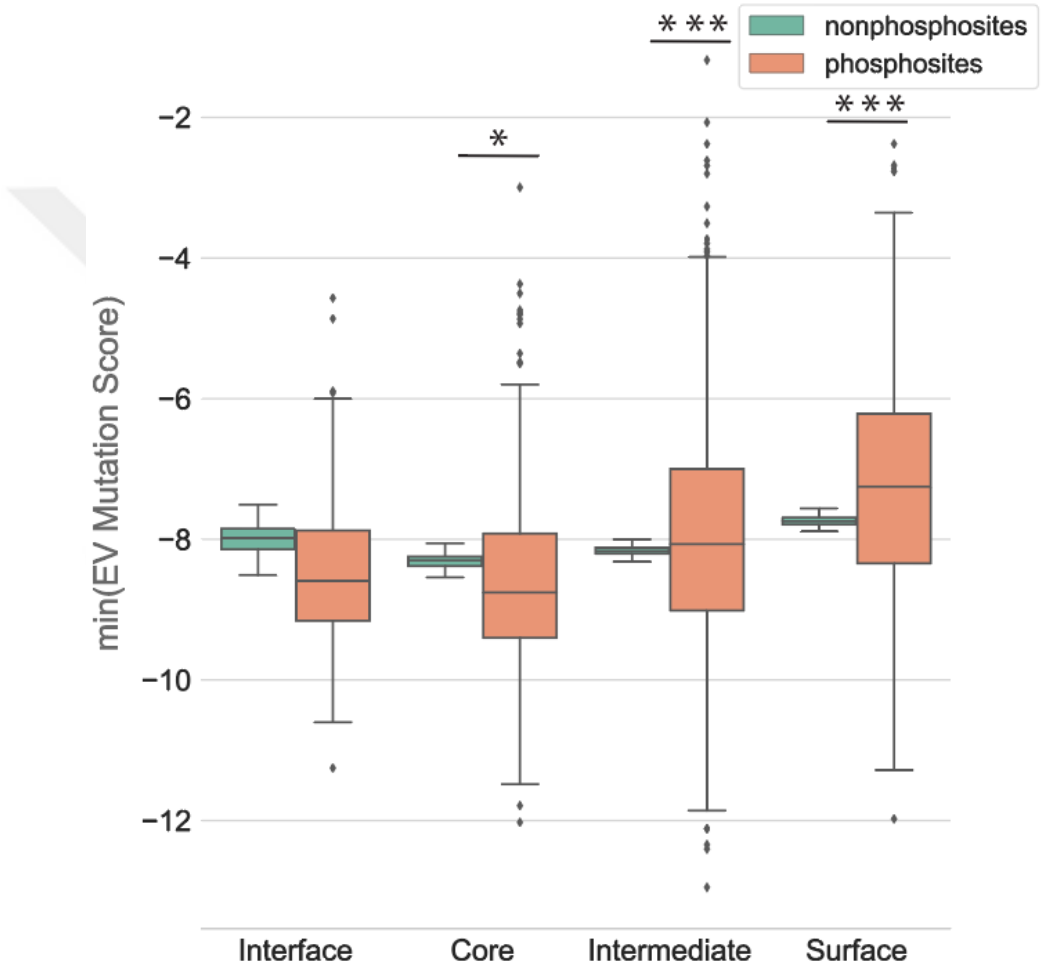



**Figure 15. Distribution of minimum mutation score (EVscore) across 3D phosphorylation groups (orange) and mean 3D nonphospho-sites groups (green)**

Distribution of mean 3D non phospho-sites is acquired by the random selection of non phospho-sites 100 times (Welch's t-test | * ≤ 0.05, **≤ 0.01, ***≤ 0.001).

As a result, we found that phosphorylation sites have different characteristics when structural information is considered. In general, high-throughput phosphoproteomic datasets provide information about which residues are phosphorylated. By integrating

protein structure information, we are able to investigate how these residues are organized and their regional and functional impact.

## 4.2 Core phosphorylation sites possess two distinct groups based on their dynamicity

Core regions of proteins are highly packed and rigid. Having a mutation in the core region is less frequent however when this occurs it has a dramatic impact on protein stability. The same rules are also valid for phospho-sites in the core region. Additionally, core phosphorylation sites are found to be less functional. When we consider protein structure-motion-function paradigm, we further analyzed the core phospho-sites in terms of possible dynamic changes to comment about their functionalities. Based on the hypothesis that core sites can only be phosphorylated after a conformational change, we searched for NMR models and multiple X-ray structures of each protein having at least one core phospho-site. In **Figure 16**, we compared the RSA values of 32 core phosphorylation sites that were detected in NMR models. Among those, 7 of them that map to a very small portion of the whole protein were filtered out. Out of 25 phosphorylation sites, 11 core phosphorylation sites alter their location in at least one model which were labelled as dynamic core sites.

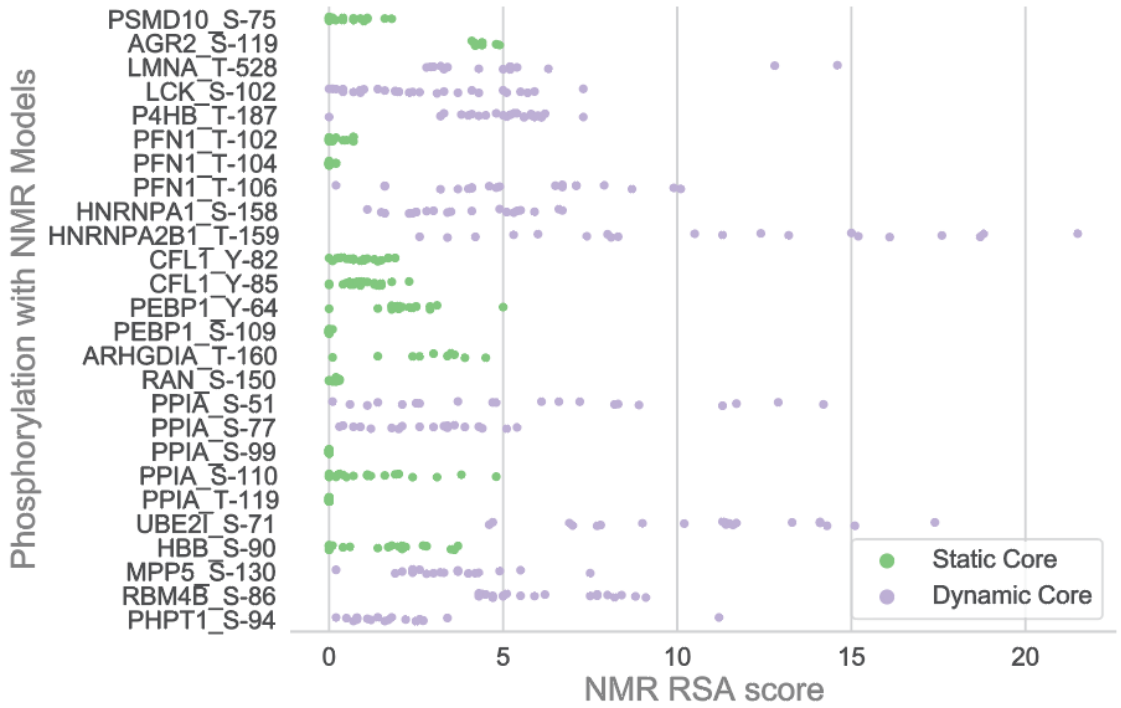



**Figure 16. Classification of core sites as dynamic or static based on their NMR models**

Phospho-sites are catalogued as dynamic if a residue changes its 3D phosphorylation group from core to any.

NMR results suggest that a certain portion of core phosphorylation sites display dynamicity which probably allows them to be accessed by kinases and phosphatases. However, the limited number of NMR structures in PDB restricts a more comprehensive dynamicity analysis. An additional layer of information is present as different conformations of a protein in multiple X-ray crystallography data. We next used proteins having at least three different conformations in PDB and checked if the corresponding core phospho-site changes its status from core to intermediate or surface across the conformations. In total 321 phosphorylation sites were obtained, 149 of which were dynamic as evident by their maximum RSA score change **(Figure 17)**.

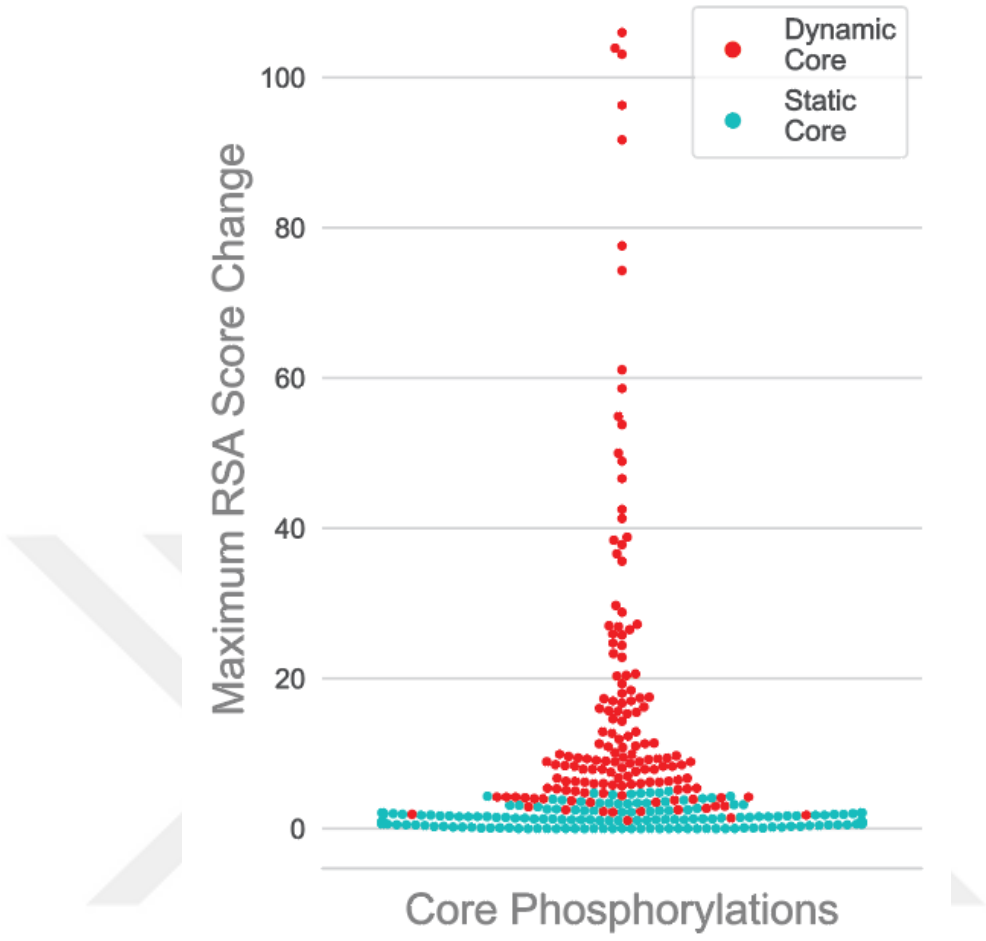



**Figure 17. Classification of core sites as dynamic or static based on their multiple X-ray structures**

Maximum RSA score differences are plotted for dynamic core (red) and static core (blue) phosphorylation sites.

Based on this analysis, we found out that the static and dynamic core phosphorylation sites are significantly different in terms of their RSA in the optimal structure **(Figure 18)**. Core phosphorylation sites which have higher RSA score are significantly more prone to be dynamic (Welch's t-test | p-value<3.32.E-16).

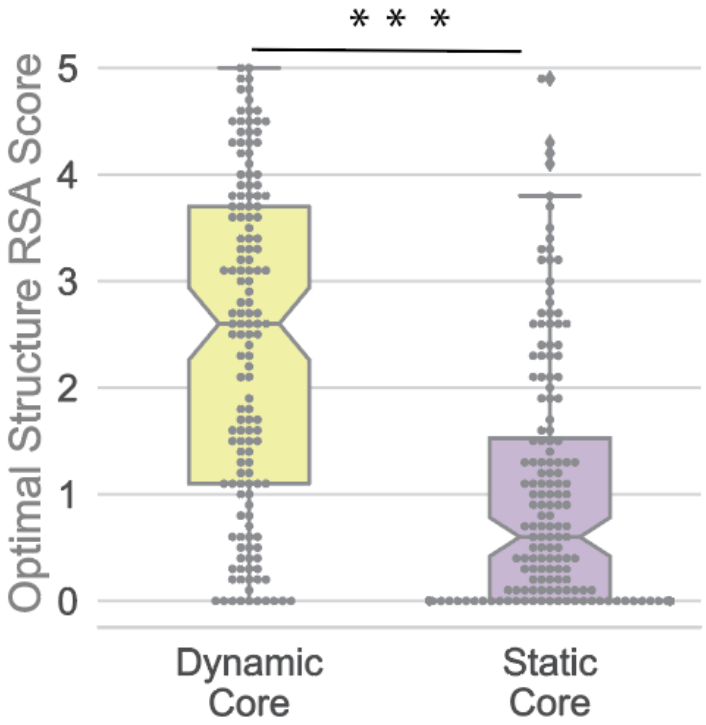



**Figure 18. Comparison of RSA scores between static (purple) and dynamic (yellow) core sites in selected optimal structures**

Optimal structures are detected as having the best coverage and resolution (Welch's t-test | p-value<3.32.E-16).

To test mobility differences between two core groups, we used normalized B-factor. However, due to multiple X-ray structures for the same protein, we continued with the average value of normalized B-factor of all structures **(Figure 19).**

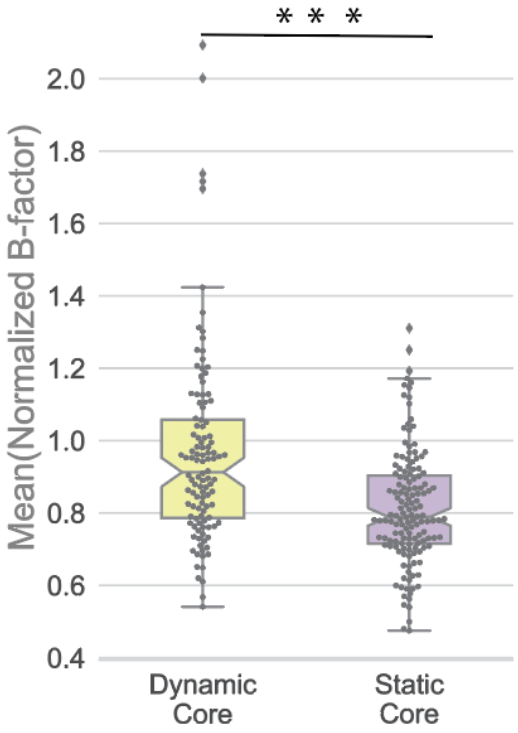



**Figure 19. Comparison of average normalized B-factors from different X-ray structures for each phosphorylation sites between dynamic and static groups**

(Welch's t-test | p-value<3.6.E-07)

Comparison of mean B-factor shows that dynamic cores have significantly higher atomic mobility (Welch's t-test | p-value<3.6.E-07). We tested whether the increase in B-factor is a natural consequence of RSA score rise, however, there is no correlation between B-factor and RSA score **(Figure 20)**.

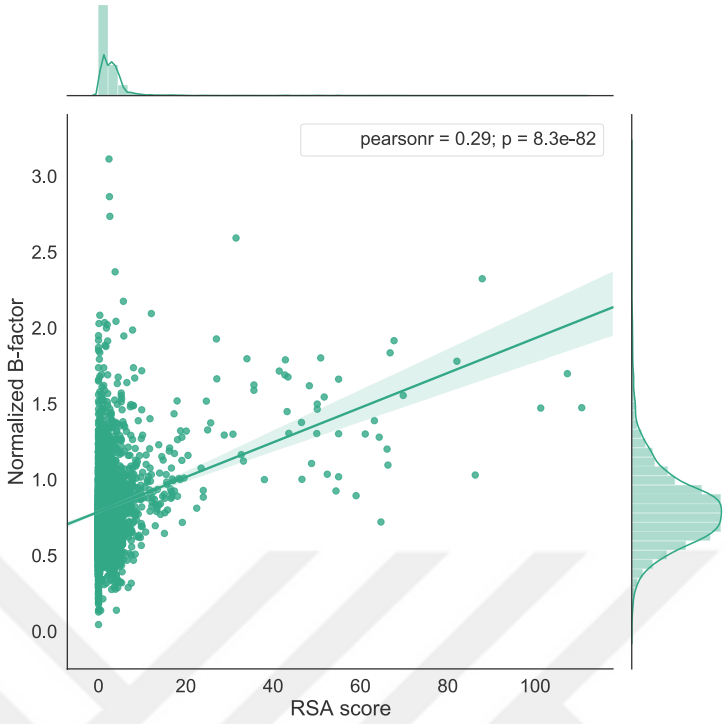



**Figure 20. The Pearson correlation between RSA score and normalized B-factor**

The physicochemical properties of the adjacent region around phospho-sites are similar to the intrinsically disordered regions in proteins (Iakoucheva et al., 2004). Therefore, we used NetSurfP-2.0 web tool to predict the disorder probability of each phospho-site (Klausen et al., 2019). NetSurfP uses a deep learning model trained on the structural features and sequence profiles to predict multiple structural properties including the disorder probability of each residue. Disorder score gives probability of the residue to be in an ordered region where 0 means high probability to be in ordered regions whereas 1 means disordered region. Even though the phospho-sites in both core groups have low probability of having disorder profile, dynamic core dataset have significantly more phospho-sites with high probability of disorder (Welch's t-test | p-value<0.00025) **(Figure 21)**.

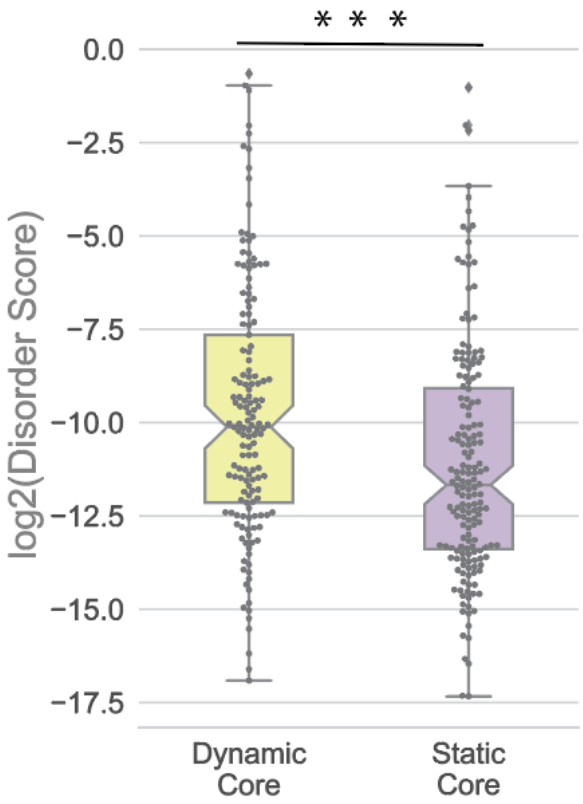



**Figure 21. Comparison of disorder scores in $\log_2$ scale between dynamic and static groups**

(Welch's t-test | p-value<0.00025)

We additionally examined the secondary structure tendency of core phosphorylation sites **(Figure 22)**. Dynamic core sites are enriched in the coil region whereas static core sites were more located in β-sheets.

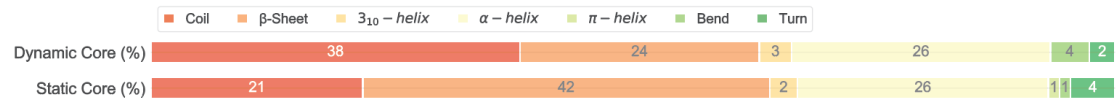



**Figure 22. Mapping phospho-sites from dynamic and static groups to the corresponding secondary structures**

Bar plot represents the percentage of each secondary structure indicated in the legend.

Finally, we asked whether these structurally well-defined two core phosphorylation subgroups have any difference in being functional or not, based on the scoring schema by Ochoa et al. (Ochoa et al., 2020). Importantly, dynamic core sites are significantly more functional than static cores (ANOVA and Tukey HSD | p-value=0.001). We compared all

3D phosphorylation types to test whether the difference between the functional score of core phosphorylation sites is negligible when compared to others (Figure 23).

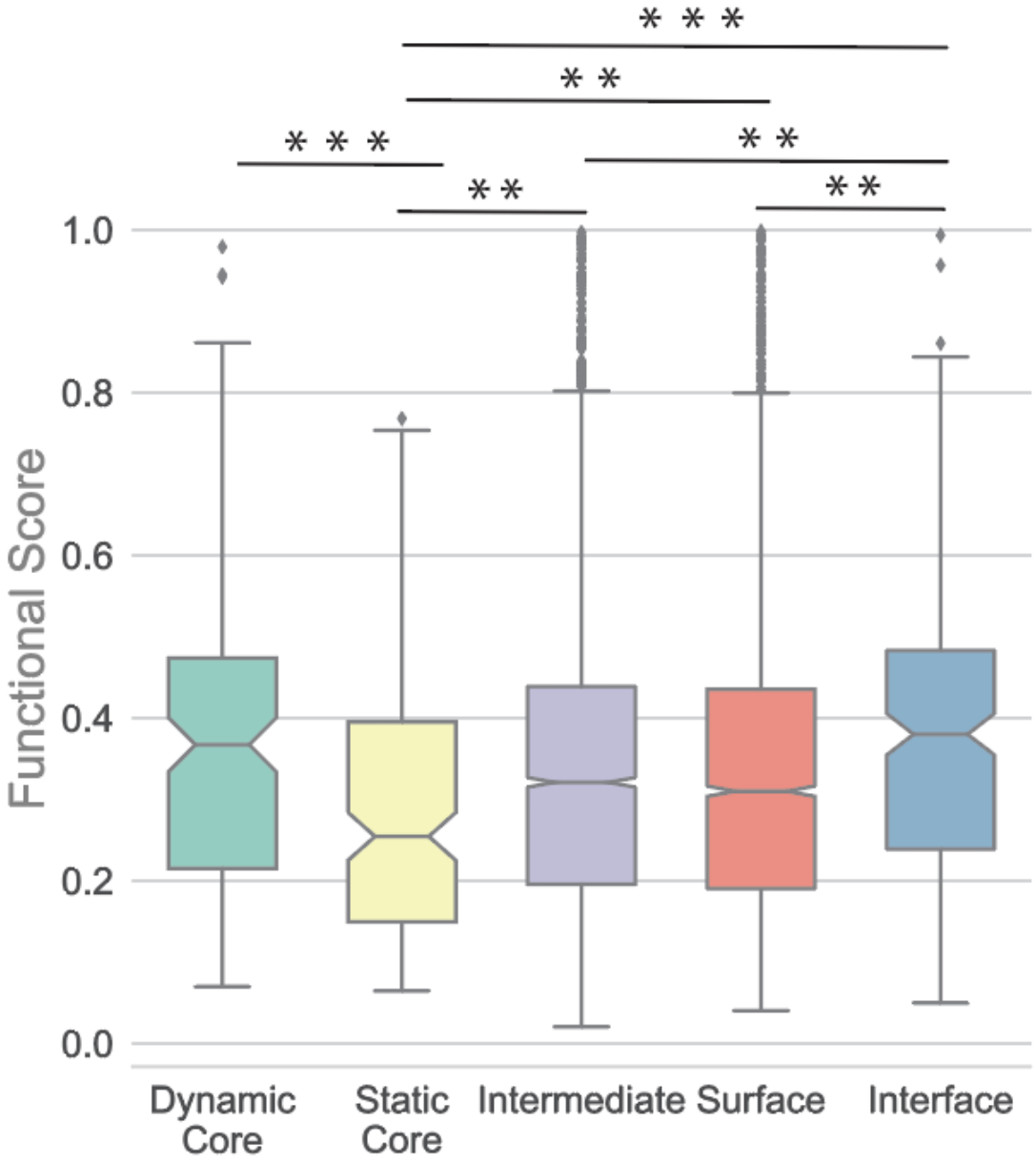



**Figure 23. Comparison of functional scores across all 3D phosphorylation groups including static and dynamic core sites**

ANOVA and Tukey HSD | * ≤ 0.05, **≤ 0.01, ***≤ 0.001

Static core phosphorylation sites are significantly less functional than all other types, whereas dynamic core phosphorylation sites are functional as much as interface sites. In line with this finding, we observed that dynamic core phosphorylation sites are

significantly more registered to PhosphoSitePlus (PSP) (Hornbeck et al., 2015) database than static ones (Figure 24).

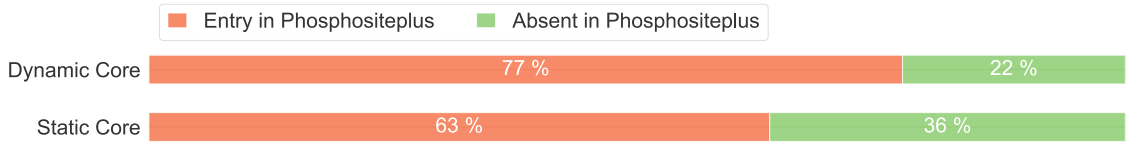



**Figure 24. Percentage of core phosphorylation sites which is entered to Phosphositeplus database**

Dynamic core seems more functional than static. Since PhosphoSitePlus is a curated database which registers the functional phosphorylation sites with kinase and functional knowledge, the difference in functional score could be the main reason for the registration difference in PhosphoSitePlus. Therefore, not all identified phospho-sites are registered without preliminary control. To determine a certain family of kinases that specifically phosphorylate dynamic or static core phosphorylations we employed GPS 5.0, a kinase prediction tool (Wang et al., 2020). However, there was not any specific kinase family preference other than atypical kinase specificity to dynamic core sites (Figure 25).

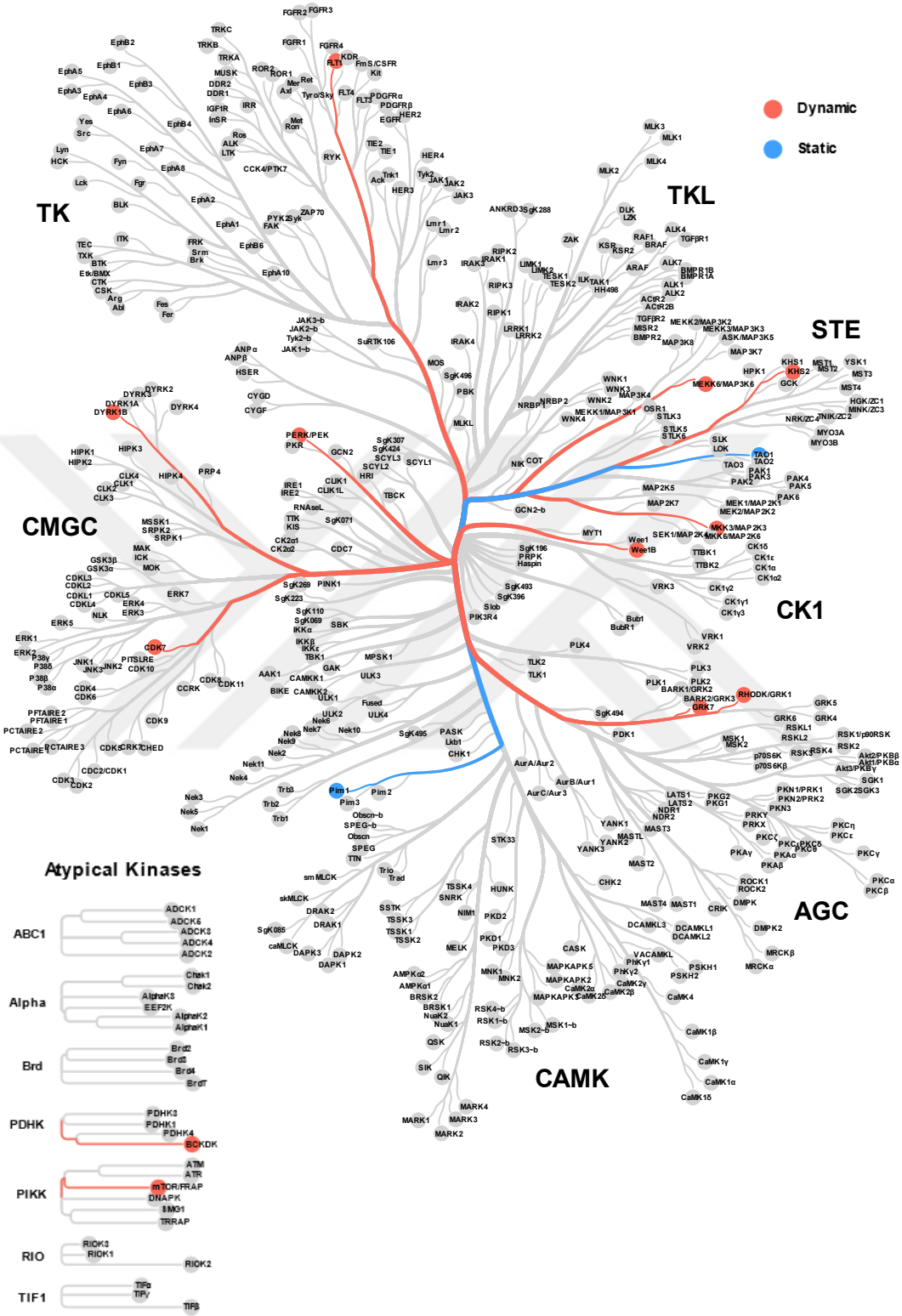



**Figure 25. Unique kinases that phosphorylate each core phosphorylation sites in kinome-tree**

Red: Kinases for dynamic core phosphorylation sites, Blue: Kinases for static core sites

## 4.3 Dynamic core sites are regulated during cell cycle

Our results revealed that core phosphorylation sites can be divided into two main groups according to the change in their location from core to intermediate or surface region. Further analysis shows that dynamic core phosphorylation sites are more functional than static core sites, and mostly on disordered regions.

The cell cycle is tightly regulated via protein phosphorylation events. Our recent study (Karayel et al., 2018) identified phosphorylation abundance alteration as cells progress into interphase, mitosis, and cytokinesis. By using these data, we investigated the 3D structure of the identified cell cycle regulated phospho-sites to tackle the functional relevance of 3D phosphorylation groups. Grouping of phosphorylation sites based on their 3D position revealed that all five detected interface phosphorylation sites are cell cycle dependent whereas other 3D phosphorylation groups have both cell cycle dependent and independent sites **(Figure 26)**.

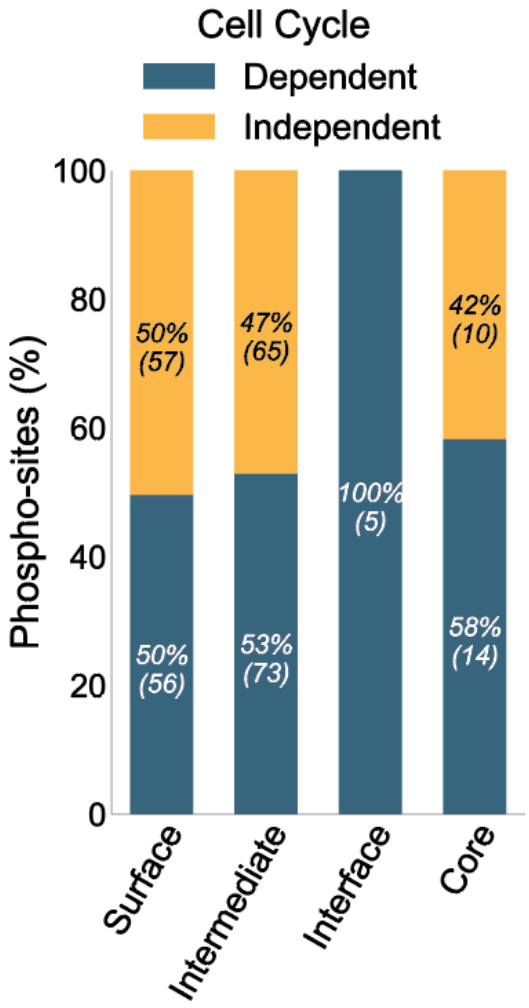



**Figure 26. Percentages of cell cycle dependent (blue) and independent (yellow) phosphorylation sites in 3D phosphorylation groups**

When we focused on core sites with multiple X-ray structures, we observed that cell cycle dependent core phosphorylation sites are more dynamic (based on the RSA distribution) than cell cycle independent phosphorylation sites which support our previous findings **(Figure 27)**.

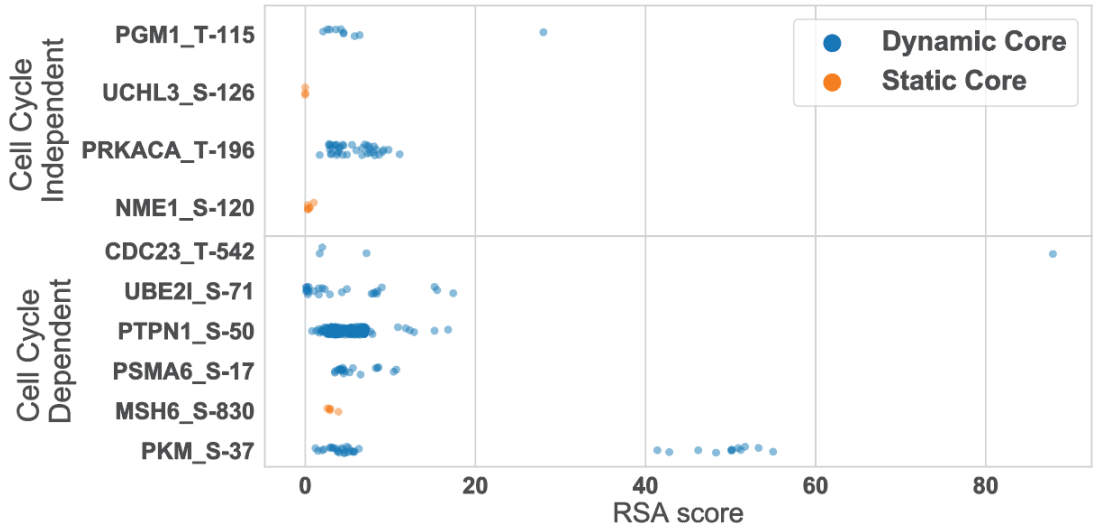



**Figure 27. Plotting of RSA scores of cell cycle dependent (bottom) and independent (top) phosphorylation sites from dynamic (blue) and static (orange) groups**

We closely examined the phosphopeptide abundance of selected 3D phosphorylation groups (interface, dynamic, and static core) during interphase, mitosis, and cytokinesis **(Figure 28)**. Abundance profiles suggest that dynamic core and interface phosphopeptides are more regulated as a cell enters mitosis whereas static core sites stay steadier which support the functional relevance of dynamic core and interface phosphorylation during the cell cycle.

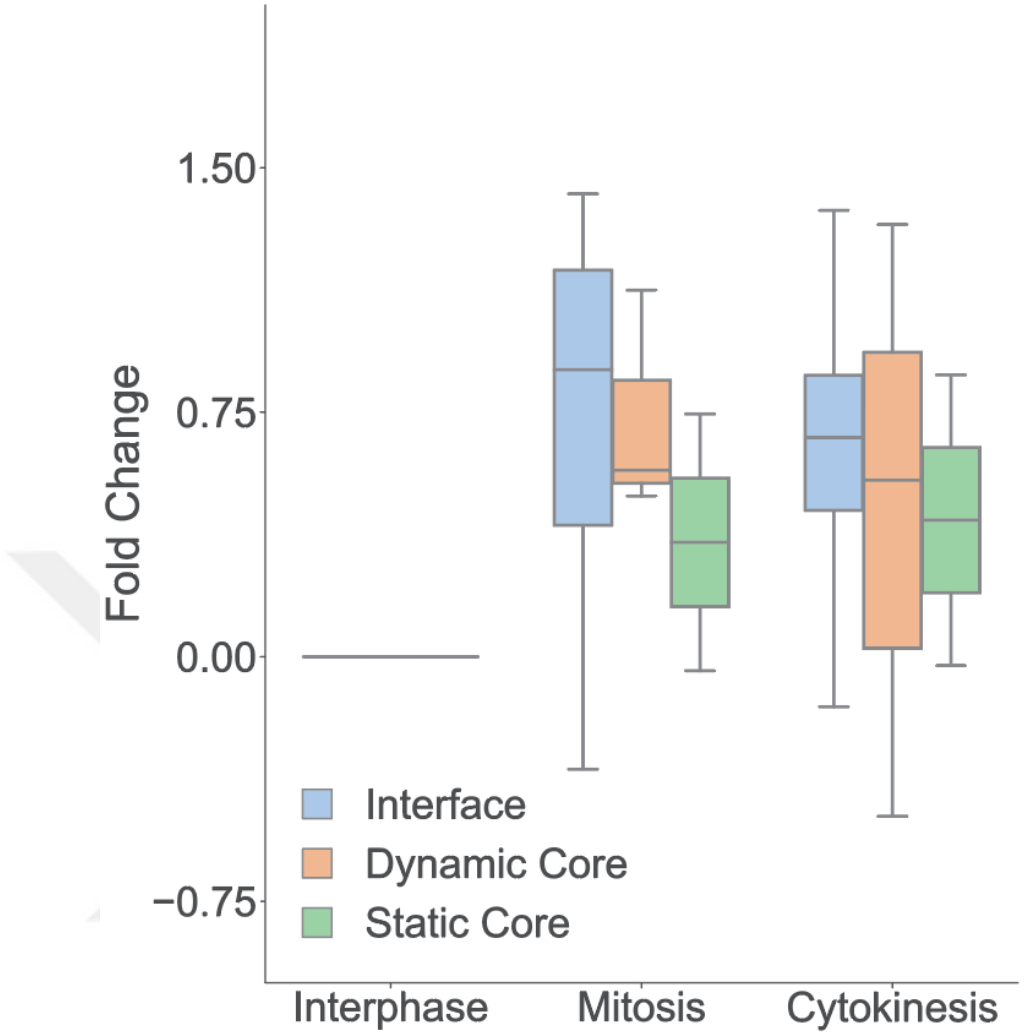



**Figure 28. Cell cycle dependent phosphorylation dynamics of interface (orange), dynamic (blue) and static (green) core groups based on their fold changes during interphase, mitosis, and cytokinesis**

### 4.4 Conformational change during phosphorylation is associated with functionality

We obtained multiple differential features of core phospho-sites in our characterization analysis. To rank these features according to their importance in their functionality, we leveraged a learning-based approach. We need to note that the training dataset is unbalanced, and small which prevented us from constructing a prediction model. Therefore, we completely focused on features which are essential for functionality. Our training set consists of positive labels from the functional phospho-sites which has a score greater than 0.5 according to Ochoa *et al.* and the rest has the negative labels. We trained

our model with different supervised learning algorithms so that we can determine the best model for feature importance. Random forest classifier performed the best among them. The ranked list of 25 features for functionality prediction is presented in descending order **(Figure 29)**. Maximum RSA change between X-ray crystallography results is the most important feature to detect functionality of core sites. Our categorization of dynamic and static core phosphorylation is the seventh. Our results imply that the more prone to conformational change are the more functional core phospho-sites.

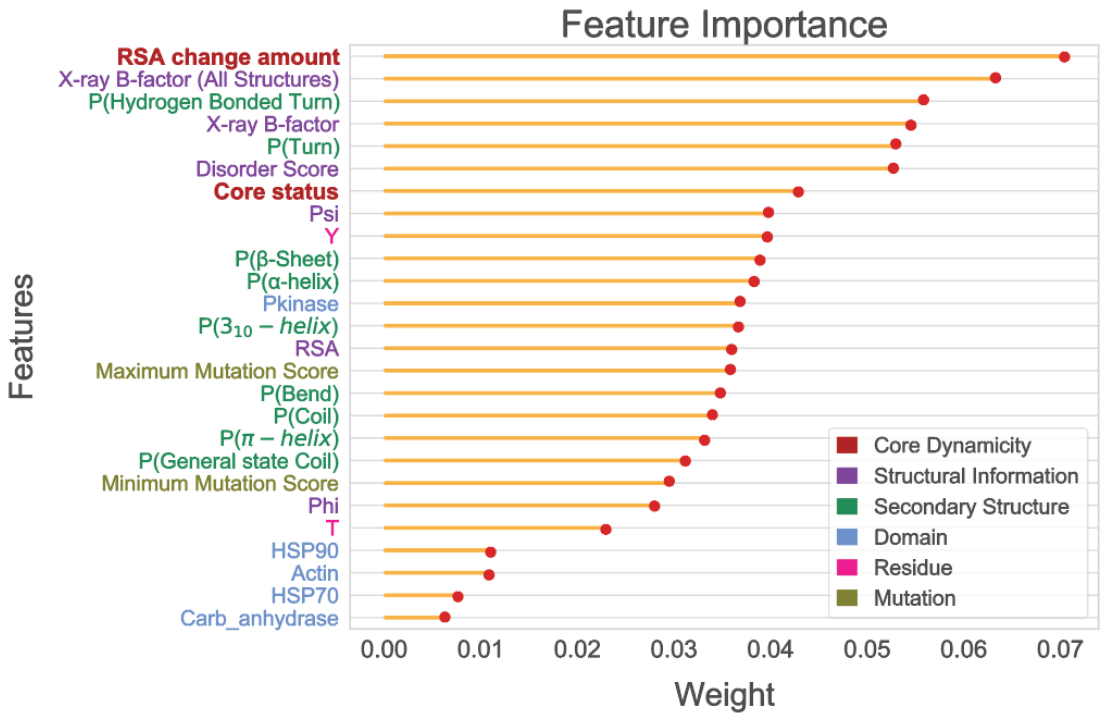



**Figure 29. Ranking of the features based on their weight in functionality prediction of core phosphorylation sites**

Features are categorized based on their general features; core dynamicity (red), structural information (purple), secondary structure (green), domain information (blue), residue (pink), and mutation (brown).

We present an example in **Figure 30** to illustrate the structural change in core phosphorylation sites, where two structures with major and minor location change are shown. We selected highly functional phosphorylation sites so that we can show the role of the dynamicity. We colored the secondary structure of proteins according to their B-factor values. Pyruvate kinase PKM opens its phosphorylation sites to the surface. Serine/threonine-protein kinase NEK2 undergoes small changes which bring the

phosphorylation site slightly closer to the surface. 3D structures clearly show that dynamic core sites undergo structural change and increase their accessibility.

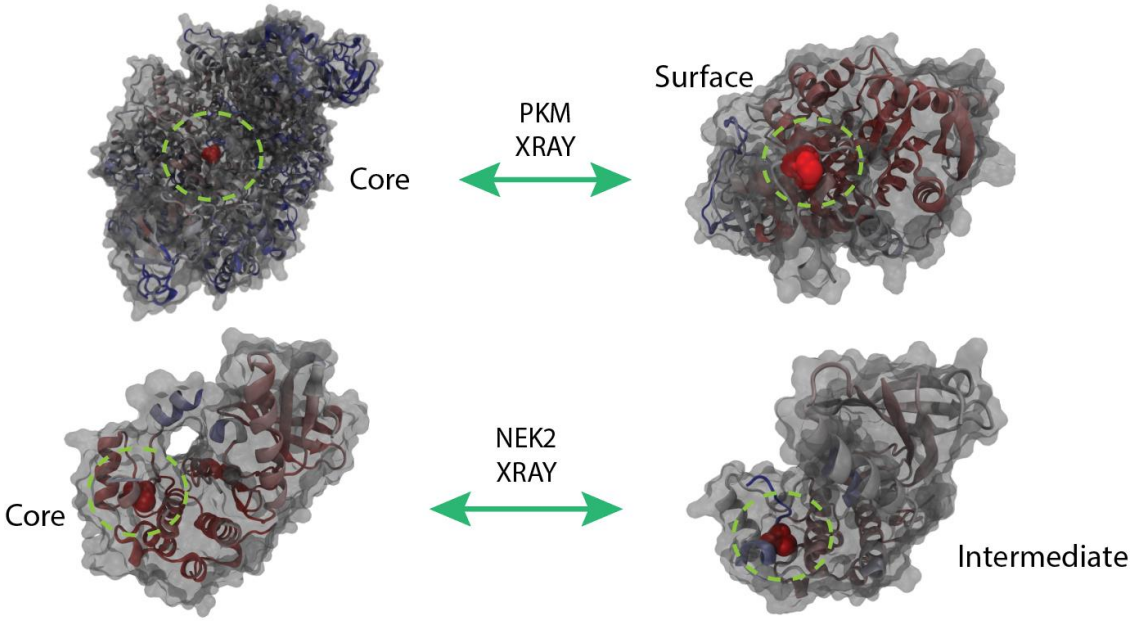



**Figure 30. 3D illustration of PKM (Phos-S37) and NEK2 (Phos-S184) in different conformations**

They are visualized by using VMD Software. Proteins are represented in surface. Phosphorylation sites are in van der Walls representation (red).

Protein phosphorylation often changes the conformation of the proteins. To examine phosphorylation dependent motion of a protein, we seek such an example having both phosphorylated and unphosphorylated structure in PDB. Insulin-like growth factor 1 receptor (IGF1R) has structures before and after phosphorylation at Y-1161 position which is a dynamic core phosphorylation site. Phosphorylated structure has been detected by electron microscopy whereas not phosphorylated structure from X-ray crystallography. Y-1161 residue translocated from core to surface in the phosphorylated form **(Figure 31)**. B-factor increase from 30.85 (normalized 1) to 55.42 (normalized 1.35) in tyrosine residue after phosphorylation which shows effect of phosphorylation of mobility. The phosphorylation mechanism needs to be further studied; however, the comparison of phosphorylated and non-phosphorylated state of IGF1R structures shows that conformation of the loop where the tyrosine is located changes and opens up in phosphorylated structure so that Y-1161 moves from core to the surface.

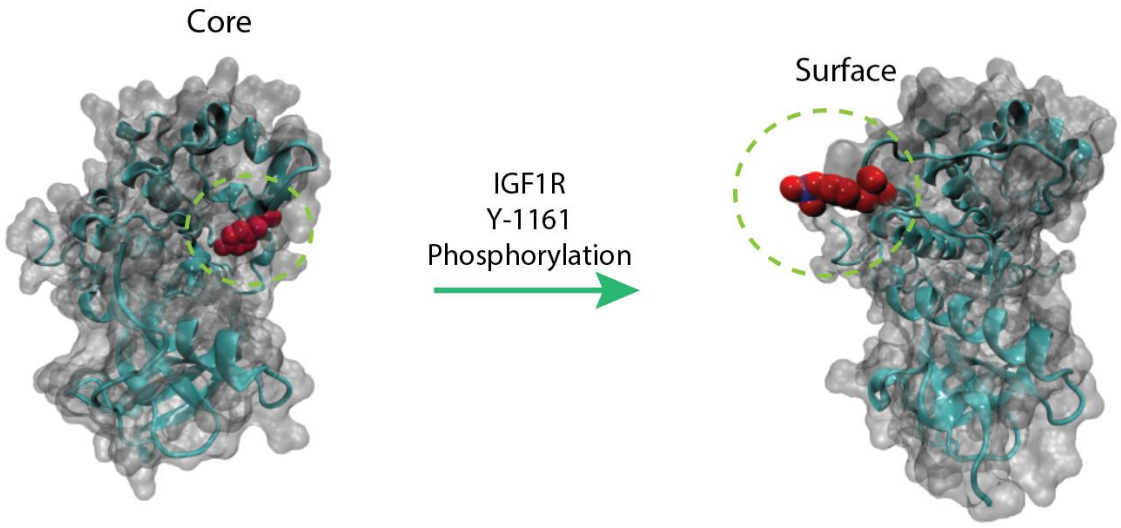



**Figure 31. 3D illustration of IGF1R (Phos-Y1161) before and after phosphorylation**

They are visualized by using VMD Software. Proteins are represented in surface. Phosphorylation sites are in van der Walls representation (red). Phosphor group is represented with blue.

# Chapter 5

# DISCUSSION

Proteins are dynamic molecules. Besides having multiple conformations, they may be phosphorylated in a context-dependent manner. Phosphorylation sites are prone to highly occurring in loop and disordered regions (Durek et al., 2009, Iakoucheva et al., 2004), and protein structures are altered after being phosphorylated (Solan and Lampe, 2009, Groban et al., 2006). High-throughput proteomic studies led us to know about at which sequence position a possible phosphorylation event occurs under the selected physiological condition. However, it does not provide information about at which conformation of the protein this event can occur. Therefore, there is a necessity to integrate more data with the proteomic studies to gain a better insight about phosphorylation events and to potentially filter out false positive hits. Given the structure-motion-function paradigm, one phospho-site may be located in the core region of the protein in its one conformation while it may be exposed to the solvent in another conformation which can form a reaction surface. In this study, we approached this problem to detect the state-dependency of phosphorylation events by integrating protein structures with proteomic studies to understand better the functionality of the phospho-sites and possibly detect the false positives in the set.

Many MS-based techniques such as cross-linking-mass spectrometry (XL-MS), limited proteolysis (LiP-MS) and hydrogen-deuterium exchange (HDX)-MS proved that MS can effectively provide information about protein structure (Iacobucci et al., 2018, Schopper et al., 2017, Yan and Maier, 2009) and push the MS's capacity beyond the determination of protein abundance. To determine protein functional alterations and regulatory events protein structural data has been shown to be complementary to protein abundance (Cappelletti et al., 2020). Similarly, we suggest that the functional relevance of a phosphorylation site can be determined by combining alteration at the protein structure and abundance level. In recent years there are remarkable efforts towards integration of

protein structure and phosphorylation events (Ramasamy et al., 2020, Ochoa et al., 2020, Tyanova et al., 2013).

We previously demonstrated that stratification of mutations observed in glioblastoma multiforme tumours as 3D patches reduces inter-patient heterogeneity and provides distinct druggable 3D mutation signatures (Dincer et al., 2019). Grouping phosphorylation sites based on their 3D position would provide distinct functional signatures. To this end, we divided phospho-sites based on their location as core, intermediate, surface and interface. To minimize the noise caused by the false positive phospho-sites we used quality-controlled Ochoa et al. dataset. However, we cannot rule out the noise caused by false phospho-sites or localization information. Interestingly, we found that passing the quality control stage of the method developed by Ochoa *et al.* is very restricted for the core phosphorylation sites. Only 9% of the core phospho-sites in the intersection of dbPAF and Ochoa dataset are in the high-quality set which implies the depletion in being functional and the power of solvent accessibility measures in quality check. However, static core sites in that portion are still as good quality as the rest of the 3D groups when we compared their peptide and localization scores **(Figure 32)**. We note that the number of known structures is limited. As the data accumulates, we will have a better understanding in evaluating these static core sites. As known, protein core regions are highly packed, conserved and rigid compared to the rest, and any mutation coincides with the core region may have a dramatic impact in the protein stability. The same rules are also followed by phospho-sites in the core region.

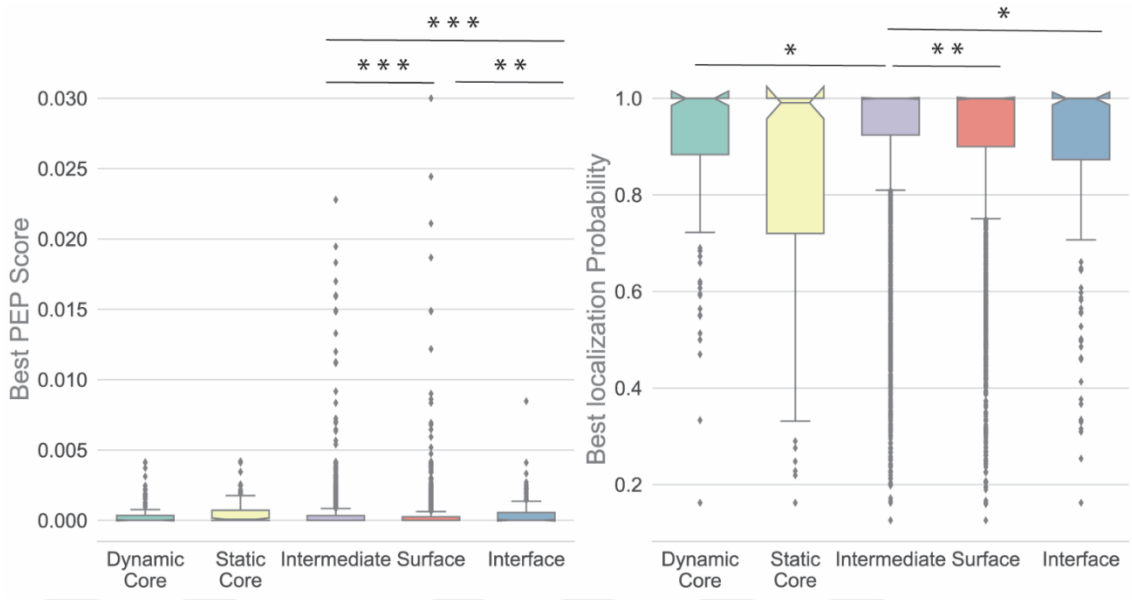



**Figure 32. Distribution of posterior error probability (PEP) score and best localization score of all 3D phosphorylation groups**

Kruskal-Wallis | Dunn's test | * ≤ 0.05, **≤ 0.01, ***≤ 0.001

Structural stratification of human phosphorylation sites revealed their signature properties such as their functionality, amino acid type tendency. Interface phospho-sites were the highly significant type of phosphorylation, probably due to their location in protein interaction sites. We observed that core phosphorylation sites are less functional and mobile than other types. However, among those, dynamic core phosphorylation sites are highly functional like other types of phosphorylation sites. Dynamic core phosphorylation sites are more accumulated in disordered regions and loops in at least one pose of the protein where it increases the possibility that dynamic core phosphorylation sites are highly functional phosphorylation sites. Our previous analysis of cell-cycle dependent phosphorylation analysis supported the functionality of dynamic core sites in comparison with static ones (Karayel et al., 2018). Although our dataset is small to draw a definite conclusion, based on our limited data we observed that the cell cycle dependent core sites are structurally more dynamic and static core sites stay steadier as a cell progresses into mitosis **(Figure 4)**. Discovery of two types of core phosphorylation sites might lead further research for the role of core phosphorylation sites in cellular mechanisms. Existence of static core phosphorylation sites could lead the creation of extra filtration steps which reckon with phosphorylation site structure for phosphoproteomic research.

To have an intuition which feature has the highest weight in describing the functionality of core phospho-sites, we used a learning-based approach. The difference between RSA of two conformations of the same protein is ranked as the most important feature that can guide about the functionality of the core phospho-sites. Because of the unbalanced and small size of the dataset, it is daunting to construct a prediction model. This type of learning-based approaches can be better trained and used for functionality prediction if the training set could be enlarged. This can be achieved by incorporating newly accumulated structural data in the future or integrating some other data resources such as the results of available molecular dynamic simulations.

The main limitation of our study is the number of known protein structures and obtaining multiple conformations. As the number of structures in PDB increases the success of this method will also improve. Additionally, molecular dynamic simulations of proteins can be utilized in the future to overcome the limited number of multiple conformations of proteins.

In conclusion, we demonstrate that stratification of phospho-sites based on their 3D position provides different groups with distinct features including functional relevance. A key challenge in the phosphoproteome field is to predict functionality of detected phosphorylation sites. This task can be tackled by making use of prediction of phosphorylation behavior by also using their 3D position information through machine learning methods.

# Bibliography

ADCOCK, S. A. & MCCAMMON, J. A. 2006. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem Rev,* 106**,** 1589-615.

ALQURAISHI, M. 2019. AlphaFold at CASP13. *Bioinformatics,* 35**,** 4862-4865.

ARDITO, F., GIULIANI, M., PERRONE, D., TROIANO, G. & LO MUZIO, L. 2017. The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy. *International Journal of Molecular Medicine,* 40**,** 271-280.

BHULLAR, K. S., LAGARON, N. O., MCGOWAN, E. M., PARMAR, I., JHA, A., HUBBARD, B. P. & RUPASINGHE, H. P. V. 2018. Kinase-targeted cancer therapies: progress, challenges and future directions. *Mol Cancer,* 17**,** 48.

BIRCK, C., MOUREY, L., GOUET, P., FABRY, B., SCHUMACHER, J., ROUSSEAU, P., KAHN, D. & SAMAMA, J. P. 1999. Conformational changes induced by phosphorylation of the FixJ receiver domain. *Structure,* 7**,** 1505-15.

BLAGUS, R. & LUSA, L. 2013. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics,* 14**,** 106.

BODENMILLER, B., WANKA, S., KRAFT, C., URBAN, J., CAMPBELL, D., PEDRIOLI, P. G., GERRITS, B., PICOTTI, P., LAM, H., VITEK, O., BRUSNIAK, M. Y., ROSCHITZKI, B., ZHANG, C., SHOKAT, K. M., SCHLAPBACH, R., COLMAN-LERNER, A., NOLAN, G. P., NESVIZHSKII, A. I., PETER, M., LOEWITH, R., VON MERING, C. & AEBERSOLD, R. 2010. Phosphoproteomic analysis reveals interconnected system-wide responses to perturbations of kinases and phosphatases in yeast. *Sci Signal,* 3**,** rs4.

BORK, P. 1991. Shuffled domains in extracellular proteins. *FEBS Lett,* 286**,** 47-54.

BOTHWELL, J. H. F. & GRIFFIN, J. L. 2011. An introduction to biological nuclear magnetic resonance spectroscopy. *Biological Reviews,* 86**,** 493-510.

CAPPELLETTI, V., HAUSER, T., PIAZZA, I., PEPELNJAK, M., MALINOVSKA, L., FUHRER, T., LI, Y., DÖRIG, C., BOERSEMA, P., GILLET, L., GROSSBACH, J., DUGOURD, A., SAEZ-RODRIGUEZ, J., BEYER, A., ZAMBONI, N., CAFLISCH, A., DE SOUZA, N. & PICOTTI, P. 2020. Dynamic 3D proteomes reveal protein functional alterations at high resolution in situ. *Cell*.

CARRONI, M. & SAIBIL, H. R. 2016. Cryo electron microscopy to determine the structure of macromolecular complexes. *Methods,* 95**,** 78-85.

COHEN, P. 2000. The regulation of protein function by multisite phosphorylation--a 25 year update. *Trends Biochem Sci,* 25**,** 596-601.

CONIBEAR, A. C. 2020. Deciphering protein post-translational modifications using chemical biology tools. *Nature Reviews Chemistry*.

COX, J. & MANN, M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol,* 26**,** 1367-72.

DE GRAAF, E. L., KAPLON, J., MOHAMMED, S., VEREIJKEN, L. A., DUARTE, D. P., REDONDO GALLEGO, L., HECK, A. J., PEEPER, D. S. & ALTELAAR, A. F. 2015. Signal Transduction Reaction Monitoring Deciphers Site-Specific PI3K-mTOR/MAPK Pathway Dynamics in Oncogene-Induced Senescence. *J Proteome Res,* 14**,** 2906-14.

DINCER, C., KAYA, T., KESKIN, O., GURSOY, A. & TUNCBAG, N. 2019. 3D spatial organization and network-guided comparison of mutation profiles in

Glioblastoma reveals similarities across patients. *PLoS Comput Biol,* 15**,** e1006789.

DOLL, S. & BURLINGAME, A. L. 2015. Mass spectrometry-based detection and assignment of protein posttranslational modifications. *ACS Chem Biol,* 10**,** 63-71.

DUREK, P., SCHUDOMA, C., WECKWERTH, W., SELBIG, J. & WALTHER, D. 2009. Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins. *BMC Bioinformatics,* 10**,** 117.

EDWARDS, A. M., ISSERLIN, R., BADER, G. D., FRYE, S. V., WILLSON, T. M. & YU, F. H. 2011. Too many roads not taken. *Nature,* 470**,** 163-165.

ELIUK, S. M., MALTBY, D., PANNING, B. & BURLINGAME, A. L. 2010. High resolution electron transfer dissociation studies of unfractionated intact histones from murine embryonic stem cells using on-line capillary LC separation: determination of abundant histone isoforms and post-translational modifications. *Mol Cell Proteomics,* 9**,** 824-37.

FICARRO, S. B., MCCLELAND, M. L., STUKENBERG, P. T., BURKE, D. J., ROSS, M. M., SHABANOWITZ, J., HUNT, D. F. & WHITE, F. M. 2002. Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. *Nat Biotechnol,* 20**,** 301-5.

FUHS, S. R. & HUNTER, T. 2017. pHisphorylation: the emergence of histidine phosphorylation as a reversible regulatory modification. *Curr Opin Cell Biol,* 45**,** 8-16.

GANESAN, K., KULANDAISAMY, A., PRIYA, S. B. & GROMIHA, M. M. 2019. HuVarBase: A human variant database with comprehensive information at gene and protein levels. Plos One, 14.

GLOTZER, M. 2005. The molecular requirements for cytokinesis. *Science,* 307**,** 1735-9.

GOVINDARAJAN, S., RECABARREN, R. & GOLDSTEIN, R. A. 1999. Estimating the total number of protein folds. *Proteins,* 35**,** 408-14.

GROBAN, E. S., NARAYANAN, A. & JACOBSON, M. P. 2006. Conformational changes in protein loops and helices induced by post-translational phosphorylation. *Plos Computational Biology,* 2**,** 238-250.

GRUHLER, A., OLSEN, J. V., MOHAMMED, S., MORTENSEN, P., FAERGEMAN, N. J., MANN, M. & JENSEN, O. N. 2005. Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol Cell Proteomics,* 4**,** 310-27.

HOPF, T. A., GREEN, A. G., SCHUBERT, B., MERSMANN, S., SCHARFE, C. P. I., INGRAHAM, J. B., TOTH-PETROCZY, A., BROCK, K., RIESSELMAN, A. J., PALMEDO, P., KANG, C., SHERIDAN, R., DRAIZEN, E. J., DALLAGO, C., SANDER, C. & MARKS, D. S. 2019. The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics,* 35**,** 1582-1584.

HOPF, T. A., INGRAHAM, J. B., POELWIJK, F. J., SCHARFE, C. P., SPRINGER, M., SANDER, C. & MARKS, D. S. 2017. Mutation effects predicted from sequence co-variation. *Nat Biotechnol,* 35**,** 128-135.

HORNBECK, P. V., ZHANG, B., MURRAY, B., KORNHAUSER, J. M., LATHAM, V. & SKRZYPEK, E. 2015. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res,* 43**,** D512-20.

HUANG, J., RAUSCHER, S., NAWROCKI, G., RAN, T., FEIG, M., DE GROOT, B. L., GRUBMULLER, H. & MACKERELL, A. D., JR. 2017. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods,* 14**,** 71-73.

HUGHES, C. A., MANDELL, J. G., ANAND, G. S., STOCK, A. M. & KOMIVES, E. A. 2001. Phosphorylation causes subtle changes in solvent accessibility at the interdomain interface of methylesterase CheB. *J Mol Biol,* 307**,** 967-76.

HUMPHREY, W., DALKE, A. & SCHULTEN, K. 1996. VMD: visual molecular dynamics. *J Mol Graph,* 14**,** 33-8, 27-8.

IACOBUCCI, C., GOTZE, M., IHLING, C. H., PIOTROWSKI, C., ARLT, C., SCHAFER, M., HAGE, C., SCHMIDT, R. & SINZ, A. 2018. A cross-linking/mass spectrometry workflow based on MS-cleavable cross-linkers and the MeroX software for studying protein structures and protein-protein interactions. *Nat Protoc,* 13**,** 2864-2889.

IAKOUCHEVA, L. M., RADIVOJAC, P., BROWN, C. J., O'CONNOR, T. R., SIKES, J. G., OBRADOVIC, Z. & DUNKER, A. K. 2004. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res,* 32**,** 1037-49.

JENSEN, O. N. 2006. Interpreting the protein language using proteomics. *Nat Rev Mol Cell Biol,* 7**,** 391-403.

JOHNSON, L. N. 2009. Protein kinase inhibitors: contributions from structure to clinical compounds. *Quarterly Reviews of Biophysics,* 42**,** 1-40.

JONES, D. T., SINGH, T., KOSCIOLEK, T. & TETCHNER, S. 2015. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics,* 31**,** 999-1006.

KARAYEL, O., SANAL, E., GIESE, S. H., URETMEN KAGIALI, Z. C., POLAT, A. N., HU, C. K., RENARD, B. Y., TUNCBAG, N. & OZLU, N. 2018. Comparative phosphoproteomic analysis reveals signaling networks regulating monopolar and bipolar cytokinesis. *Sci Rep,* 8**,** 2269.

KEMPF, J. G. & LORIA, J. P. 2003. Protein dynamics from solution NMR: theory and applications. *Cell Biochem Biophys,* 37**,** 187-211.

KLAUSEN, M. S., JESPERSEN, M. C., NIELSEN, H., JENSEN, K. K., JURTZ, V. I., SONDERBY, C. K., SOMMER, M. O. A., WINTHER, O., NIELSEN, M., PETERSEN, B. & MARCATILI, P. 2019. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins,* 87**,** 520-527.

KOLLAREDDY, M., DZUBAK, P., ZHELEVA, D. & HAJDUCH, M. 2008. Aurora Kinases: Structure, Functions and Their Association with Cancer. *Biomedical Papers-Olomouc,* 152**,** 27-33.

KUHLBRANDT, W. 2014. Biochemistry. The resolution revolution. *Science,* 343**,** 1443-4.

KUHLMAN, B. & BRADLEY, P. 2019. Advances in protein structure prediction and design. *Nat Rev Mol Cell Biol,* 20**,** 681-697.

LEE, B. & RICHARDS, F. M. 1971. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol,* 55**,** 379-400.

LI, F. Y., FAN, C. S., MARQUEZ-LAGO, T. T., LEIER, A., REVOTE, J., JIA, C. Z., ZHU, Y., SMITH, A. I., WEBB, G. I., LIU, Q. Z., WEI, L. Y., LI, J. & SONG, J. N. 2020. PRISMOID: a comprehensive 3D structure database for post-translational modifications and mutations with functional impact. *Briefings in Bioinformatics,* 21**,** 1069-1079.

LI, H., CHANG, Y. Y., YANG, L. W. & BAHAR, I. 2016. iGNM 2.0: the Gaussian network model database for biomolecular structural dynamics. *Nucleic Acids Res,* 44**,** D415-22.

LIM, S. & KALDIS, P. 2013. Cdks, cyclins and CKIs: roles beyond cell cycle regulation. *Development,* 140**,** 3079-93.

MANN, M., KULAK, N. A., NAGARAJ, N. & COX, J. 2013. The coming age of complete, accurate, and ubiquitous proteomes. *Mol Cell,* 49**,** 583-90.

MANNING, G., WHYTE, D. B., MARTINEZ, R., HUNTER, T. & SUDARSANAM, S. 2002. The protein kinase complement of the human genome. *Science,* 298**,** 1912-34.

METZ, K. S., DEOUDES, E. M., BERGINSKI, M. E., JIMENEZ-RUIZ, I., AKSOY, B. A., HAMMERBACHER, J., GOMEZ, S. M. & PHANSTIEL, D. H. 2018. Coral: Clear and Customizable Visualization of Human Kinome Data. *Cell Syst,* 7**,** 347-350 e1.

MITTERNACHT, S. 2016. FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Res,* 5**,** 189.

MOULT, J., PEDERSEN, J. T., JUDSON, R. & FIDELIS, K. 1995. A large-scale experiment to assess protein structure prediction methods. *Proteins,* 23**,** ii-v.

NABEL, E. G. 2002. CDKs and CKIs: molecular targets for tissue remodelling. *Nat Rev Drug Discov,* 1**,** 587-98.

NEEDHAM, E. J., PARKER, B. L., BURYKIN, T., JAMES, D. E. & HUMPHREY, S. J. 2019. Illuminating the dark phosphoproteome. *Sci Signal,* 12.

NEWMAN, J. 2006. A review of techniques for maximizing diffraction from a protein crystal in stilla. *Acta Crystallogr D Biol Crystallogr,* 62**,** 27-31.

NISHI, H., SHAYTAN, A. & PANCHENKO, A. R. 2014. Physicochemical mechanisms of protein regulation by phosphorylation. *Front Genet,* 5**,** 270.

OCHOA, D., JARNUCZAK, A. F., VIEITEZ, C., GEHRE, M., SOUCHERAY, M., MATEUS, A., KLEEFELDT, A. A., HILL, A., GARCIA-ALONSO, L., STEIN, F., KROGAN, N. J., SAVITSKI, M. M., SWANEY, D. L., VIZCAINO, J. A., NOH, K. M. & BELTRAO, P. 2020. The functional landscape of the human phosphoproteome. *Nat Biotechnol,* 38**,** 365-373.

OLSEN, J. V., BLAGOEV, B., GNAD, F., MACEK, B., KUMAR, C., MORTENSEN, P. & MANN, M. 2006. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell,* 127**,** 635-648.

PARDEE, A. B. 1989. G1 events and regulation of cell proliferation. *Science,* 246**,** 603-8.

PAULING, L., COREY, R. B. & BRANSON, H. R. 1951. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A,* 37**,** 205-11.

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, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research,* 12**,** 2825-2830.

PERKINS, J. R., DIBOUN, I., DESSAILLY, B. H., LEES, J. G. & ORENGO, C. 2010. Transient protein-protein interactions: structural, functional, and network properties. Structure, 18, 1233-43.

PETRONCZKI, M., GLOTZER, M., KRAUT, N. & PETERS, J. M. 2007. Polo-like kinase 1 triggers the initiation of cytokinesis in human cells by promoting recruitment of the RhoGEF Ect2 to the central spindle. *Dev Cell,* 12**,** 713-25.

PIEPER, U., WEBB, B. M., DONG, G. Q., SCHNEIDMAN-DUHOVNY, D., FAN, H., KIM, S. J., KHURI, N., SPILL, Y. G., WEINKAM, P., HAMMEL, M., TAINER, J. A., NILGES, M. & SALI, A. 2014. ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res,* 42**,** D336-46.

PINKSE, M. W. H., UITTO, P. M., HILHORST, M. J., OOMS, B. & HECK, A. J. R. 2004. Selective isolation at the femtomole level of phosphopeptides from proteolytic digests using 2D-nanoLC-ESI-MS/MS and titanium oxide precolumns. *Analytical Chemistry,* 76**,** 3935-3943.

PINNA, L. A. & RUZZENE, M. 1996. How do protein kinases recognize their substrates? *Biochimica Et Biophysica Acta-Molecular Cell Research,* 1314**,** 191-225.

PRABAKARAN, S., LIPPENS, G., STEEN, H. & GUNAWARDENA, J. 2012. Post-translational modification: nature's escape from genetic imprisonment and the basis for dynamic information encoding. *Wiley Interdiscip Rev Syst Biol Med,* 4**,** 565-83.

RAMASAMY, P., TURAN, D., TICHSHENKO, N., HULSTAERT, N., VANDERMARLIERE, E., VRANKEN, W. & MARTENS, L. 2020. Scop3P: A Comprehensive Resource of Human Phosphosites within Their Full Context. *J Proteome Res,* 19**,** 3478-3486.

RANKIN, N. J., PREISS, D., WELSH, P., BURGESS, K. E. V., NELSON, S. M., LAWLOR, D. A. & SATTAR, N. 2014. The emergence of proton nuclear magnetic resonance metabolomics in the cardiovascular arena as viewed from a clinical perspective. *Atherosclerosis,* 237**,** 287-300.

REINDERS, J. & SICKMANN, A. 2005. State-of-the-art in phosphoproteomics. *Proteomics,* 5**,** 4052-61.

RENAUD, J. P., CHARI, A., CIFERRI, C., LIU, W. T., REMIGY, H. W., STARK, H. & WIESMANN, C. 2018. Cryo-EM in drug discovery: achievements, limitations and prospects. *Nat Rev Drug Discov,* 17**,** 471-492.

RICHARDS, A. L., HEBERT, A. S., ULBRICH, A., BAILEY, D. J., COUGHLIN, E. E., WESTPHALL, M. S. & COON, J. J. 2015. One-hour proteome analysis in yeast. *Nat Protoc,* 10**,** 701-14.

RICHARDSON, J. S. 1981. The anatomy and taxonomy of protein structure. *Adv Protein Chem,* 34**,** 167-339.

ROSE, P. W., PRLIC, A., ALTUNKAYA, A., BI, C., BRADLEY, A. R., CHRISTIE, C. H., COSTANZO, L. D., DUARTE, J. M., DUTTA, S., FENG, Z., GREEN, R. K., GOODSELL, D. S., HUDSON, B., KALRO, T., LOWE, R., PEISACH, E., RANDLE, C., ROSE, A. S., SHAO, C., TAO, Y. P., VALASATAVA, Y., VOIGT, M., WESTBROOK, J. D., WOO, J., YANG, H., YOUNG, J. Y., ZARDECKI, C., BERMAN, H. M. & BURLEY, S. K. 2017. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res,* 45**,** D271-D281.

ROSKOSKI, R. 2015. A historical overview of protein kinases and their targeted small molecule inhibitors. *Pharmacological Research,* 100**,** 1-23.

RUDOLPH, J. D., DE GRAAUW, M., VAN DE WATER, B., GEIGER, T. & SHARAN, R. 2016. Elucidation of Signaling Pathways from Large-Scale Phosphoproteomic Data Using Protein Interaction Networks. Cell Syst, 3, 585-593 e3.

SANGER, F. 1952. The arrangement of amino acids in proteins. *Adv Protein Chem,* 7**,** 1-67.

SCHAFER, K. A. 1998. The cell cycle: a review. *Vet Pathol,* 35**,** 461-78.

SCHOPPER, S., KAHRAMAN, A., LEUENBERGER, P., FENG, Y., PIAZZA, I., MULLER, O., BOERSEMA, P. J. & PICOTTI, P. 2017. Measuring protein structural changes on a proteome-wide scale using limited proteolysis-coupled mass spectrometry. *Nat Protoc,* 12**,** 2391-2410.

SCHOTTE, F. 2003. Watching a Protein as it Functions with 150-ps Time-Resolved X-ray Crystallography. *Science,* 300**,** 1944-1947.

SCHWARTZ, P. A. & MURRAY, B. W. 2011. Protein kinase biochemistry and drug discovery. *Bioorg Chem,* 39**,** 192-210.

SENIOR, A. W., EVANS, R., JUMPER, J., KIRKPATRICK, J., SIFRE, L., GREEN, T., QIN, C. L., Z?DEK, A., NELSON, A. W. R., BRIDGLAND, A., PENEDONES, H., PETERSEN, S., SIMONYAN, K., CROSSAN, S., KOHLI, P., JONES, D. T., SILVER, D., KAVUKCUOGLU, K. & HASSABIS, D. 2019. Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins-Structure Function and Bioinformatics,* 87**,** 1141-1148.

SENIOR, A. W., EVANS, R., JUMPER, J., KIRKPATRICK, J., SIFRE, L., GREEN, T., QIN, C. L., ZIDEK, A., NELSON, A. W. R., BRIDGLAND, A., PENEDONES, H., PETERSEN, S., SIMONYAN, K., CROSSAN, S., KOHLI, P., JONES, D. T., SILVER, D., KAVUKCUOGLU, K. & HASSABIS, D. 2020. Improved protein structure prediction using potentials from deep learning. *Nature,* 577**,** 706-+.

SHARMA, K., D'SOUZA, R. C., TYANOVA, S., SCHAAB, C., WISNIEWSKI, J. R., COX, J. & MANN, M. 2014. Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. Cell Rep, 8, 1583-94.

SOLAN, J. L. & LAMPE, P. D. 2009. Connexin43 phosphorylation: structural changes and biological effects. *Biochemical Journal,* 419**,** 261-272.

SRIVASTAVA, A., NAGAI, T., SRIVASTAVA, A., MIYASHITA, O. & TAMA, F. 2018. Role of Computational Methods in Going beyond X-ray Crystallography to Explore Protein Structure and Dynamics. *Int J Mol Sci,* 19.

STRUMILLO, M. J., OPLOVA, M., VIEITEZ, C., OCHOA, D., SHAHRAZ, M., BUSBY, B. P., SOPKO, R., STUDER, R. A., PERRIMON, N., PANSE, V. G. & BELTRAO, P. 2019. Conserved phosphorylation hotspots in eukaryotic protein domain families. Nat Commun, 10, 1977.

SURYADINATA, R., SADOWSKI, M. & SARCEVIC, B. 2010. Control of cell cycle progression by phosphorylation of cyclin-dependent kinase (CDK) substrates. *Biosci Rep,* 30**,** 243-55.

THINGHOLM, T. E., JENSEN, O. N., ROBINSON, P. J. & LARSEN, M. R. 2008. SIMAC (sequential elution from IMAC), a phosphoproteomics strategy for the rapid separation of monophosphorylated from multiply phosphorylated peptides. *Mol Cell Proteomics,* 7**,** 661-71.

TYANOVA, S., COX, J., OLSEN, J., MANN, M. & FRISHMAN, D. 2013. Phosphorylation variation during the cell cycle scales with structural propensities of proteins. *PLoS Comput Biol,* 9**,** e1002842.

UHLEN, M., FAGERBERG, L., HALLSTROM, B. M., LINDSKOG, C., OKSVOLD, P., MARDINOGLU, A., SIVERTSSON, A., KAMPF, C., SJOSTEDT, E., ASPLUND, A., OLSSON, I., EDLUND, K., LUNDBERG, E., NAVANI, S., SZIGYARTO, C. A., ODEBERG, J., DJUREINOVIC, D., TAKANEN, J. O., HOBER, S., ALM, T., EDQVIST, P. H., BERLING, H., TEGEL, H., MULDER, J., ROCKBERG, J., NILSSON, P., SCHWENK, J. M., HAMSTEN, M., VON FEILITZEN, K., FORSBERG, M., PERSSON, L., JOHANSSON, F., ZWAHLEN, M., VON HEIJNE, G., NIELSEN, J. & PONTEN, F. 2015. Proteomics. Tissue-based map of the human proteome. *Science,* 347**,** 1260419.

ULLAH, S., LIN, S., XU, Y., DENG, W., MA, L., ZHANG, Y., LIU, Z. & XUE, Y. 2016. dbPAF: an integrative database of protein phosphorylation in animals and fungi. *Sci Rep,* 6**,** 23534.

UNIPROT, C. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res,* 47**,** D506-D515.

VARADI, M., VRANKEN, W., GUHAROY, M. & TOMPA, P. 2015. Computational approaches for inferring the functions of intrinsically disordered proteins. *Front Mol Biosci,* 2**,** 45.

VIZCAINO, J. A., DEUTSCH, E. W., WANG, R., CSORDAS, A., REISINGER, F., RIOS, D., DIANES, J. A., SUN, Z., FARRAH, T., BANDEIRA, N., BINZ, P. A., XENARIOS, I., EISENACHER, M., MAYER, G., GATTO, L., CAMPOS, A., CHALKLEY, R. J., KRAUS, H. J., ALBAR, J. P., MARTINEZ-BARTOLOME, S., APWEILER, R., OMENN, G. S., MARTENS, L., JONES, A. R. & HERMJAKOB, H. 2014. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol,* 32**,** 223-6.

WANG, C., XU, H., LIN, S., DENG, W., ZHOU, J., ZHANG, Y., SHI, Y., PENG, D. & XUE, Y. 2020. GPS 5.0: An Update on the Prediction of Kinase-specific Phosphorylation Sites in Proteins. *Genomics Proteomics Bioinformatics,* 18**,** 72-80.

WANG, H. W. & WANG, J. W. 2017. How cryo-electron microscopy and X-ray crystallography complement each other. *Protein Science,* 26**,** 32-39.

WEISS, S. 1999. Fluorescence spectroscopy of single biomolecules. *Science,* 283**,** 1676-83.

WETLAUFER, D. B. 1973. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci U S A,* 70**,** 697-701.

WILLEMS, E., DEDOBBELEER, M., DIGREGORIO, M., LOMBARD, A., LUMAPAT, P. N. & ROGISTER, B. 2018. The functional diversity of Aurora kinases: a comprehensive review. *Cell Div,* 13**,** 7.

YAN, X. & MAIER, C. S. 2009. Hydrogen/deuterium exchange mass spectrometry. *Methods Mol Biol,* 492**,** 255-71.

YANG, J., YAN, R., ROY, A., XU, D., POISSON, J. & ZHANG, Y. 2015. The I-TASSER Suite: protein structure and function prediction. *Nat Methods,* 12**,** 7-8.

YU, H. 1999. Extending the size limit of protein nuclear magnetic resonance. *Proc Natl Acad Sci U S A,* 96**,** 332-4.

ZHANG, H. & GE, Y. 2011. Comprehensive analysis of protein modifications by top-down mass spectrometry. *Circ Cardiovasc Genet,* 4**,** 711.

ZHANG, J. M., YANG, P. L. & GRAY, N. S. 2009. Targeting cancer with small molecule kinase inhibitors. *Nature Reviews Cancer,* 9**,** 28-39.