TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

EVALUATION OF THE RELEVANCE AND IMPACT OF KINASE DYSFUNCTION IN NEUROLOGICAL DISORDERS

Through proteomics and phosphoproteomics bioinformatics analysis

Ye Hong

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. F OSA – TOM. 26 | TECHNICA – INFORMATICA| TURKU 2023

University of Turku

Faculty of Technology
Department of Computing
Computer Science
Doctoral programme in Bioinformatics

Supervised by

Professor Laura L. Elo
Turku Bioscience Centre
University of Turku

Professor Eleanor T. Coffey
Turku Bioscience Centre
Åbo Akademi University

Reviewed by

Professor, Sampsa Hautaniemi
University of Helsinki

Docent, Sepinoud Azimi Rashti
Åbo Akademi University

Opponent

Professor, Jacques Colinge
University of Montpellier

ISBN 978-951-29-9409-0 (PRINT)
ISBN 978-951-29-9408-3 (PDF)
ISSN 2736-9390 (Print)
ISSN 2736-9684 (Online)
Painosalama, Turku, Finland 2023

UNIVERSITY OF TURKU
Faculty of Technology
Department of Computing
Computer Science
YE HONG: Evaluation of the relevance and impact of kinase dysfunction in neurological disorders through proteomics and phosphoproteomics bioinformatics analysis
Doctoral Dissertation, 192 pp.
Doctoral Programme in Technology
August 2023

ABSTRACT

Phosphorylation is an important post-translational modification that is involved in various biological processes and its dysregulation has in particular been linked to diseases of the central nervous system including neurological disorders. The present thesis characterizes alterations in the phosphoproteome and protein abundance associated with schizophrenia and Parkinson's disease, with the goal of uncovering the underlying disease mechanisms. To support this goal, I eventually created an automated analysis pipeline in R to streamline the analysis process of proteomics and phosphoproteomics data.

Mass spectrometry (MS) technology is utilized to generate proteomics and phosphoproteomics data. Study I of the thesis demonstrates an automated R pipeline, PhosPiR, created to perform multi-level functional analyses of MS data after the identification and quantification of the raw spectral data. The pipeline does not require coding knowledge to run. It supports 18 different organisms, and provides analyses of MS intensity data from preprocessing, normalization and imputation, through to figure overviews, statistical analysis, enrichment analysis, PTM-SEA, kinase prediction and activity analysis, network analysis, hub analysis, annotation mining, and homolog alignment.

The LRRK2-G2019S mutation, a frequent genetic cause of late onset Parkinson's disease, was investigated in Study II and III. One study investigated the mechanism of LRRK2-G2019S function in brain, and the other identified proteins with significantly altered overall translation patterns in sporadic and LRRK2-G2019S patient samples. Specifically, study II identified that LRRK2 is localized to the small 40S ribosomal subunit and that LRRK2 activity suppresses RNA translation, as validated in cell and animal models of Parkinson's disease and in patient cells. Study III utilized bio-orthogonal non-canonical amino acid tagging to label newly translated proteins in order to identify which proteins were affected by repressed translation in patient samples, using mass spectrometry analysis. The analysis revealed 33 and 30 nascent proteins with reduced synthesis in sporadic and LRRK2-G2019S Parkinson’s cases, respectively. The biological process "cytosolic signal recognition particle (SRP)-dependent co-translational protein targeting to

membrane" was functionally significantly affected in both sporadic and LRRK2-G2019S Parkinson's, while "Tubulin/FTsz C-terminal domain superfamily network" was only significantly enriched in LRRK2-G2019S Parkinson's cases. The findings were validated bytargeted proteomics and immunoblotting.

Study IV is conducted to investigate the role of JNK1 in schizophrenia. Wild type and *Jnk1-/-* mice were used to analyze the phosphorylation profile using LC-MS/MS analysis. 126 proteins associated with schizophrenia were identified to overlap with the significantly differentially phosphorylated proteins in *Jnk1-/-* mice brain. The NMDAR trafficking pathway was found to be highly enriched, and surface staining of NMDAR subunits in neurons showed that surface expression of both subunits in *Jnk1-/-* neurons was significantly decreased. Further behavioral tests conducted with MK801 treatment have associated the *Jnk1-/-* molecular and behavioral phenotype with schizophrenia and neuropsychiatric disease.

KEYWORDS: Phosphoproteomics, proteomics, PhosPiR, JNK1, schizophrenia, Parkinson's disease, MS analysis

TURUN YLIOPISTO
Teknillinen tiedekunta
Tietotekniikan laitos
Tietojenkäsittelytieteet
YE HONG: Evaluation of the relevance and impact of kinase dysfunction in neurological disorders through proteomics and phosphoproteomics bioinformatics analysis
Väitöskirja, 192 pp.
Teknologian tohtoriohjelma
Elokuu 2023

TIIVISTELMÄ

Fosforylaatio on tärkeä translaation jälkeinen muokkaus, jolla on rooli useissa biologisissa prosesseissa. Fosforylaatioon liittyvillä säätelyhäiriöillä on erityinen yhteys keskushermoston sairauksiin, mukaan lukien neurologiset häiriöt. Tämä väitöskirja kuvaa fosfoproteomin ja proteiinitasojen muutoksia, jotka liittyvät skitsofreniaan ja Parkinsonin tautiin, tavoitteena paljastaa näiden tautien taustalla olevat mekanismit. Tätä tavoitetta tukemaan loin automatisoidun analyysisovelluskokoonpanon R:ssä proteomiikan ja fosfoproteomiikan datan analysointiprosessin virtaviivaistamiseksi.

Massaspektrometri (MS) -teknologiaa käytetään proteomiikan ja fosfoproteomiikan datan tuottamiseen. Väitöskirjan I tutkimus esittelee automatisoidun R-sovelluksen, PhosPiR:n, joka on luotu suorittamaan moniulotteisia toiminnallisia analyysejä MS-datalle raakaspektridatan tunnistamisen ja kvantifioinnin jälkeen. Sovelluksen ajaminen ei vaadi ohjelmointitaitoa. Se tukee 18:aa eri organismia ja tarjoaa MS-intensiteettidatan analyysit esikäsittelystä, normalisoinnista ja imputoinnista aina kuvaesittelyihin, tilastolliseen analyysiin, rikastamisnalyysiin, PTM-SEA:han, kinaasi-ennustamiseen ja -toiminta-, verkko- ja solmuanalyysiin, sekä annotaatiolouhimiseen ja homologien vertailuun.

LRRK2-G2019S-mutaatiota, joka on yleinen geneettinen syy myöhäisen alkamisajan Parkinsonin taudille, tutkittiin tutkimuksessa II ja III. Tutkimus II tarkasteli LRRK2-G2019S:n toiminnan mekanismia aivoissa ja Tutkimus III tunnisti proteiineja, joiden translaatio oli muuttunut satunnaisissa ja LRRK2-G2019S Parkinsonin taudin potilasnäytteissä. Tutkimus II:ssa havaittiin, että LRRK2 sijaitsee ribosomin 40S-alayksikössä ja LRRK2-aktiivisuus säätelee RNA-translaatiota. Tämä vahvistettiin Parkinsonin taudin solu- ja eläinmalleissa, sekä potilasperäisissä soluissa. Tutkimus III käytti bio-ortogonaalista ei-kanonista aminohappoleimausta vastasyntetisoitujen proteiinien havaitsemisessa. Tämä leimaus mahdollisti translaatiohäiriöstä kärsivien proteiinien tunnistamisen potilasnäytteistä MS-analyysissä. Analyysi paljasti 33 ja 30 vastasyntetisoitua proteiinia, joiden synteesi oli alentunut satunnaisissa ja LRRK2-G2019S Parkinsonin taudin näytteissä. Biologinen prosessi "cytosolic signal recognition particle (SRP)-dependent co-

translational protein targeting to membrane" oli merkittävästi muuttunut sekä satunnaisissa että LRRK2-G2019S Parkinsonin tapauksissa, kun taas "Tubulin/FTsz C-terminal domain superfamily network" oli merkittävästi rikastunut vain LRRK2-G2019S Parkinsonin tapauksissa. Löydökset vahvistettiin kohdennetulla MS-analyysillä ja Western-blot menetelmällä.

Tutkimus IV:ssä tutkittiin JNK1:n roolia skitsofreniassa. Villityypin ja Jnk1-/- -hiiriä käytettiin fosforylaatioprofiilin analysointiin MS-analyysillä. 126 skitsofreniaan liittyvää proteiinia oli merkittävästi päällekkäisiä Jnk1-/- -hiiren fosfoproteiinie kanssa. NMDAR-trafikointireitin havaittiin olevan merkittävästi rikastunut, ja NMDAR alayksiköiden pintavärjäys neuroneissa osoitti, että molempien alayksiköiden pintailmentyminen Jnk1-/- neuroneissa oli merkittävästi vähentynyt. Lisäksi MK801-hoidolla suoritetut yhdistivät Jnk1-/- molekulaarisen ja -käyttäytymisfenotyypin skitsofreniaan ja neuropsykiatriseen sairauteen.

ASIASANAT: Fosfoproteomiikka, proteomiikka, PhosPiR, JNK1, skitsofrenia, Parkinsonin tauti, MS-analyysi

Table of Contents

Abbreviations

ABL	Abelson Murine Leukemia Viral Oncogene Homolog
ACN	Acetonitrile
AGC	Protein Kinase A, G, And C Families
AHA	L-Azidohomoalanine
AKT	Protein Kinase B
ALS	Amyotrophic Lateral Sclerosis
ATP	Adenosine 5'-Triphosphate
AUC	Area under the curve
BLOSUM	Block Amino Acid Substitution Matrices
BONCAT	Bio-Orthogonal Non-Canonical Amino Acid Tagging
CaMK	Ca2+/Calmodulin-Dependent Protein Kinases
CAMKII	Ca2+-Calmodulin-Dependent Protein Kinase II
CDK	Cyclin-Dependent Kinases
CK1	Casein Kinase 1
CK2	Casein Kinase 2
CLK	Cyclin-Dependent Kinase Like Kinases
CMGC	Cyclin-Dependent Kinases, Mitogen-Activated Protein Kinases, Glycogen Synthase Kinase-3S, and Dual Specificity Protein Kinase CLKs
CML	Chronic Myeloid Leukemia
cMyBP-C	Cardiac Myosin-Binding Protein-C
CV	Coefficients Of Variation
DDA	Data-Dependent-Acquisition
DIA	Data-Independent-Acquisition
DNA	Deoxyribonucleic Acid
DSK	Dual-Specificity Kinases
DYRK1A	Dual-Specificity Tyrosine Phosphorylation-Regulated Kinase 1A
EGFR	Epidermal Growth Factor Receptor
EM	Expectation-Maximization
ER	Endoplasmic Reticulum
ERK	Extracellular Signal Regulated Kinase

ESI	Electrospray Ionization
FDR	False Discovery Rate
GO	Gene Ontology
GPCR	G-Protein Coupled Receptors
GSEA	Gene Set Enrichment Analysis
GSK3	Glycogen Synthase Kinase-3S
GUI	Graphical User Interface
GWAS	Gene-Wide Association Study
Her2	Human Epidermal Growth Factor Receptor 2
HMW-MAP2	High Molecular Weight Forms of Microtubule-Associated Protein 2
IMAC	Immobilized Metal Ion Affinity Chromatography
IRAK	Interleukin-1 Receptor-Associated Kinase
JAK	Janus Tyrosine Kinase
JNK	C-Jun N-Terminal Kinases
JNK1	C-Jun N-Terminal kinase 1
KEGG	Kyoto Encyclopedia of Genes And Genomes
KNN	K Nearest Neighbors
LC-MS	Liquid Chromatography Mass Spectrometry
LLS	Local Least-Squares Imputation
LOD	Limit of Detection
LRRK	Leucine-Rich Repeat Serine/Threonine-Protein Kinase
LSA	Least-Squares Adaptive Imputation
MAP	Microtubule-Associated Proteins
MAPK	Mitogen-Activated Protein Kinases
MAR	Missing At Random
MBI	Model-Based Imputation
MCAR	Missing Completely At Random
MEK	Mitogen-Activated Protein Kinase Kinase
MLE	Maximum Likelihood Estimates
MNAR	Missing Not At Random
MS	Mass Spectrometry
MSA	Multiple System Atrophy
mTOR	Mammalian Target Of Rapamycin
NA	Not Available
NINDS	National Institute of Neurological Disorders
NMDAR	N-methyl-D-aspartate receptor
$p15^{INK4B}$	Cyclin-Dependent Kinase 4 Inhibitor B
p21	Cyclin-Dependent Kinase Inhibitor 1
PAK	P21-Activated Kinase

PEP	Posterior Error Probability
PKA	Protein Kinase A
PKC	Protein Kinase C
PKG	Protein Kinase G
PPCA	Probabilistic Principal Component Analysis
PPI	Paired Pulse Inhibition
pRb	Retinoblastoma Protein Phosphorylation
PRM	Parallel Reaction Monitoring
PSM	Peptide Spectrum Match
PTM	Post Translational Modification
PTM-SEA	Post Translational Modification Set Enrichment Analysis
PTSD	Post-Traumatic Stress Disorder
Raf	Rapidly Accelerated Fibrosarcoma
REM	Regularized Expectation Maximization Algorithm
ROC curve	Receiver operating characteristic curve
ROCK	Rho-Associated Coiled-Coil-Containing Protein Kinases
RTI	Random Tail Imputation
SAPK	Stress-Activated Protein Kinases
SCHEMA	Schizophrenia Exome Meta-Analysis Consortium
SCX	Strong Cation Exchange Chromatography
SERCA2A Ca2+	Sarcoplasmic/Endoplasmic Reticulum Ca2+
SMKI	Small-Molecule Kinase Inhibitor
Src	Proto-Oncogene Tyrosine-Protein Kinase Sarcoma
SRP	Cytosolic Signal Recognition Particle
STE	Sterile Kinases
STK	Serine/Threonine Kinases
Tau	Tubulin Associated Unit
TGF β	Transforming Growth Factor B
TK	Tyrosine Kinases
TKL	Tyrosine Kinase-Like
TNF	Tumor Necrosis Factor
TNGB	Telethon Network of Genetics Biobanks
TUH	Turku University Hospital
UPDRS	Unified Parkinson's Disease Rating Scale
VEGFR	Vascular Endothelial Growth Factor Receptor
Vsn	Variance Stabilization Normalization

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I Ye Hong, Dani Flinkman, Tomi Suomi, Sami Pietilä, Peter James, Eleanor Coffey, Laura Elo. PhosPiR: an automated phosphoproteomic pipeline in R. *Briefings in Bioinformatics*, 2022; 23(1): bbab510.

II Prasannakumar Deshpande, Dani Flinkman, Ye Hong, Elena Goltseva, Valentina Siino, Lihua Sun, Sirkku Peltonen, Laura Elo, Valtteri Kaasinen, Peter James, Eleanor Coffey. Protein synthesis is suppressed in sporadic and familial Parkinson's disease by LRRK2. *FASEB J.*, 2020; 34(11): 14217-14233.

III Dani Flinkman, Ye Hong, Jelena Gnjatovic, Prasannakumar Deshpande, Zuzsanna Ortutay, Sirkku Peltonen, Valtteri Kaasinen, Peter James, Eleanor Coffey. Regulators of proteostasis are translationally repressed in fibroblasts from patients with sporadic and LRRK2-G2019S Parkinson's disease. *NPJ Parkinson's Disease*, 2023; 9(20).

IV Ye Hong, Jismi John, Artemis Varidaki, Nikita Tiwari, Paolo Cifani, Anni-Maija Linden, Raghavendra Mysore, Esa Korpi, Laura Elo, Peter James, Eleanor Coffey. Jun kinase regulates 126 schizophrenia polygenes providing a model for schizophrenia. *Manuscript*.

1 Introduction

1.1 Overview of the objectives of this thesis

Phosphorylation is a highly prevalent and essential post-translational modification that plays a critical role in various biological processes. Dysregulation of phosphorylation signaling has been implicated in the pathogenesis of various neurological disorders, including chronic depression, Alzheimer's disease, Parkinson's disease, and schizophrenia. By Investigating the interplay between phosphorylation and changes in protein expression, underlying disease mechanisms can be elucidated, and potential drug targets can be identified. The present thesis focuses on characterizing alterations in the phosphoproteome and protein abundance associated with two such disorders, schizophrenia (Study IV) and Parkinson's disease (Study II and III), with the aim of uncovering disease mechanisms and associated regulatory networks and pathways. To streamline the analysis process, an automated R pipeline was developed (Study I). This integrated various analysis methods from the previous studies with additional useful phosphoproteomics analysis methods, saving weeks of analysis work for the users, and without a requirement for coding knowledge.

1.2 Kinase and phosphorylation

This section provides an overview of the biochemical process of phosphorylation, a specific protein post-translational modification (PTM) catalyzed by proteins known as "protein kinases." Various PTMs are elucidated, emphasizing the significance of phosphorylation in cellular contexts and its essential role in diverse biological processes. The implications of kinase activity in pathological conditions, and the therapeutic potential of targeting kinases in drug development are also discussed.

1.2.1 Phosphorylation overview

Phosphorylation is a type of protein post translational modification (PTM) that occurs on proteins in a cellular context. PTMs are covalent, biochemical

modifications to proteins which result in the addition of a chemical moiety [1]. They include acetylation, glycosylation, methylation, phosphorylation, ubiquitination, nitrosylation, sumoylation, carboxylation, hydroxylation, proteolytic cleavage, amidation, and disulfide bond formation [2]. PTMs are crucial regulators of protein function and play a key role in diverse biological processes [1]. Among them, phosphorylation is one of the most common, and widely studied PTMs that is essential for biological function [3].

Phosphorylation is a fully reversible process which is catalyzed by a group of proteins known as "protein kinases". During a phosphorylation event, the γ-phosphate (PO_4) from adenosine 5'-triphosphate (ATP) is added to the polar group R of different amino acid residues [3] [4] (Figure 1). Commonly modified residues include serine (Ser or S), threonine (Thr or T), and tyrosine (Tyr or Y). Together they make up more than one third of all phosphorylation events [3]. Of those, serine residue is most favored, constitutes 86.4% of the phosphorylation events, followed by threonine, which constitutes 11.8%, and tyrosine, which is the least common out of the three, accounting for only 1.8% [5]. Other than these three residues, noncanonical residues such as histidine (His or H) and aspartate (Asp or D) have also been found to be phosphorylated [3]. It was believed these residues are rare and less stable than the three common residues, however, recent studies have shown histidine phosphorylation, in particular, in fact partakes critical roles in cellular regulatory mechanisms, and is surprisingly common in bacteria, constituting 10% of Escherichia coli phosphorylation events for example [4].

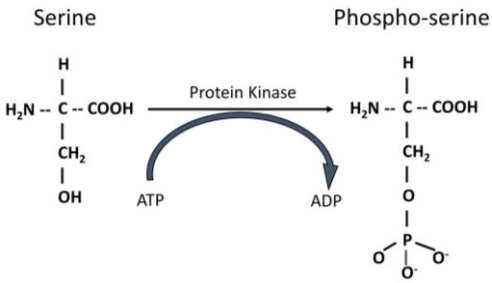

Figure 1. An example of a phosphorylation reaction. In this example, a serine residue is phosphorylated to phospho-serine residue.

1.2.2 Kinase regulation families

Protein kinases are responsible for adding the phosphate group to different phosphosites. Their activities are subjected to regulation in one of three ways: by the kinase itself through autophosphorylation, by binding with another protein known as activator, leading to transphosphorylation (allosteric regulation), or by controlling

its localization in relation to its substrates [6]. There are 518 human protein kinases discovered, and all of them are categorized by their substrate R group residue. Serine/threonine kinases (STKs) phosphorylate both serine and threonine residues [7], tyrosine kinases (TKs) phosphorylate tyrosine residues [8], and dual-specificity kinases (DSKs) phosphorylate all three residues [8]. STKs are most well-known, at least 125 human kinases belong to this category. They target the OH group of serine and threonine, and are activated by a variety of physiological events such as deoxyribonucleic acid (DNA) damage or chemical signals from e.g. cAMP [7].

Within the three main categories, kinases are further divided into subfamilies, particularly, CaMK, CK1, TK, STE, AGC, TKL, and CMGC subfamilies (Figure 2). CaMK stands for Ca^{2+}/calmodulin-dependent protein kinases. They respond to an increase in intracellular Ca^{2+} concentration. Once activated, they phosphorylate the serine or threonine residues of several transcription factors, making their activity crucial for many gene expression regulations [9]. CK1 stands for casein kinase 1, or cell kinase 1. This subfamily has seven members, and each is a monomeric enzyme which phosphorylates serine or threonine specifically (serine/threonine-selective). They regulate signal transduction pathways such as circadian rhythms, DNA repair and DNA transcription [10]. TK stands for tyrosine kinases. They are cell surface receptors that takes care of surface related functions, and they only phosphorylate tyrosine residues [11]. STE stands for sterile kinase. This family consists of three main groups, Ste7, Ste20 and Ste11, which cascades to eventually activate the mitogen-activated protein kinases (MAPK) [12]. AGC stands for protein kinase A, G, and C families (PKA, PKC, and PKG). They are a subgroup of STKs with similar catalytic kinase domains [13]. TKL stands for tyrosine kinase-like. They are kinases which are similar in sequence with the TK subfamily, however, they belong to the STKs category. Interleukin-1 receptor-associated kinase (IRAK), leucine-rich repeat serine/threonine-protein kinase (LRRK), and RAF proto-oncogene serine/threonine-protein kinase (RAF) are a few examples of kinases from this subfamily [14]. CMGC stands for cyclin-dependent kinases (CDK), mitogen-activated protein kinases (MAPK), glycogen synthase kinase-3s (GSK3) and dual specificity protein kinase CLKs (CLK). These four sets are well studied and participate in important regulatory functions. CDK regulates the various phases of cell cycle. MAPK regulates cellular processes such as proliferation, differentiation, and death, and is closely related to oncogenic pathology [15]. GSK3 kinases α and β were originally known as key enzymes in glycogen metabolism, before they were understood to be kinases with a diverse assembly of roles. They are especially important during the embryonic development period [16]. CLK kinases are involved in regulating pre-mRNA processing, and indirectly modulating splice site selection [17].

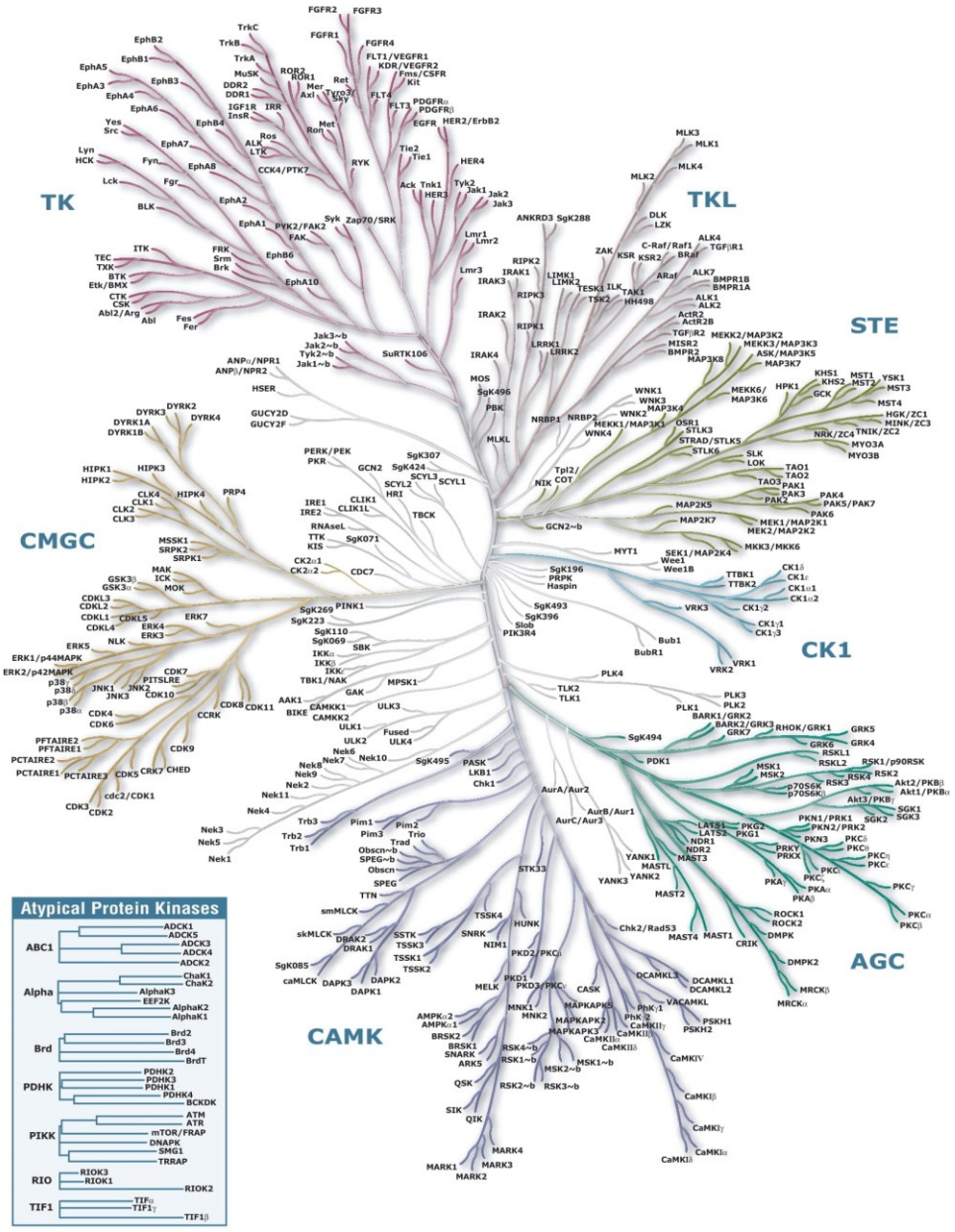

Figure 2. An illustration of the human kinome. The dendrogram shows the sequence similarity of kinase domains. Illustration reproduced courtesy of Cell Signaling Technology, Inc. (www.cellsignal.com).

1.2.3 Kinase function

It is revealed that more than ten thousand distinct phosphorylation events take place in human cells [4]. Moreover, greater than two thirds of the proteins encoded by the human genome are phosphorylated, many on multiple sites, with an estimation that 90% of all proteins will be found to be subject to phosphorylation with future research [3]. The ubiquitous nature of phosphorylation alone is an indication of its functional importance. The addition of a phosphate group transforms the local polarity of a protein, converting it from hydrophobic apolar to hydrophilic polar, in turn changing the confirmation of the protein, and allowing it to actively bind other molecules [18]. The assembly of protein complexes through phosphorylation has established the foundation for the intricate network of protein-protein interactions. Altering the phosphorylation state of a single protein could modify the activity, localization, and interactions of a chain of proteins linked by its interaction network. Due to the widespread influence, protein phosphorylation is of vital importance in virtually all cellular processes, protein synthesis, cell growth, signal transduction, cell division, and aging are just a few instances whose activation are regulated by phosphorylation from specific kinases [19].

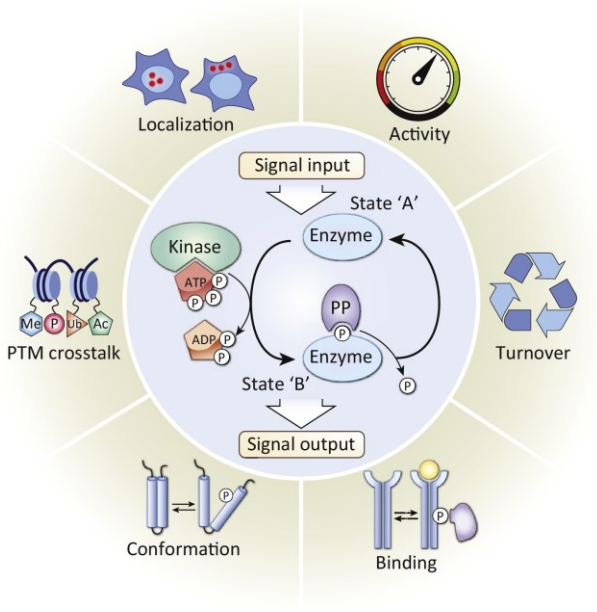

Figure 3. Six ways in which phosphorylation modulates protein function, including enzymatic activity, protein turnover, interactions, conformation, localization, and crosstalk with other PTMs. Figure is reprinted from [20] with permission from Elsevier.

As illustrated in Figure 3 protein phosphorylation plays a critical role in regulating various protein functions, which enable the execution of diverse biological processes. The phosphorylation regulatory mechanism can be classified into several categories. It could, first, serve as a molecular switch, whereby proteins become activated upon phosphorylation, and carry on their intended functions. Processes in cell survival and cell growth are regulated in this way [21], [22]. Phosphorylation could facilitate temporary protein-protein interactions, where only after phosphorylation of the protein would it interact with another protein to form a functional complex. This regulation provided the means to adjust many signaling pathways [23]. Another regulation strategy is to trigger subcellular translocation by protein phosphorylation, to send proteins to/from their functional sites. As an example, apoptosis of T and B cells is regulated in this way [24]. Phosphorylation is involved in the ATP production cycle, which gave it a key role in reactions which require energy [25]. And lastly, phosphorylation can regulate another PTM. Its involvement in the insulin signaling pathway utilizes this mechanism [26]. These methods of regulation are intertwined through the phosphorylation signaling network comprised of protein kinases, phosphatases, and their substrate binding sites [27].

The phosphorylation database PhosphoSitePlus have documented over 850 unique binding sites referred to as phosphosites and kept a record of another 1,000 plus phosphosites from predictions [3]. Many different sites can belong to the same protein; however, it was observed not all of them are functional. Hence two categories of phosphorylation exist. The first category is stable phosphorylation, it is believed that all stable phosphosites are functional. The second category is transitory phosphorylation, these phosphorylations are unstable and believed to have no functional effect [28], [29]. The determination of a phosphosite's stability depends solely on the site itself without environmental factors playing a role [30]. Hence, the study of the phosphorylation signaling network would guarantee an accurate pinpointing of local phosphorylation functions, which would be beneficial to the understanding of pathological mechanisms.

1.2.4 Kinase disease relevance

Dysregulation at any point on a given phosphorylation signaling network could create a ripple effect that leads to unwanted consequences for the cell. Therefore, phosphorylation anomalies are hallmarks of many diseases, including numerous cancers [31], cardiac diseases [32], neurological disorders [33], and even the recent Coronavirus pandemic [34].

Utilizing the field of cancer biology research as example, cancers are well known for their genetic mutation mechanisms; however, epigenetic changes have also been

a crucial mechanism in cancer [35]. More than 1,000 alternation patterns in kinase expression from human tumors have been revealed so far [31], the most commonly known alterations associated with cancer pathology are concentrated within a few subfamilies of kinases, such as tyrosine kinases, mitogen-activated protein kinases (MAPK), and cyclins [36]. Tyrosine kinases are activated by growth factors and hormones. Once activated, they auto-phosphorylate and phosphorylate downstream proteins to regulate intercellular communication and homeostasis [37]. Aberration of these kinases could cause uncontrolled cell growth and cell division, creating a recipe for oncogenesis. The first discovered proto-oncogene in vertebrates, Proto-oncogene tyrosine-protein kinase sarcoma (Src), belongs to this subfamily of kinase [38]. HER2 [39] and mammalian target of rapamycin (mTOR) [40] are also tyrosine kinases. MAPK are involved in many cellular processes including proliferation, differentiation, and apoptosis [41]. These kinases interact with each other to form a complexed signaling pathway which is regulated with precision by phosphorylation and dephosphorylation. Changes in regulation of the MAPK cascade are often found in cancer. Rapidly accelerated fibrosarcoma (Raf) and Mitogen-activated protein kinase kinase (MEK) are MAPKs that are well known for their involvement in cancer progression [42]. Cyclins regulate the cell cycle. Disruption in their function is found in a variety of human cancers. Cyclin D1 belongs to this group; phosphorylation of this kinase activates its transportation from nucleus, and degradation in cytoplasm. Disruption in its phosphorylation causes its accumulation in nucleus, which increases oncogenic potential, and is known to be associated with esophageal cancer [43] (Figure 4). Transforming growth factor β (TGF β) also belongs to this kinase group. TGF β deactivates retinoblastoma protein by preventing its phosphorylation [44], at the same time, activates the synthesis of Cyclin-dependent kinase 4 inhibitor B ($p15^{INK4B}$) and cyclin-dependent kinase inhibitor 1 (p21), which promote retinoblastoma protein phosphorylation (pRb) by blocking cyclin-CDK complexes [45]. pRb hypophosphorylation halts the cell cycle in G1 phase by inhibiting the expression of genes which signals the cell to transit into S phase [46]. TGF β which takes a key role in balancing the phosphorylation of pRb is often found with altered activity in human cancers [45].

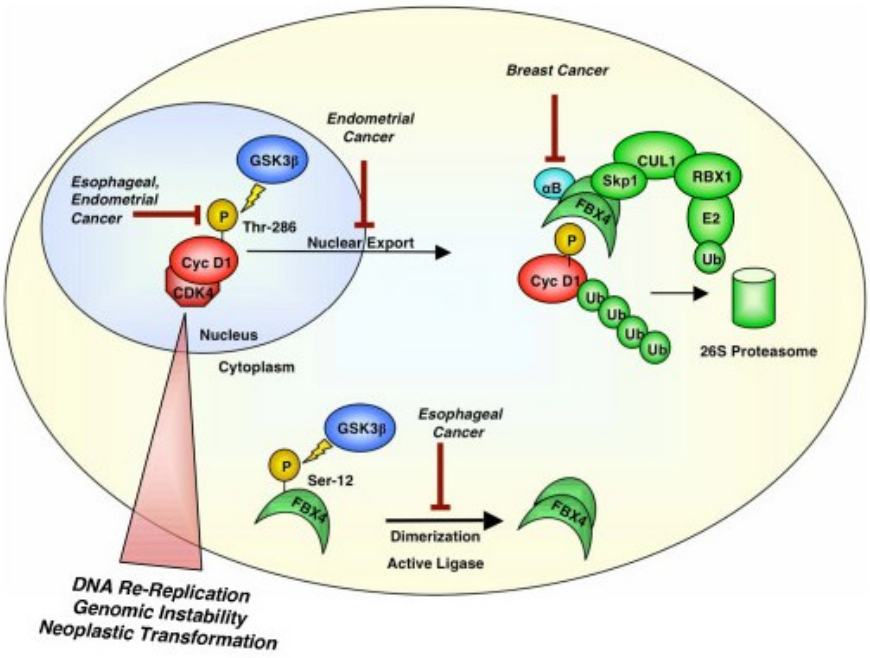

Figure 4. Illustration of Cyclin D1's association with human cancers. Cyclin D1 phosphorylation disruption in the nucleus is associated with esophageal cancer [47].

Neurological disorders, of course, share a similar mechanistic connection to phosphorylation regulation and their associated kinases as do other diseases. A well-known case is tubulin associated unit (Tau) and its relation to Alzheimer's disease (Figure 5). The signature of neurofibrillary tangles present in Alzheimer's patient brains is the result of Tau hyperphosphorylation [33]. Another kinase, dual-specificity tyrosine phosphorylation-regulated kinase 1A (DYRK1A), according to one study, is dysfunctional in a variety of human neurological disorders, including Down syndrome, dementia, Parkinson's disease and autism [48]. The current thesis work focuses on JNK phosphorylation in the brain, and Parkinson's disease protein expression and phosphorylation changes, hence these specific phosphorylation aberration in the brain will be discussed in detail in the result section.

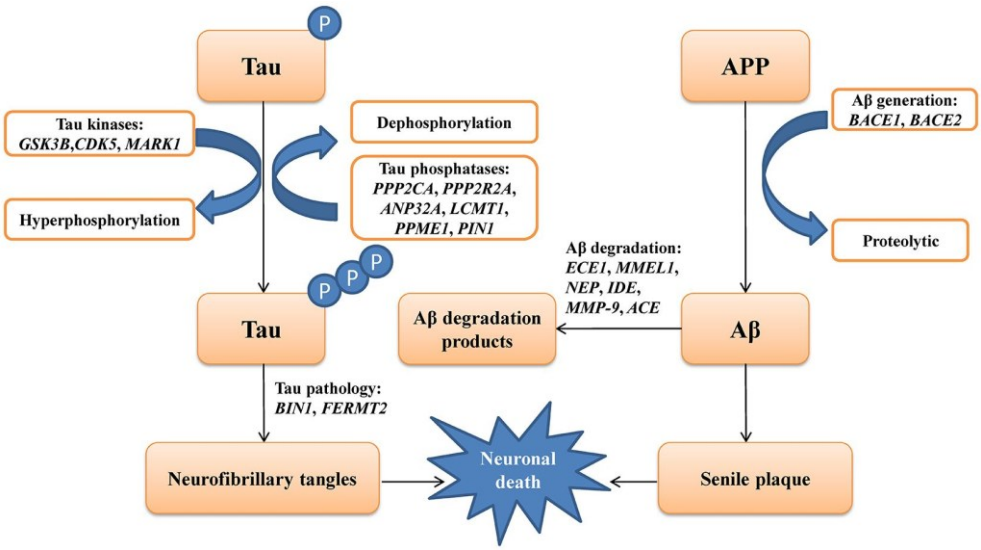

Figure 5. A schematic diagram showing the proteins involved in Tau hyperphosphorylation and amyloid-β metabolic pathway which contributies to neuronal death in Alzheimer's disease [49].

1.2.5 Kinase as drug targets

Due to their widespread influence in disease pathologies, kinases have captured the attention of the pharmaceutical industries since the 1980s [50]. However, due to technology limitations at that time, the first small-molecule kinase inhibitor (SMKI) was not available on the market until the 1990s. The first approved SMKI was Fasudil, approved in Japan in 1995, the drug treats cerebral vasospasm by inhibiting Rho-associated coiled-coil-containing protein kinases 1 and 2 (ROCK1 and ROCK2) [50]. The first kinase inhibitor approved by the US FDA was Sirolimus, which reached US market in 1999, and was purposed to prevent organ rejection [51]. The most impactful addition to the FDA approval list of SMKI is perhaps Imatinib, which was approved in 2001. It is intended to treat chronic myeloid leukemia (CML) by inhibiting the tyrosine kinase abelson murine leukemia viral oncogene homolog (ABL) [52]. The previous interferon treatment had a patient resistant rate of 95%; when switched to Imatinib, patients experienced complete hematological response, and had an 89% estimated progression-free survival rate [53]. Due to the enormous success of this particular drug, a surge of new SMKI intended for oncology therapy had flooded to the market, many of which are also tyrosine kinase inhibitors.

Around 89% of FDA approved SMKIs are purposed for oncology treatment, and TK subfamily is the most targeted group for these drugs to this day [50]. Beyond oncology treatment, SMKIs are also purposed for immune system related therapies, 11 SMKIs are approved for such purpose [50]. Ruxolitinib, approved in 2011, was the first SMKI approved by FDA that was not oncology related. It is intended to treat

patients with intermediate to high risk of myelofibrosis by inhibiting Janus tyrosine kinase 1 and 2 (JAK1 and JAK2) [54]. A small number of SMKIs are targeting diseases aside from oncology and immunology. For example, Everolumus treats tuberous sclerosis complex-associated partial-onset seizures by inhibiting mTOR [55], and Nintedanib treats idiopathic pulmonary fibrosis by inhibiting vascular endothelial growth factor receptor (VEGFR) [56]. Following the approval of the first SMKI in 1995, a total of 71 SMKIs have since been approved. Figure 6 lists the first SMKIs that targets a specific kinase family in their respective year of validation.

Due to recent advancement in phosphoproteome and kinase studies, the number of approved SMKIs have doubled in the past five years and comprises 15% of all approved novel drugs during this time [50]. Despite the substantial increase in the SMKI numbers, it is indicated that at least 70% of all kinases are still unexplored [50], leaving room for new studies to discover novel therapeutic options in a broader range of kinases. Furthermore, drug response variation and side effects can stem from complexed regulation network alternations, which can be better understood through the study of phosphorylation signaling network, where kinase activities and its influence through the signaling network is analyzed and revealed, hence providing the mechanistic insights for drug efficacy and potential side effects in vivo. Such dependency highlights the importance of phosphoproteome studies in drug discovery, as it is a more accurate predictor of the phosphorylation signaling network than genomic landscape studies, it is essential to the development of better therapeutic practices in the future [4].

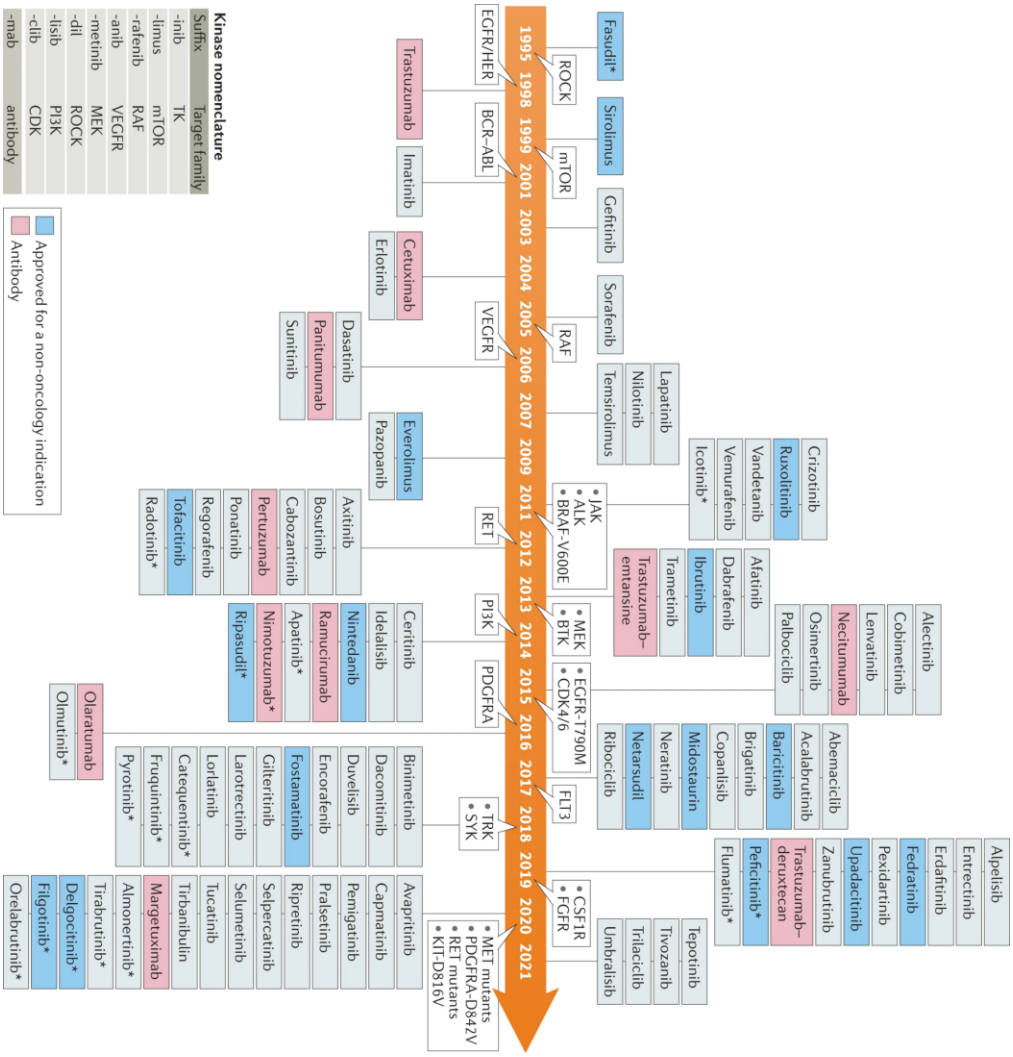

Figure 6. A timeline illustrating the first small-molecule kinase inhibitors (SMKI) that targets a new kinase family in their respective year of being validated [50]. This timeline is reproduced with permission from SNCSC.

1.3 c-Jun N-terminal kinase (JNK) and MAPK signaling

This section provides a comprehensive insight into c-Jun N-terminal kinases (JNKs), part of the mitogen-activated protein kinase (MAPK) family, elucidating their role, structure, and the complexity of the signaling transduction cascade known as the "three-tiered" MAPK pathways. Subsection 1.1.1 introduces JNKs, detailing their discovery, initial categorization, structure, and multifaceted activation in response to

various stimuli. Subsection 1.1.2 explores the relevance of JNK to schizophrenia, discussing its regulatory roles, its connection to schizophrenia symptoms, and its potential as a therapeutic target. The subsequent subsection, 1.1.3, emphasizes the prominence of JNK, highlighting its precise regulation, conservation across species, and its significance in physiological and pathophysiological mechanisms. The entire section collectively paints an intricate picture of JNK's diverse functions, regulatory complexity, and potential implications in neurological disorders and pharmaceutical targeting.

1.3.1 c-Jun N-terminal kinase (JNK) introduction

c-Jun N-terminal kinases (JNKs) belong to the mitogen-activated protein kinase (MAPK) family and is one of the six sub-families that included JNKs, extracellular signal regulated kinase (ERKs) 1 and 2, ERK 3 and 4, ERK5 and BMK1, ERK 7 and 8, and p38 MAPKs [57]. JNKs were originally discovered in the mouse liver in an experiment where the mouse liver was treated with cycloheximide, which instigated inflammation and cell death [58]. JNKs were initially categorized as the stress-activated protein kinases (SAPKs) but was later renamed to JNKs due to the well-known function of phosphorylating c-Jun [57]. Figure 7 illustrates the structure and splice isoforms of JNK.

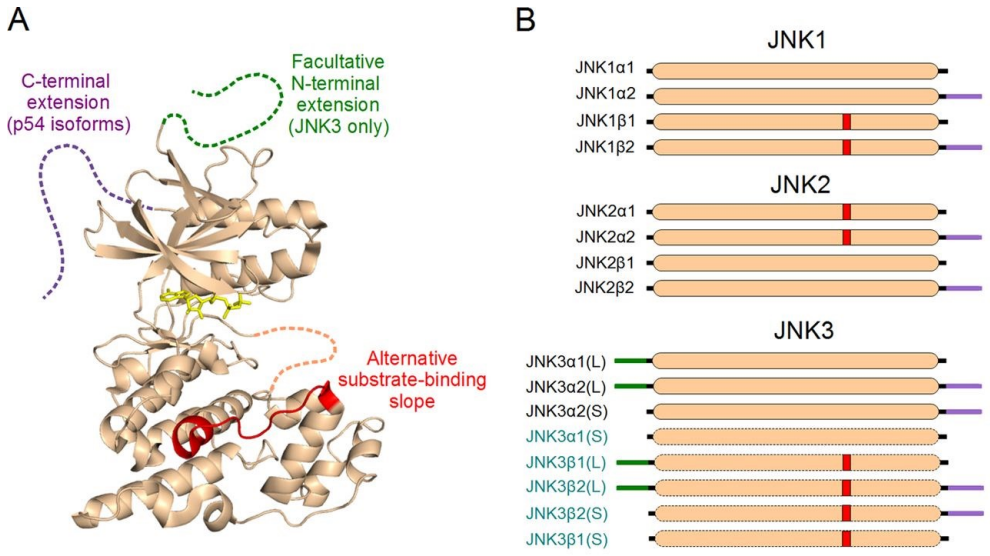

Figure 7. Part A depicts the crystal structure of JNK, with the common structure shown in beige, while variable regions from different splice isoforms are depicted in green, red and purple. Part B shows the different splice isoforms of a human JNK gene [59].

JNKs are components of the signaling transduction cascade known as the "three-tiered" MAPK pathways (Figure 8). At the top tier, MAP3Ks can be activated via interactions with e.g. small GTP-binding proteins. Activated MAP3Ks in turn phosphorylate and activate MAP2Ks in the middle tier of the cascade. MAP2Ks then phosphorylate MAPKs, making them active and ready to interact with downstream substrates [60]. The JNK signaling pathway can be activated in response to a variety of extracellular and intracellular stimuli such as pathogens, inflammation, oxidative stress, DNA damage or cytoskeletal changes, this activation serves as the downstream signaling cascade of receptors including G-protein coupled receptors (GPCRs), Wnt receptors, tumor necrosis factor (TNF) receptors, and Toll receptors [59].

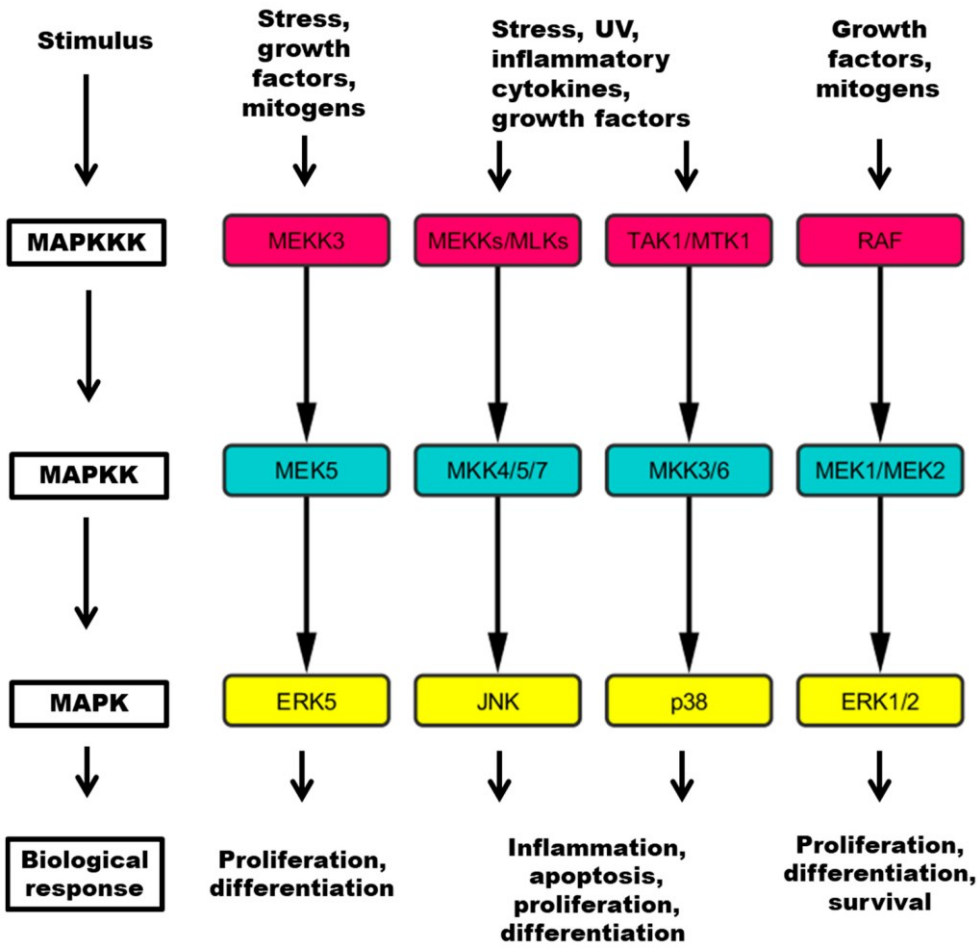

Figure 8. An illustration of the "three-tiered" MAPK pathways and proteins involved at each layer of the signaling pathway [61].

When JNKs are phosphorylated by the MAP2Ks, two phosphorylation events take place typically within the Thr-x-Tyr motif in its activation loops, where a confirmational change occurs and realigns the N- and C-terminal domains to create a functional active site [59]. Once JNKs are activated, they are translocated from the cytoplasm to the nucleus, where JNKs phosphorylate their substrates by interacting and forming a ternary complex with the substrate and catalyzing the transfer of the γ-phosphate from ATP [62] (Figure 9). The number of validated substrates of JNKs are close to 100 to date. The most iconic is c-Jun, where JNK phosphorylates the N-terminal Ser 63 and 73 positions to activate, and stabilize according to some research, the transcriptional activities of c-Jun [63], [64].

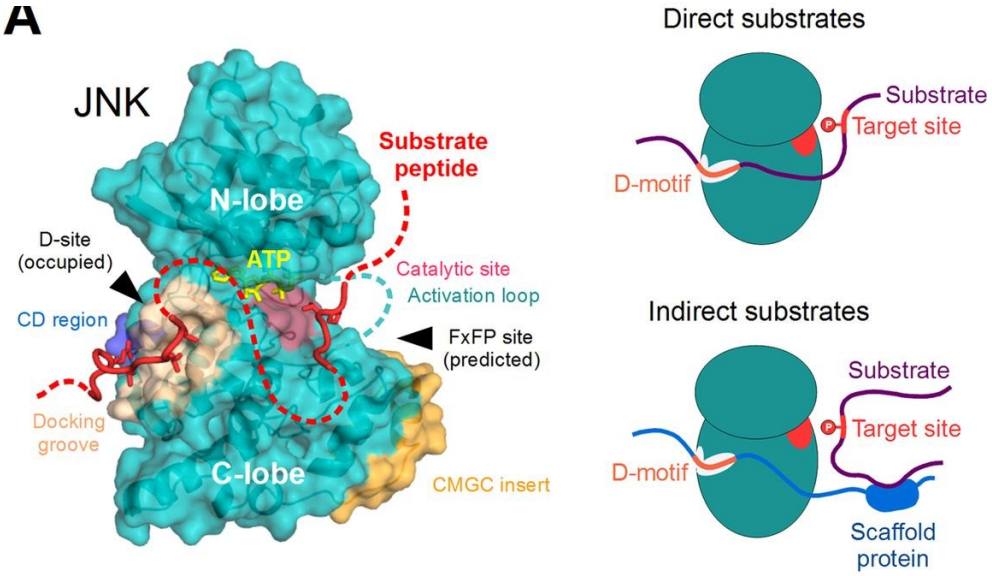

Figure 9. An illustration of the JNK substrate docking site. The CD region and docking groove form the major docking site (D-site) of JNK proteins, which are crucial for substrate recruitment. Both direct and indirect substrates bind to the D-site. Direct substrates (depicted on the top) contain a linear motif that enables them to bind directly to the D-site, while indirect substrates (depicted on the bottom) engage in heterologous interactions with a third protein that has the necessary motif for their recruitment to the D-site [59].

1.3.2 JNK and schizophrenia

Sensitive to a substantial number of stress stimuli, and with multiple substrates means that JNK has a multitude of roles, including regulatory neuronal functions, immunological actions and more. Research from our own lab showed that one of these roles is in regulating dendrite arborization in neurons through phosphorylation

of high molecular weight forms of microtubule-associated protein 2 (HMW-MAP2). The resulting grey matter loss, synapse regression coupled with dendrite reduction, and motor deficits are signatures of schizophrenia, all of which were consistent with our findings in *Jnk1-/-* mice [65].

A study from Openshaw et al. have linked JNK to schizophrenia through MAP2K7, one of the regulators of JNK. Schizophrenia patients are reported to have reduced level of MAP2K7 transcripts, and Openshaw et al. explored the relationship between MAP2K7 and schizophrenia using Map2k7+/- mice and ketamine and dextroamphetamine (D-amphetamine), which are drugs that induce schizophrenia-like symptoms. They have concluded that both brain imaging endophenotypes and behavioral phenotypes of Map2k7-/- mice resembled those of schizophrenia [66].

Furthermore, a review from Ansarey has linked JNK to the Niacin skin flush test for schizophrenia. Niacin (vitamin B3) exposure results in skin flush response in healthy population, whereas in most of the schizophrenia population, this response is diminished. The Niacin skin flush test could be utilized to distinguish schizophrenia patients from patients of other disorders such as depression or bipolar disorder at a prodromal stage. According to the review, several factors contributed to the altered skin flush response in schizophrenia patients. The protein expression levels in the GPR109A-COX-prostaglandin pathways are altered along with their receptors and downstream products. An inflammatory imbalance could also contribute to the altered response, which could be caused by environmental factors such as oxidative stress, which in turn reduces receptor bonding by changing receptor confirmations. It is likely both microglia and neurons were involved and affected. JNK regulates neuronal apoptosis, and interacts with M1, NF-κB, IL-1B, TNF-α, cPLA2, COX-2, and PPAR-γ, all of which were components of the mechanisms discussed in the review that altered the skin flush response. Hence JNK was recommended as a suitable therapeutic target for schizophrenia [67].

Schizophrenia is one of the top 25 leading disabilities with one percent of the global population suffering from it. The World Health Organization has estimated a spending of 94 million to 102 billion dollars on this disorder [68], [69]. A distribution of the cost of schizophrenia in the U.S. in 2019 is shown in Figure 10. Currently schizophrenia is diagnosed by the onset symptoms including positive symptoms such as hallucination and disoriented thoughts, negative symptoms such as apathy and social withdrawal, and cognitive symptoms such as impaired memory [67]. Since schizophrenia is a heterogeneous psychiatric disorder, the exact mechanism leading to its development is yet to be fully understood [70]. A gene-wide association study (GWAS) published in Nature journal have pinpointed 108 genetic hits that are closely associated with schizophrenia, many of which overlap with immune-related genes [71]. This matches with one of JNK's regulatory functions [72]. As much evidence has connected JNK to schizophrenia, we decided to explore the matter

further by performing a shotgun mass spectrometry (MS) analysis on *Jnk1-/-* mice brain over four different age groups and investigate phosphorylation and mechanistic changes as well as comparing them to known schizophrenia related genes and symptoms.

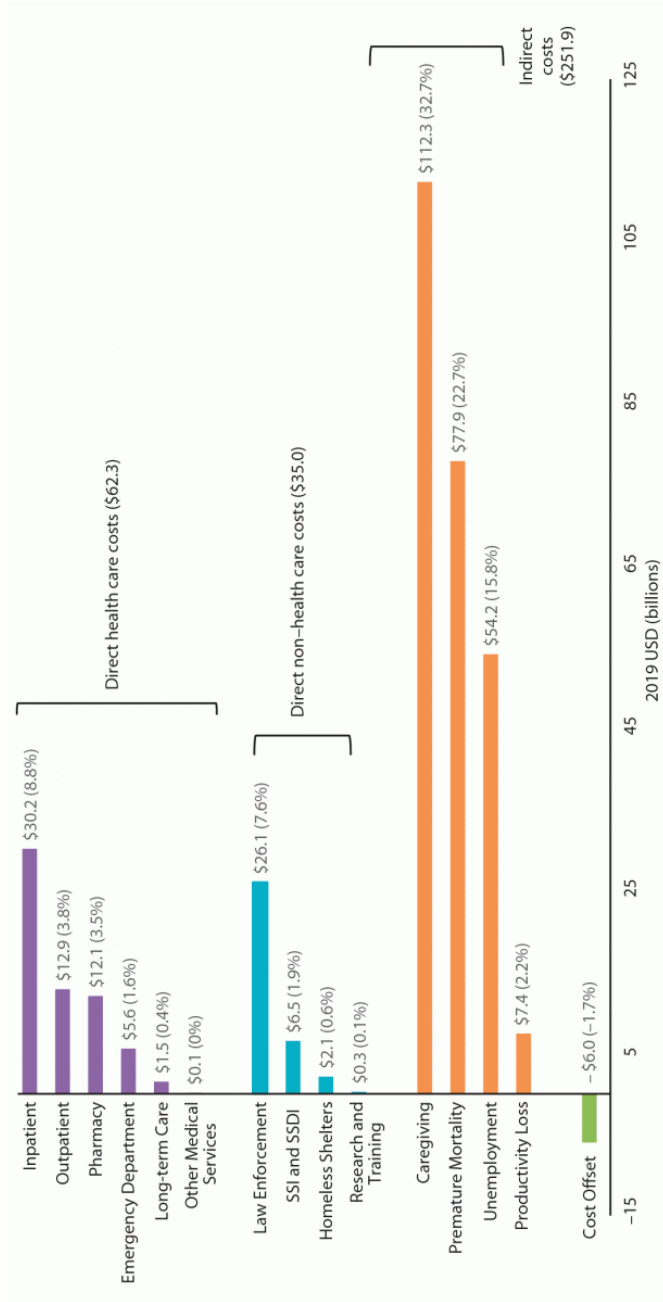

Figure 10. Distribution of excess total costs of schizophrenia in the United States in 2019 [73]. *Kadakia A, Catillon M, Fan Q, Williams GR, Marden JR, Anderson A, Kirson N, Dembek C. The Economic Burden of Schizophrenia in the United States, The Journal of Clinical Psychiatry. Vol. 83(6), page 22 m14458, 2022. Copyright 2023, Physicians Postgraduate Press. Reprinted by permission.*

1.3.3 The prominence of JNK

There are many reasons why JNK has been studied extensively. JNK pathways can remain inactivated even in the presence of stimuli, which meant it is precisely regulated not only by kinases, but also by phosphatases [74]. The precision of JNK pathway regulation does not end there, a study from Bhalla et al. have shown the phosphatases control the signal flux of JNK pathway as well, by changing the expression level of phosphatases, JNK signals can flexibly respond to stimuli in calculated proportions [75]. In addition to the complex regulation scheme, JNK pathway is highly conserved in all eukaryotes, from yeast to human [76]. Conservation among different species and precision in regulation both affirm the importance of JNK pathway for the physiological and pathophysiological mechanisms.

JNK signaling has been extensively studied for over 20 years, where numerous stimuli have been found to be associated with JNK regulation, and close to 100 substrates have been identified, yet many structural and mechanistic insight have only begun to be uncovered [59]. Already it has attracted attention as a potential pharmaceutical target, and it has successfully captured our attention through its activeness in neurological disorders. There are three JNK genes in the human genome: JNK1, JNK2 and JNK3. Structural wise, JNK1 and JNK3 are more similar to each other than JNK2, with JNK3 having an extra N-terminal extension compared to JNK1 [77]. Functionally, however, JNK1 and JNK2 are more similar with many overlapping functions. This is supported by knockout experiments, where *Jnk1/Jnk2* double knockouts are embryonic lethal, while *Jnk1/Jnk3* double knockouts and *Jnk2/Jnk3* double knockouts are feasible [78]. Even though JNK1 and JNK2 have some overlapping functionalities, differences in cellular regulation between the two can be distinguished from comparing *Jnk1-/-* and *Jnk2-/-* mice. *Jnk1-/-* mice showed abnormalities in brain development and metabolic regulations, while *Jnk2-/-* showed only mild phenotype changes including epidermal hyperplasia and moderate immune disturbance [79], [80]. In addition, neurogenesis in vitro primarily requires JNK1 but not JNK2 or JNK3 [81]. Since we are interested in brain functions controlled by JNK, we focused our study on JNK1, which is the physiologically active JNK isoform.

1.4 Parkinson's disease

Parkinson's Disease is the second most common motor disorder after Alzheimer disease [82]. It was first mentioned by James Parkinson, a general practitioner in London, in 1817 in an essay that described it as an involuntary tremulous motion [82], [83]. It was estimated that 1.5 million people suffers from Parkinson's disease

in US alone and is affecting 1-2% of the entire world population [84]. The distribution of Parkinson's disease in different global regions is displayed in Figure 11. Parkinson's disease is mainly caused by the progressive loss of dopaminergic neurons in the substantia nigra of the middle brain, which leads to alterations in downstream basal ganglia circuitry [82]. Symptoms of Parkinson's disease consist of both motor and non-motor types. Motor symptoms include bradykinesia, resting tremor, rigidity, and postural instability, while non-motor symptoms include anxiety, depression, fatigue, and sleep disorders [84].

Currently, the diagnosis of Parkinson's disease poses a major challenge for clinicians and scientists [84], [85]. The state-of-the-art diagnosis is symptom-based assessment referred to as the Unified Parkinson's Disease Rating Scale (UPDRS). This often results in late detection of the disease [85]. In addition, it can sometimes be underdiagnosed or misdiagnosed due to drugs, Wilson's disease, and other similar neurological disorders manifesting seemingly identical symptoms [84]. To date there is no good biomarker with high enough sensitivity and specificity for the diagnosis of Parkinson's disease [84], [85]. As such, we have gathered Parkinson's patient samples from the Nordic area, and performed MS analysis on the samples in hope of learning more about Parkinson's disease and discover the biomarkers that are critically important for improving diagnostic strategies.

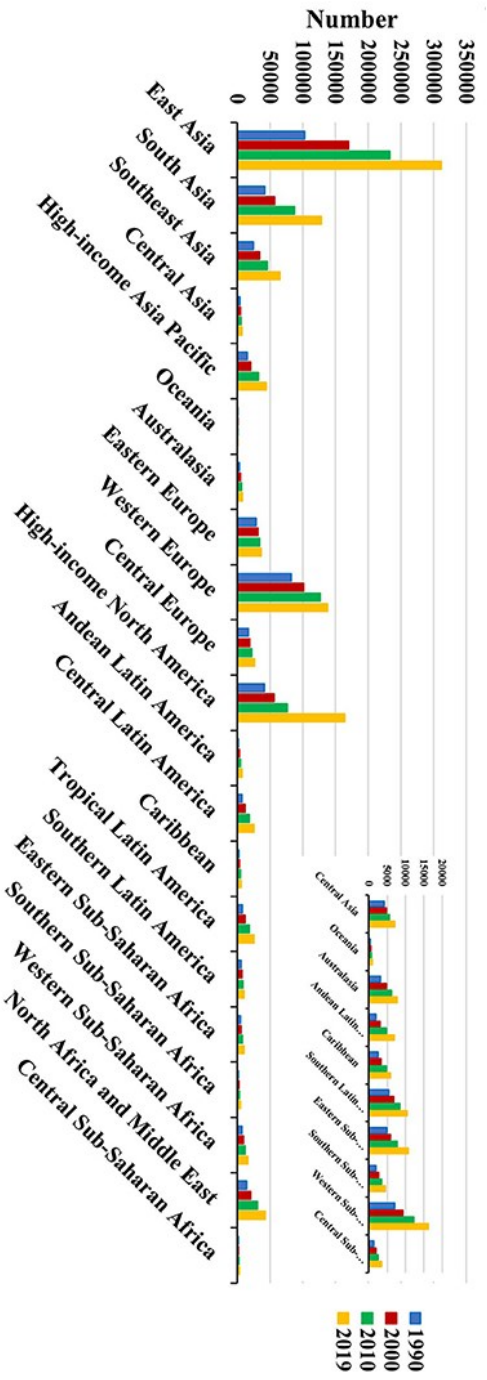

Figure 11. The distribution of Parkinson's disease cases by geographical regions from 1990 to 2019 [86].

1.5 Mass spectrometry technology

This section delves into the technological aspects of mass spectrometry, a powerful tool for the detection and quantification of proteins, with a specific emphasis on shotgun proteomics workflow.

1.5.1 Mass spectrometry and shotgun proteomics workflow

Mass spectrometry methodology enables the detection and quantification of thousands of proteins from multiple samples, especially in the last few decades, rapid development of the technology allows for the routine study of proteomics and post-translational modifications such as phosphorylation. Mass spectrometry-based proteomics branches into top-down and bottom-up approaches, for our analysis, bottom-up proteomics, or shotgun proteomics, where proteins are digested into peptides before being analyzed with a mass spectrometer, were utilized [87].

The typical workflow for shotgun proteomics (Figure 12) involves 1) digesting protein samples from cell or tissue lysate with proteases such as trypsin and cleaving the protein into peptides at specific positions; 2) fractionating the mixture into multiple portions based on parameters such as charge, size, or polarity; 3) separating each portion with liquid chromatography; and 4) running the eluted peptides through the mass spectrometer [88]. Before entering the mass spectrometer, the concentrated positively charged peptide droplet travels through a voltage area where it breaks surface tension with coulombic repulsion and explodes into the gas phase in a process called electrospray ionization (ESI). The ionized peptides then enter the mass spectrometer, where they can be detected or filtered based on their mass-to-charge (m/z) ratio. The read out from this detection is called MS1 spectrum, the height of the signals correlate to the number of detected ions for the peptide. The peptide ions can be further fragmented by colliding with inert gases, the result readout is called MS2 or MS/MS spectrum. The MS1 and M2 readouts can identify the amino acid sequence and post-translational modifications associated with the peptide when compared to the theoretical spectra of possible peptides and assigned the identity of the best matching peptide [88]. The detected peptides can also be quantified; however, this quantification is relative rather than absolute. The ionization efficiency can differ considerably for different peptides, therefore the number of ions formed does not reflect the number of proteins in the original sample. However, the same peptide (in different samples) can be comparable due to having the same ionization efficiency. Absolute quantification is achievable with added spiked-in as control, with known concentration [88].

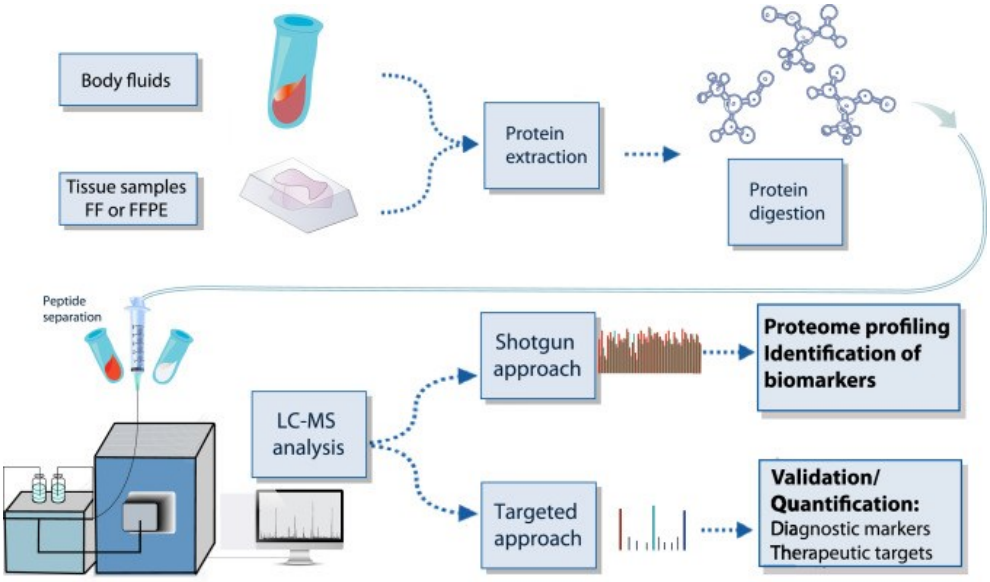

Figure 12. An illustration of the MS workflow. Reprinted from [89] with permission from Elsevier.

1.5.2 An extra step in phosphorylation site identification and quantification

Studies which focus on phosphoproteomics data usually go through the shotgun workflow with an additional enrichment step before running through liquid chromatography. In an equilibrium state, there are much less phosphorylated sites than their unphosphorylated counterparts, and phosphorylation site identification and quantification could easily suffer as a result due to the undersampling effect. The enrichment step is designed to extract and isolate the phosphorylated peptides and increase its concentration before going through the mass spectrometry analysis [90]. There are quite a few enrichment methods, such as strong cation exchange chromatography (SCX), immobilized metal ion affinity chromatography (IMAC), and titanium dioxide affinity purification (TiO_2). We have employed TiO_2 enrichment method for our phosphoproteomics MS analysis. The general protocol includes binding peptides to the TiO2 beads, removing unphosphorylated peptides by washing in glycolic acid solution and 50% acetonitrile (ACN), eluting the phosphopeptides with NH4OH, then acidifying and drying before running through the liquid chromatography [90].

1.5.3 Data-independent-acquisition (DIA) method alleviates the missing value issue for label-free approach

This typical workflow described for the shotgun proteomics in section 1.4.1, known as label-free approach, has a few drawbacks. The method requires multiple runs, and running samples separately results in poor reproducibility when MS1 and MS2 spectra are obtained separately. The median protein coefficients of variation (CVs) between replicates are somewhere around 20%, and worse with less abundant peptides. In addition, a portion of the peptides are not detected in every sample due to undersampling, even for replicates, and this results in the missing value problem [88]. An implementation that alleviates the missing value problem is the data-independent-acquisition (DIA) method. The typical label-free proteomics adapts a common feature called data-dependent-acquisition (DDA), where the instrument chooses the largest signals from MS1 spectrum for MS2 spectra acquisition and peptide identification. The signals chosen tend to differ from run to run, which contributes to the replicate variance and missing values. DIA label-free proteomics, however, collects MS2 spectra continuously, and covering the entirety of MS1 spectrum. This coverage advantage greatly lessens the missing value problem compared to DDA approach [96]. Figure 13 demonstrates the methodology difference between DIA and DDA.

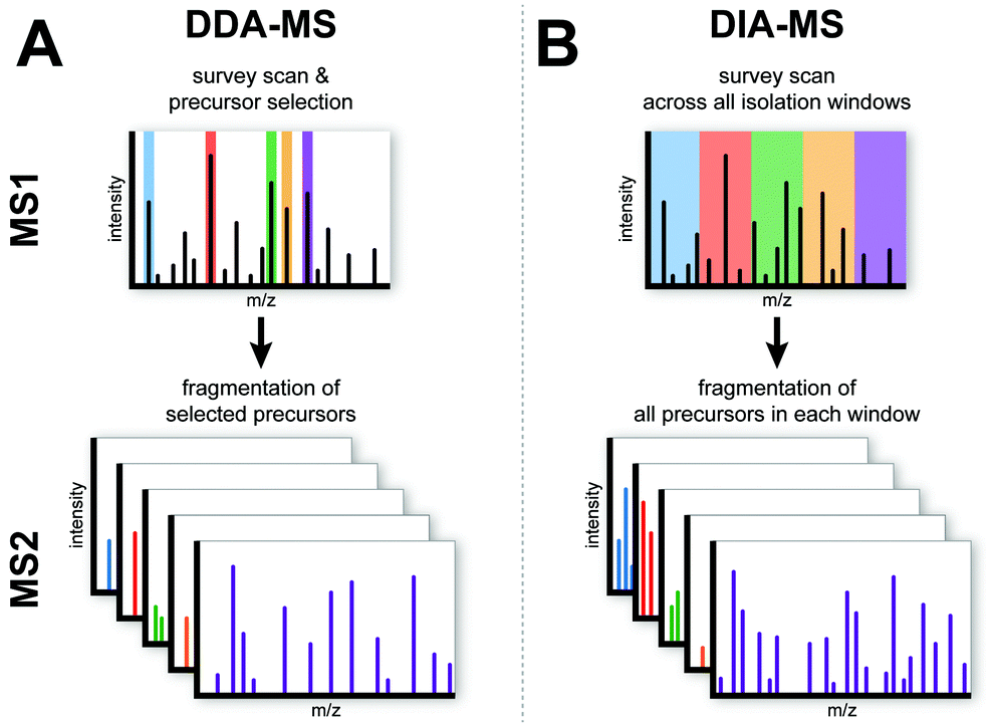

Figure 13. DIA vs DDA methology overview. In DDA, MS1 survey scan picks up the n most abundant precursor ions and subject them to fragmentation in MS2. In contrast, DIA employs a predefined wide isolation window to select all precursor ions within the m/z range of the window in MS1, and fragments all precuror ions within each isolation window in MS2. Figure reproduced from [91] with permission from the Royal Society of Chemistry.

1.5.4 Other approaches in mass spectrometry

Aside from shotgun discovery proteomics, other MS variants exist. Multiplexed proteomics with isobaric labeling is a popular approach, samples can all be run at the same time instead of one by one with this approach, hence it improves reproducibility between samples and prevents missing at random (MAR) and missing completely at random (MCAR) missing values [88]. MS for specifically targeted proteins is another variant. For example, L-Azidohomoalanine (AHA) labeling method followed by enrichment, is used in this thesis to extract newly synthesized proteins [92]. When treated with AHA, cultured cells incorporate it into proteins during active protein synthesis, and a click reaction between an azide from AHA and an alkyne from alkyne-tagged biotin enables enrichment of azido modified proteins specifically. Essentially MS analysis of this detects newly synthesized proteins [93]. Another example of targeted approach is parallel reaction monitoring (PRM), which can validate results from a proteomic shotgun analysis. It uses prior

information to target specific peptides in the sample for high resolution quantification [94].

1.5.5 Peptide identification and quantification

Mass spectrometry generates raw spectral data which undergoes subsequent processing, typically via a quantitative proteomics software, for peptide and protein identification and quantification. Peptide identification involves matching the peptide precursor mass-to-charge ratio and its fragment ions to known peptide sequences from comprehensive protein databases utilizing search algorithms such as Mascot or SEQUEST, while quantification involves tallying the number of spectra corresponding to each identified peptide sequence [95].

1.6 Downstream bioinformatics analysis of MS data

After the identification and quantification of proteomics and phosphoproteomics data, downstream bioinformatics analysis can be performed. Bioinformatics analysis refers to analysis performed with the aid of computational software on large biological datasets. The goal of the analysis is to find useful patterns from the expression level or phosphorylation level of the identified proteins, to improve mechanistic understanding of the disease, treatment, mutation, and any other topic the data entails. Such knowledge would find application in improving the current clinical and therapeutic technology threshold.

1.6.1 Quality control

Quality control is crucial in providing an unbiased research space to study the data that was produced. Quality checks are already implemented in the spectral identification software to minimize errors in protein discovery and intensity calibration. For example, MaxQuant utilizes a target-decoy search strategy to control false identity discovery. The concept of posterior error probability (PEP) is employed in the target-decoy strategy, where peptide properties such as charge, and number of modifications are considered to assess the quality of a peptide spectrum match (PSM). In addition, FDR calculations are implemented at protein group and PTM site level to further control the quality of identified peptide or PTMs. Furthermore, “match between runs” option is provided in case there is no sufficient information in one run to identify/quantify some sequences; and normalization option is available to reduce individual fraction bias introduced by fractionation step

[96]. However, between spectral software output and the start of sample analysis, further quality control takes place with a specific focus of removing or improving low quality entries. To start off, entries marked as potential contaminants should be checked manually before removing true contamination entries. For example, keratin is usually marked as contaminant, for skin samples, however, it could be a target of interest and naturally occurring compound from the samples. Based on sample-specific biological evidence, entries falsely marked as contaminants should be removed from the contamination list. Entries marked as reverse sequences should be removed. It is also customary to only accept protein entries with more than 1 peptide identification.

Data processing is part of the quality control procedure where the numerical data is examined, low quality entries are identified, and decisions are made for these low-quality entries to reduce biases that could contribute to the overall analysis. Depending on the technical protocols carried out to produce the data, several, or all, of the following steps can be employed in data processing, they are filtering, normalization, imputation, and batch correction.

1.6.1.1 Filtering

When dealing with low quality data, one option is to remove these data from the analysis all together to guarantee the integrity of the analysis conclusion, making certain it is drawn only from high quality data. That is the exact role of the filtering step. Of course, maintaining a balance between the completeness of the data and the quality of the data is very important, hence filtering thresholds should be set carefully. Only suspected contaminations or gross errors, whether coming from biological, technical, or human sources, should be filtered out. When a subset of data displays concentrated outlier values on either tail of the data distribution, it usually signals contamination or an error, and is indicative of low-quality data entries.

Filtering can be done on both the sample level (column) or the peptide level (row). For sample filtering, the easiest way to spot problematic samples is to plot data overview figures. Outlier samples can be easily spotted with boxplot, heatmap, or PCA plot (Figure 14). For this very reason, filtering should be done prior to normalization or imputation steps, outliers can possibly be “corrected” and difficult to catch when they are processed with either step. For peptide filtering, overview figures are not as helpful, given the large quantity of entries involved. In this case, a good indication of peptide quality is the proportion of missing values included in its intensity distribution. High number of missing values in a single peptide entry is usually indicative of low data quality for this entry. A missing value count threshold can be installed to pick out the low-quality peptides. Such threshold should be customized considering the nature of the data source, the focus of the study, and the

technical procedures which produced the data. For example, a dataset with 2 genotypes and 2 treatment groups could have 4 thresholds per peptide, one for each unique group combination. The numerical value of each threshold can be set with reference to the median and standard deviation of the missing value distribution from the corresponding group. This would remove peptides with excess NA values and assure the data quality of all groups for each peptide that survived filtering.

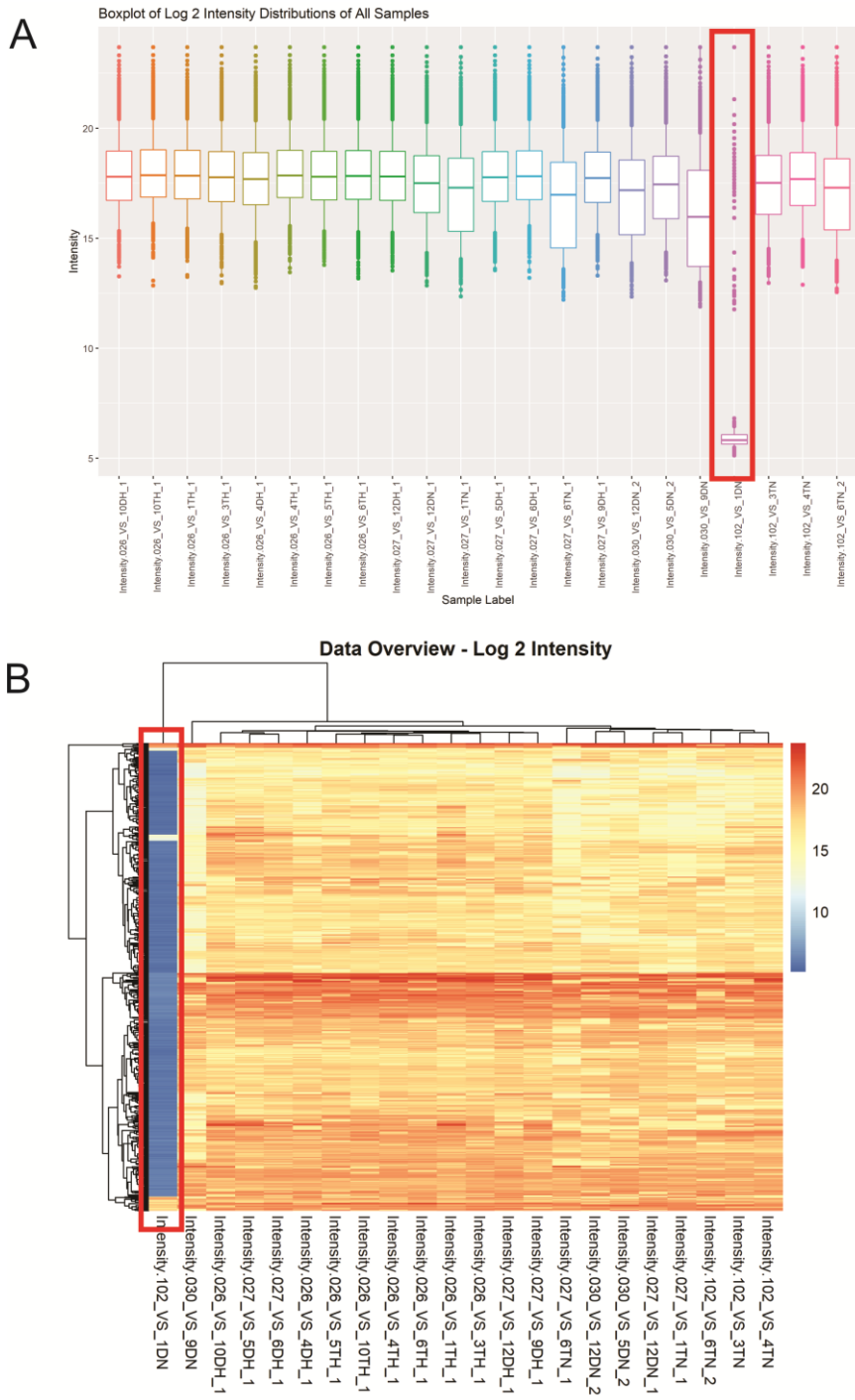

Figure 14. Example of an outlier sample spotted in boxplot (A) and heatmap (B). The outlier sample is enclosed by red rectangles. Represented in the figures are samples from healthy and JNK knock down mouse brain log 2 MS intensity data.

1.6.1.2 Normalization

The goal of normalization is to reduce variation between technical or biological replicates [97]. Small increments of variations are introduced during the course of MS workflows, they accumulate to be significant enough by the end of the MS intensity quantification, that normalization is usually required to adjust this variation bias between samples, even for high quality data. Normalization is different from batch correction, where it is preferred even when all the samples are processed in the same batch throughout the workflow. This is to correct any spontaneous variations contributed by the accepted error range of each machine and methods.

The initial normalization methods for MS generated data are based on methods developed originally for DNA microarray technology [98]. For example, cyclic loess method and quantile normalization was originally used on microarray data [99]. Later methods emerged which would take into account MS specific steps, such as phospho-peptide enrichment, for the formulation of the normalization concept. For example, Kauko et al. utilizes normalization which takes into account the phospho-peptide abundance before and after the enrichment step, to address the major source of variation introduced by the MS specific step of TiO2, and to accommodate global phosphorylation alterations [100] This approach is available as an R package, Phosphonormalizer, which performs pairwise normalization using non-enriched phosphopeptide as references to scale the final phosphopeptide intensities [101].

One of the challenges of proteomics is to decide on a normalization method. Välikangas et al. have conducted a study to compare the different normalization methods available for MS generated data (Figure 15). 11 different methods were tested on three spike-in datasets and one mouse proteomics dataset, namely, log 2, fast loess, cyclic loess, linear regressions Rlr, RlrMA and RlrMA clyclic, variance stabilization normalization (Vsn), quantile, median, Progenesis provided normalization, and EigenMS normalization. In the end, they have concluded that Vsn performs the best in terms of reducing variation between technical replicates, and consistently maintains low error rates in differential expression analysis. Fast loess, Rlr and RlrMA also performed well in differential expression analysis [97]. However, to pick the most suitable normalization method, the nature of the data should be carefully considered before making a decision. For example, if the non-enriched peptides have wildly different identities from the enriched phospho-peptides (this happens quite often from experience), phosphonormalizer should not be utilized when there are too few matched pairs, this would promote inaccurate scaling predictions for the normalization [101].

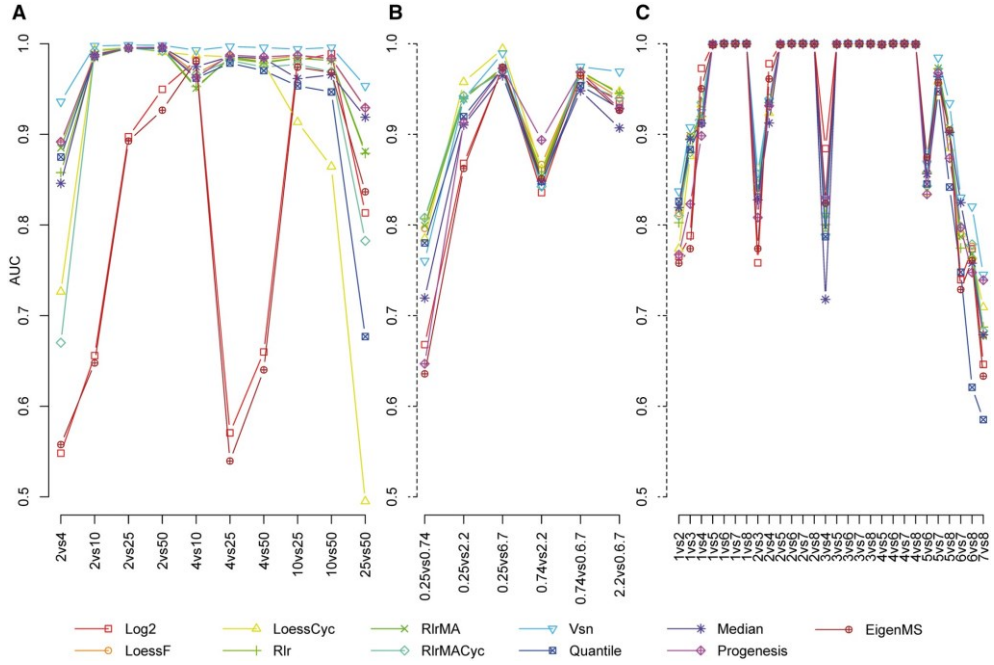

Figure 15. Performance of 11 normalization methods on 3 separate datasets is shown in the area under the curve (AUC) plots. The performance metric is calculated from the ROC curves of differential expression analyses in (A) UPS1 data, (B) CPTAC data and (C) SGSD data after applying the test normalization method. This figure is reproduced from the study result of Välikangas et al [97].

1.6.1.3 Batch Correction

A batch effect is a systematic difference between data due to technical or environmental factors [102]. The intensity measurements derived from mass spectrometry and spectral interpretation software can be affected by e.g. the length of incubation time, handler change, reagent batch or instrument differences between experiments. The variances introduced by such technical variables create a batch effect. This is problematic as it can mask real biological significance within the data [103]. Batch effect lowers the quality of the data for all the follow-up data analyses, hence batch correction algorithms have subsequently been developed to solve this issue. Especially in recent years, due to technical advancement in the ability to handle large proteomic datasets [104]–[106], the issue of batch effect magnifies as it is very difficult to carry out experiment protocols on all of the samples at the same time.

It is important to note that normalization and batch correction are two separate steps. Normalization adjusts samples to bring them to a comparable scale, however, it is on a global scale, i.e. it is applied to the entire dataset [102]. Normalization does not correct for feature specific batch effects, in fact, L. Zhou et al. has demonstrated

with quantile normalization algorithm, that batch effect is not removed from the calculated result, on the contrary, it contributes to the ranked means after quantile normalization [107]. Batch correction algorithms, on the other hand, aim specifically to reduce variance associated with technical and environmental factors for each feature across all samples [108].

Batch correction starts with an initial assessment step to evaluate the severity of the batch effect, the nature of the affected data, and possible sources for the batch effect. Overview figures are a good way to detect batch effects. For example, PCA plots with each plot using separate colors for different batches, or sample correlation plots with the same color scheme. Determining the nature of the data helps selecting a normalization method. Batch correction algorithms are usually coupled by normalization to set all samples to the same scale. For example, if the total amount of material in the samples are similar to each other, then quantile normalization can be used [109]. In samples where the total amount of proteins should vary significantly, a different normalization may take place, quantile normalization in this case would introduce errors by enforcing quantile-centering for all the samples. Batch correction is performed after normalization and will benefit from a suitable normalization method.

Batch effect can be continuous or discrete. Continuous effect could signify a signal drift (Figure 16), which often occurs in large sample size data. This drift could be corrected by fitting a curve to the data, such as LOESS fit [108]. Discrete effect shifts samples from each batch more uniformly. In this case, mean and median centering algorithms should be utilized, such as ComBat, which is a modified mean centering method where empricial Bayes framework is employed to estimate batch effect parameters [110].

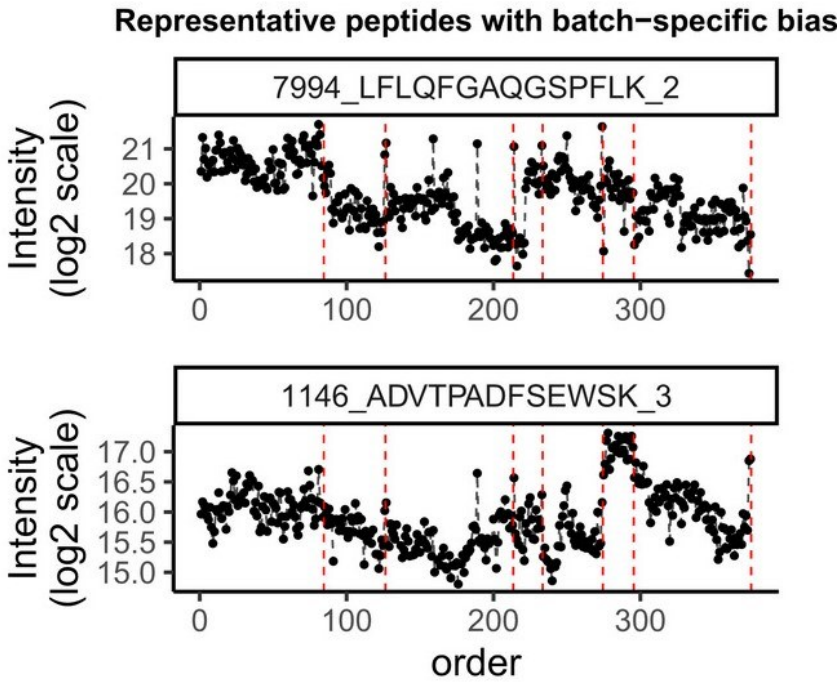

Figure 16. Example of signal drift batch effect [108]. Two peptide entries are plotted where the log2 intensity values demonstrate MS signal drift batch effect that requires correction.

The last step of batch correction is quality control, where the resulting data is inspected for batch improvement, and for downstream effect such as how differential expression significance is altered by the correction. Before going into quality control, there are several factors that could dramatically affect the performance of the batch correction algorithms. One of them is missing values. One of the batch effect manifestations is the different number of missing values generated in different batches. ComBat, for example, has no tolerance for missing values, and cannot correct for the feature if any single batch contains missing values. A common practice to rid of missing values is imputation. However, imputation algorithms can introduce batch or peptide specific biases to the data, and can disrupt batch adjustment, resulting in seemingly higher correlation within batches, and lower correlation between replicates for batch corrected data. For this reason, imputation should be avoided, or carried out after batch correction, if batch correction is in order [108].

Another factor that affects the performance of batch correction is the confounding effect. The confounding effect describes the mix up of sample group and batch effect, from balanced, to indistinguishable. For example, a balanced

dataset would have groups A, B and C, equally distributed between batch one and batch two, each having 50% of samples from A, from B, and from C. An indistinguishable dataset would have all samples from A in batch one, and all samples from B in batch two. It is next to impossible to determine whether the variance derives from batch effects, or group effect in the case of indistinguishable datasets, and this greatly reduces the performance of the batch correction algorithms [107]. Planning the experimental design to produce an optimally balanced dataset is advised. The effect of different normalization algorithms is not one of the factors that influence batch correction performance, so it can be chosen solely based on the suitability with the data [107].

To actually evaluate the batch correction performance is rather difficult without simulated or spiked data. One method would be to check differentially expressed features separately for separate batches, where high overlap would suggest good performance [111]. This method works well with larger dataset, as smaller dataset suffers from lower predictive power, and is relatively unstable. If technical or biological repeats exist across batches, correlating these repeats would give a good indication of the batch correction performance [108]. The correlation is expected to increase compared to the dataset before correction. Since batch correction supposedly reduces variance, the downstream differential expression statistics is inevitably affected. However, a good performance does not guarantee improvement for the statistical analysis. L. Zhou et al. have evaluated several batch correction algorithms to determine their performance, as well as their influence on statistical analysis. It was reported for severely unbalanced datasets, that the SVA algorithm emerges as the best all-rounder [107]. Detailed results can be found from their study. Performing batch correction can be tricky, however, it can become a worthwhile step in the analysis for the right dataset, with the potential to greatly improve the quality of the data.

1.6.1.4 Missing Value (“Not Available” or “NA”)

The missing value problem in proteomics and phosphoproteomics is much more problematic than in e.g. microarray based studies. For certain global proteomics approaches, it is common to have missing values take up 50% of the entire dataset [112]. This proves to be a major difficulty in all downstream analyses, including, but not limited to, unsupervised clustering, functional inference, supervised machine learning, and interaction network prediction [113], [114]. For this reason, missing values must be dealt with in almost all proteomics/phosphoproteomics studies.

There are 3 common ways to handle missing values. The first is to completely filter out rows of data with missing values or leave only 5%-10% of missing values in the dataset. Another way is to employ analysis algorithms which are lenient on

missing value proportions. The last option is to impute the missing values based on either simple or sophisticated models of the conditions which contributes to missing data [112]. The strict filtering approach from the first option is less practical simply because of the sheer number of missing values. This could drastically reduce the data size, limiting the validity of the follow up analysis. The option to employ specific tools can also be restrictive, as the important functional analysis and network analysis usually don't tolerate missing values well. Hence the most popular option is to impute missing values. A comprehensive understanding of MS generated missing values is necessary in order to facilitate accurate predictions for each imputed value. However, it is not a simple task to unravel the complexity which contributes to the high percentage of missing values. Unlike with microarray data, where missing values comprise only five percent of the data, global proteomics data could contain 20% to 50% of missing values (Figure 17). Some of the reasons for a microarray missing value could be scratches or spotting issues [115]. For proteomics, a series of factors could cause missing data since numerous steps have taken place in a typical label-free liquid chromatography mass spectrometry (LC-MS). This includes sample-side factors such as low protein abundance, as well as experiment-side factors such as loss of sample during preparation steps, peptide mis-cleavage during digestion step, and poor ionization efficiency during MS run [116].

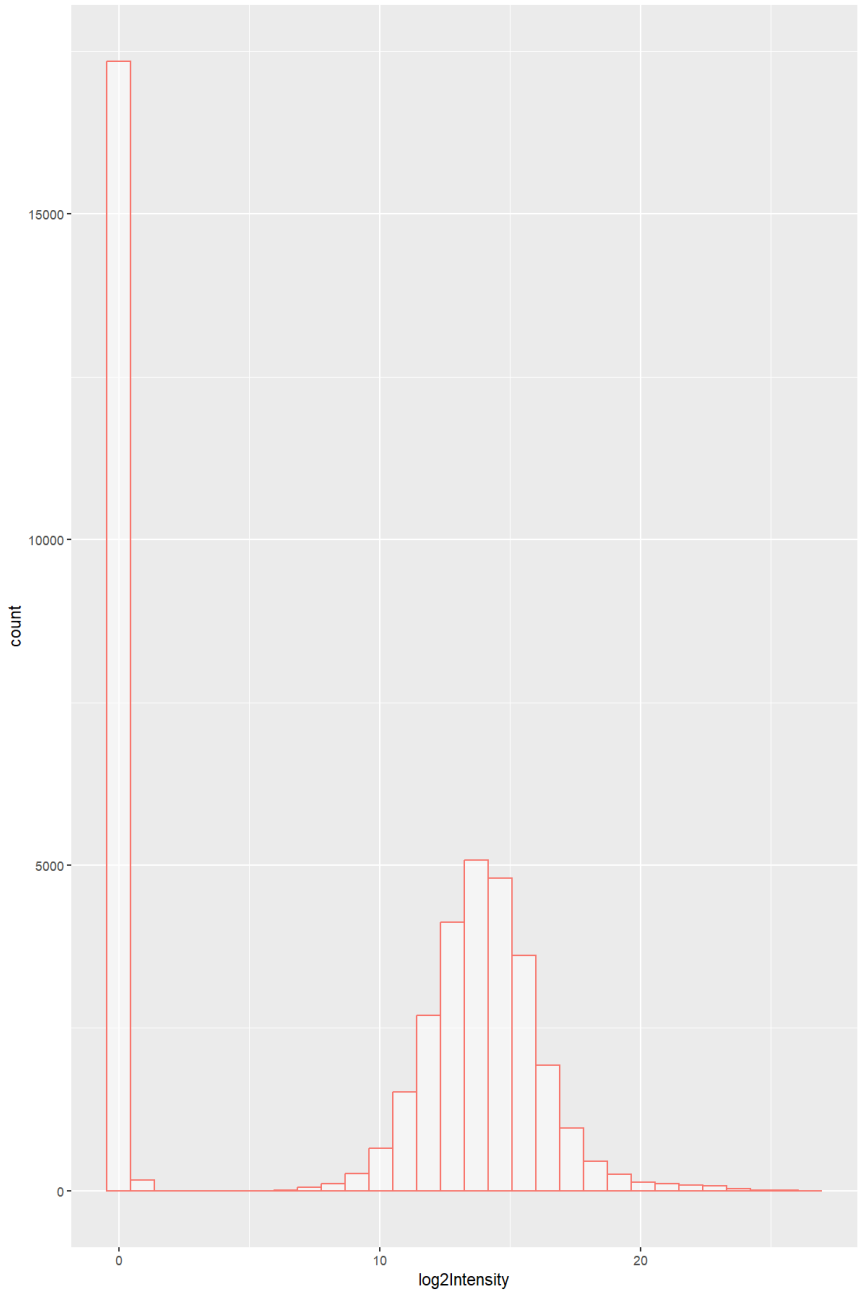

Figure 17. An example intensity distribution (in log2) of a typical 15-sample proteomic data in peptide entries. The missing value count is represented by the intensity bin at 0.

Missing values can be classified into 3 types, they are missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). if values

are completely at random for all entries in a dataset, then it is a case of MCAR. If values are missing at random only within a given group or property, the values are said to be MAR. If the missing value is neither these cases, and seems to be following a clear trend, then the missing value is MNAR [117]. Missing values in proteomics consists of both MAR and MNAR. Technical limitations and stochastic fluctuations are likely to cause MAR missing values in a compound abundance independent manner, and measurability and detectability of compounds could contribute to MNAR missing values in an abundance dependent manner [116].

1.6.1.5 Imputation

Imputation takes a missing value and assigns a numerical value to it based on inference from the rest of the dataset. There are many methods to choose from, depending on the nature of the algorithm, they can be classified into three categories [112] (Figure 18). The first category can be described as imputation by a single replacement value. The methods from this category replace a missing value with a constant or a sensible random value. This type of imputation can sometimes be found in microarray workflows, however, its performance on microarray data is much worse than some of the more complex algorithms, except in situations where the missing values are predominantly left censored, in these instances it performs rather well [118]. Data with left censored missing values have their missing values mostly concentrated on the left side tail of the data distribution, or the low intensity portion of the data. Such data can be assumed to have MNAR missing values [118]. One approach from this category is to estimate a numerical value which represents the limit of detection (LOD), and then assigning a value based on it. “Half of the global minimum” method and “half of the peptide minimum” method are considered LOD methods [119], [120]. Half of the global minimum adopts the minimal value of the entire data, whereas half of the peptide minimum adopts the lowest same-peptide intensity and takes half of this value for imputation. Random tail imputation (RTI) is also part of imputation by a single replacement value. The algorithm assumes the data forms a variant of normal distribution, and that missing values are left censored. Random values are drawn from the left tail of a truncated data distribution to fill the missing values, a limiting parameter is set to define the range where the values are drawn, so that the additional non-missing values will not spike a second peak into the normal distribution (bimodal distribution) [121], [122].

The second category of imputation methods is the local similarity approach. This approach assumes protein expressions are dependent on its interactions, and closely related proteins, either in function, regulation mechanism, or localization, can share similar expression patterns [123]. The approach exploits highly correlated protein expressions to interpolate the most appropriate value for each missing value. Two

steps are involved, the first is to select the most similar peptides, this is usually determined by similarity assessment algorithms such as distance formula or correlation formula. The second step is the actual imputation based on values of the combination of these close neighboring peptides [112]. One example of this approach is K nearest neighbors (KNN) [124], it uses Euclidean distance formula to determine 10 peptides with the most similar peak intensity profiles, also known as 10 nearest neighbors, and impute the missing value based on the nearest neighbors. In the event where all 10 neighbors also have missing values, the next 10 peptides with the closest distance would be used. Other methods belonging to this category includes the local least-squares imputation (LLS) [125], the least-squares adaptive imputation (LSA) [126], the regularized expectation maximization algorithm (REM) [127] and model-based imputation (MBI) [128].

The last imputation category is global-structure approaches. Imputation methods in this category utilizes expectation-maximization (EM) algorithm on dimension reduced data, where the maximum likelihood estimates (MLE) is determined in each iteration until the likelihood estimates does not improve anymore. One example is the probabilistic principal component analysis (PPCA) [129]. PPCA assumes the latent data points and noise are both normally distributed. The data is reduced using PCA algorithm based on non-missing data, then the reduced data and the missing values will be considered as model parameters in each iteration of the EM algorithm. Each iteration includes an expectation (E) step and a maximization (M) step, the expectation step imputes the missing values based on the model parameters, while the maximization step imposes a MLE algorithm on the imputed dataset from the expectation step and modifies the imputation parameters for the model. This process is repeated until likelihood estimates plateau [130]. msImpute method also belongs to this category. This method was selected as the default imputation method for PhosPiR pipeline from study I of this thesis. At that time of implementation, a most recent publication has indicated good performance of msImpute [131]. We made the implementation decision based on the resulting competitive performance of msImpute against other imputation methods such as Perseus-style imputation and K-Nearest Neighbors (KNN) according to Hediyeh-zadeh et al. Approaches in this imputation category can retain accurate imputations even for MNAR values [130], however, due to the computational complexity, this type of algorithm requires high processing power and is often time consuming [112].

Based on the approaches introduced above, it is apparent that making assumptions about the nature or distribution of the data is necessary to formulate the algorithmic parameters involved in each approach. These assumptions should be carefully studied to evaluate the compatibility of the algorithm and the data. In many cases, a specific type of normalization is required before carrying out the imputation process.

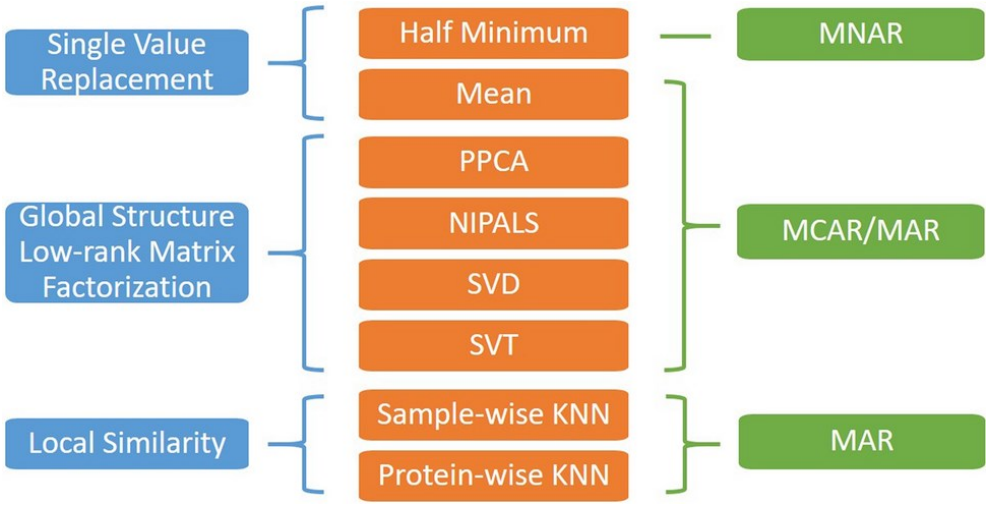

Figure 18. Illustration of the three categories of imputation methods, example methods from each category, and the missing value types for which each category is most appropriate [132].

1.6.2 Data analysis

This section introduces various analytical methodologies employed for the interpretation of mass spectrometry data.

1.6.2.1 Annotation method introduction

The information age provides a superior stage for the collaboration of individual research, and annotation databases supply the infrastructure to effectively combine knowledge from individual studies and publications and develop them into organized glossary of information that is shared between the science community, providing standard, transparency, and valuable background knowledge for new studies and novel algorithm developments.

A few examples of popular annotation databases include Ensembl, UniProt, and for human, the HUPO databases. Ensembl specializes in genome data annotation and integrates a variety of organisms data, linking them through orthology annotations [133]. UniProt specializes in protein and protein sequence annotation, with focus on known and predicted protein functional information [134]. HUPO comprises a collection of databases exclusively containing human-related biological information, it is one of the best resources for human studies [135]. While HUPO specializes in human data, many databases have their own specialized study areas. For example, Allen Brain Map [136] is extremely helpful for neuroscientists, while PhosphoSitePlus [137] is an excellent source of information for phosphorylation studies specifically.

In our analysis, annotation databases are employed directly to provide background information for our data and results, and indirectly through e.g. enrichment or network algorithms, where the databases provide the knowledge basis for association or function inferences, and is indispensable to the algorithms' operations. Annotation databases have become an essential tool in the modern-day scientific research.

1.6.2.2 Differential expression analysis

Differential expression refers to the pattern of change manifested in the experimental condition compared to the control condition. It is generally measured with statistics to take into account the range of variation that could take place between individual samples. Statistical tests typically compare 2 or more groups and determine whether the groups in the comparison are statistically different based on a probability value, or p-value. P-value threshold can be set by the test performer, but the commonly accepted value is 0.05 or 0.01. This would mean that the probability of having the observed group distributions under the assumption there is no difference between the two group is 5% or less for p-value threshold of 0.05. There are pre-conditions which need to be satisfied for statistical tests to yield reliable results. T-test, for example, assumes normality of data distribution [138]. If any group is not normally distributed, T-test should not be chosen since the distributions would not meet with the pre-conditioned assumption for its algorithm, and the resulting statistics could be misleading for this reason. Based on the distribution prerequisite, statistical tests can be classified as either parametric or non-parametric methods. Parametric methods make assumptions on the distribution of the input data; non-parametric methods do not make assumptions or make very few assumptions on data distribution, and hence can assess as intended on data that's not normally distributed. However, when normality is present for the data, parametric tests generally provide greater power than their nonparametric counterpart tests [139].

In the context of phosphoproteomics analysis, differential expression analysis compares phosphorylation changes for each phosphopeptide entry between two or more experimental conditions. A phosphopeptide is deemed significantly changing between conditions when the p-value or false discovery rate (FDR) of the test is smaller than the threshold. FDR values can be calculated from performing multiple testing corrections. The statistical analysis of proteomics data typically involves a large number of hypothesis tests due to the numerous peptide features in the dataset. Multiple testing correction methods are used to adjust for the increased likelihood of false positives that arise when multiple tests are conducted simultaneously [140]. The significantly changing phosphopeptides would become the focus of the aftermath analysis, such as enrichment analysis or network analysis. If the conditions

are disease and healthy, these phosphopeptide would hold a key role in understanding the mechanistic insight of the disease.

1.6.2.3 Enrichment analysis

Enrichment analysis refers to the process of testing whether a key term is associated with the significant phosphopeptides more than expected by chance. A key term could be gene ontology terms, pathways, cell types, or any other categories that are annotated with protein/peptide affiliations. Enrichment analysis could be broadly classified into over-representation test and gene set enrichment analysis (GSEA), which bases its method on the ranking algorithm (Figure 19).

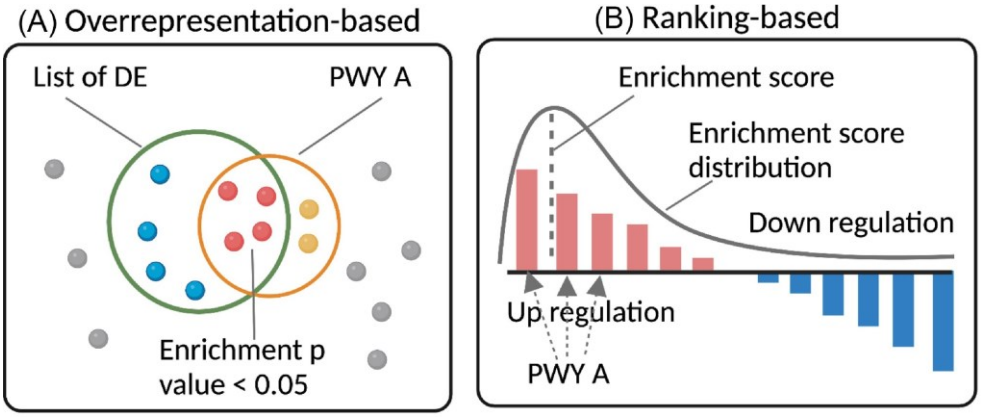

Figure 19. An outline of overrepresentation-based and ranking-based enrichment methods. Overrepresentation-based enrichment (A) examines whether the frequency of proteins from a pathway or another key term (indicated as PWY A) is higher than would be expected by chance alone in a protein list of interest when comparing it to a background set. Ranking-based enrichment (B) ranks all proteins from the entire dataset first based on the detected signals such as change of expression. The it assesses whether proteins from a pathway or another key term tend to cluster at the top or bottom of the ranked list to indicate potential enrichment. Figure reprinted from [141] with permission from Elsevier.

Over-representation test inspects associated key terms from a chosen database for each significant peptide entry and for each background/control list entry. After which the key term appearances are counted in each list. An association test such as Fisher's exact test is then utilized to compare the count value of the key terms from both lists, taking into account the list sizes, and determines whether the key term is present in the significant list more than expected with a p-value or FDR [142].

GSEA does not necessarily have to be performed on genes, rather the focus is on a "set" of data. Therefore, it can also be used for proteomics data. While over-

representation test takes a small list of significant phosphopeptide as input, GSEA takes the entire set of data as input for the analysis. GSEA ranks and orders a dataset by differential expression significance, then determines whether the key term associations are spread out through the ordered dataset or are mainly clustered at the significant end/s. The key term is said to be enriched if its associations are clustered at the significant end/s. Differential expression significance in GSEA can be described by different parameters, such as fold change, correlation, or p-values from significant tests. Based on the parameter selected and testing goals, GSEA can have either only the top end or both ends considered for associating with significant phosphopeptides [143].

1.6.2.4 Kinase identification and activity prediction analysis

"Kinase analysis" predicts the identify of kinases responsible for the phosphorylation differential expression observed between samples and control. Sometimes activity is also predicted along with identify by substrate-based algorithms. The basis for the prediction is the assumption that kinase activity changes will be reflected by the phosphorylation alterations of its substrates [144]. Hence the first step of the algorithm is to establish kinase-substrate relationship between potential kinases and phosphopeptide entries from the dataset. A library of kinase-substrate annotations is usually utilized to establish a phosphorylation pattern, or motif, for the kinase, and based on the motif the most probable substrates are assigned to each potential kinase. Kinase activities are then deduced from the differential changes of the substrates between samples and control. Sometimes not only substrates are considered for kinase activity prediction, but closely interacting proteins and sites are also considered; these algorithms which take into account indirect associations are network based and utilize annotations from network databases [145]. For the described algorithms, kinase activity prediction heavily relies on the accuracy and extensiveness of the libraries it utilizes since a well-established motif serves as the backbone to all the follow up predictions. Hence it is unfortunate that the less studied kinases would yield less precise results. In the case of a poorly studied kinase, one could turn to algorithms where kinase motifs are predicted by structural similarity to a more well-known protein or ortholog, given that a better studied match exists [146].

1.6.2.5 Network analysis

Protein functions are usually diverse, most play a role spanning multiple pathways. To consider the functional impact of a list of significant proteins, where each may be engaged in several pathways, with or without overlaps, and could interact with each other directly or indirectly, a good organization of information is necessary in

order to reach a reasonable conclusion. Network analysis describes the process of this organization. From this analysis, key proteins and driver mutations can surface as a result of interaction calculations. A network visual consists of nodes and edges (Figure 20). Nodes are proteins or another interacting agent such as a drug or a peptide; edges are representations of interactions and connects two nodes together to indicate a relationship between the two nodes. Edges can be directional (usually represented with arrows) or nondirectional. Nodes in a network can be stretched out and grouped by functions, pathways, GO terms, or other sensible systems. Edges can have weight indicating the strength of evidence supporting the interaction, or they can be separated into multiple edges, each representing a different type of interaction. Based on the interaction patterns, algorithms are developed to e.g. cluster highly interacting nodes, or identifying hub areas where one node directly and indirectly interacts with all or nearly all nodes. All of which facilitates better understanding of functional impact and key players involved [147].

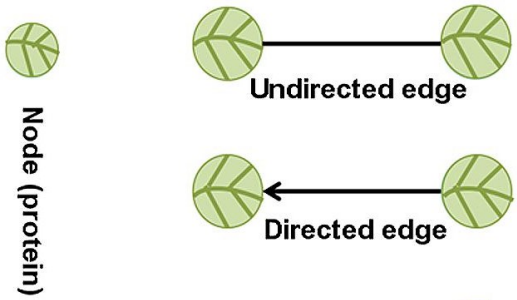

Figure 20. An illustration of node and edge in a protein-protein intereaction. A network image consists of many such interactions, and connecting the nodes into one network [148].

2 Methods

2.1 Section Content

This thesis comprises four studies, with three focused on investigating alterations in disease-associated phosphoproteomic and proteomic abundances, while one study focuses on the analysis methodology of proteomic and phosphoproteomic intensity data. The three biological exploration studies include mass spectrometry (MS) intensity data, alongside other biological data from cellular and animal models, wet lab validation, and behavioral tests. My involvement in these studies entails the analysis of MS intensity data, while co-authors of these studies analyzed other sources of biological data. The final methodology study introduces a new automated pipeline, PhosPiR, which compiles the analysis methods I utilized in the analysis of the three biological studies with additional methods for phosphoproteomics analysis. In this method section, the focus is on the specific methods used for analyzing MS intensity data in the four studies. The methods employed by my co-authors will be briefly mentioned in the results section, along with their implications. The PhosPiR pipeline integrates most of the methods employed in all the studies, and the method section will present the PhosPiR methods first, followed by additional methods utilized that are not part of the PhosPiR pipeline.

2.2 Features of PhosPiR analysis pipeline

2.2.1 Graphical user interface (GUI)

Graphical user interface (GUI) in the pipeline is supported by the R package "svDialogs" [149]. Through GUI, users can select analysis methods and sample related information such as organism, entering group orders and group names, and check short guidelines. "svDialogs" package stated that Windows, Mac and Linux operating systems are supported, for PhosPiR, Windows is fully tested, and Mac and Linux is not yet tested.

2.2.2 Input formatting

For MaxQuant generated input, the pipeline has two additional formatting options, one is to “expand” the dataset, the other one is to double check marked potential contaminants for possible false labels. Both options are coded using base R without additional packages.

Expanding the dataset set option inspects the percentage of one, two, or multiple phosphorylation sites found on each phosphopeptide, and distinguish peptide with different site count as different peptide entries. Instead of combining the intensity for the same peptide sequence, this option separates the intensity into their respective site count. This could prevent masking effect if the level of differential expression is not evenly distributed between different site counts. Table 1 demonstrates the expanded input option with an example.

For the contaminants checking option, protein IDs of each contaminant is searched against the entire dataset, and if the same ID can be found in entries not marked as contaminant, then these contaminants would be marked as ambiguous, and kept in the dataset.

Table 1. The standard PhosPiR input that is automatically formatted from the MaxQuant result by PhosPiR is shown at the top. The expanded PhosPiR input data resulting from the expanded format option is displayed at the bottom. In comparison to the standard input, the same site entry is separated into multiple entries based on the phosphorylation count of its peptide window.

Standard PhosPiR input formatted from MaxQuant result vs expanded PhosPiR input formatted from MaxQuant result

Standard PhosPiR input example								
Protein	Protein.names	Gene.nam	Amin	Position	Sequence.window	Intensity.2	Intensity.2	Intensity.2
D3YVU7	Lysine-specific demethylase 2B	Kdm2b	S	893	KTESTLAHESQQPIKSEPESENDEPKRPLSH	24706000	0	19758000
D3YVU7	Lysine-specific demethylase 2B	Kdm2b	S	897	TLAHESQQPIKSEPESENDEPKRPLSHCERP	24706000	0	19758000
D3YWA4	PEST proteolytic signal-containing nuclear protein	Pcnp	S	52	PKTLSVAAAFNEDEDSEPEEMPPEAKMRMKN	28328000	13354000	16169000
E9PWC5	MAGUK p55 subfamily member 6	Mpp6	Y	104	FQSLLEAHDIVASKCYDSPPSSPEMNIPSLN	0	4292800	5754800
E9PWE8	Dihydropyrimidinase-related protein 3	Dpysl3	Y	612	ARRKMADLHAVPRGMYDGPVFDLTTTPKGGT	0	4971400	0
Expanded PhosPiR input example								
Protein	Protein.names	Gene.nam	Amin	Position	Sequence.window	Intensity.2	Intensity.2	Intensity.2
D3YVU7	Lysine-specific demethylase 2B; phospho-count=1	Kdm2b	S	893	KTESTLAHESQQPIKSEPESENDEPKRPLSH	0	0	0
D3YVU7	Lysine-specific demethylase 2B; phospho-count=2	Kdm2b	S	893	KTESTLAHESQQPIKSEPESENDEPKRPLSH	24706000	0	19758000
D3YVU7	Lysine-specific demethylase 2B; phospho-count=3	Kdm2b	S	893	KTESTLAHESQQPIKSEPESENDEPKRPLSH	0	0	0
D3YVU7	Lysine-specific demethylase 2B; phospho-count=1	Kdm2b	S	897	TLAHESQQPIKSEPESENDEPKRPLSHCERP	0	0	0
D3YVU7	Lysine-specific demethylase 2B; phospho-count=2	Kdm2b	S	897	TLAHESQQPIKSEPESENDEPKRPLSHCERP	24706000	0	19758000
D3YVU7	Lysine-specific demethylase 2B; phospho-count=3	Kdm2b	S	897	TLAHESQQPIKSEPESENDEPKRPLSHCERP	0	0	0
D3YWA4	PEST proteolytic signal-containing nuclear protein; phospho-count=1	Pcnp	S	52	PKTLSVAAAFNEDEDSEPEEMPPEAKMRMKN	28328000	13354000	16169000
D3YWA4	PEST proteolytic signal-containing nuclear protein; phospho-count=2	Pcnp	S	52	PKTLSVAAAFNEDEDSEPEEMPPEAKMRMKN	0	0	0
D3YWA4	PEST proteolytic signal-containing nuclear protein; phospho-count=3	Pcnp	S	52	PKTLSVAAAFNEDEDSEPEEMPPEAKMRMKN	0	0	0
E9PWC5	MAGUK p55 subfamily member 6; phospho-count=1	Mpp6	Y	104	FQSLLEAHDIVASKCYDSPPSSPEMNIPSLN	0	0	0
E9PWC5	MAGUK p55 subfamily member 6; phospho-count=2	Mpp6	Y	104	FQSLLEAHDIVASKCYDSPPSSPEMNIPSLN	0	0	0
E9PWC5	MAGUK p55 subfamily member 6; phospho-count=3	Mpp6	Y	104	FQSLLEAHDIVASKCYDSPPSSPEMNIPSLN	0	4292800	5754800
E9PWE8	Dihydropyrimidinase-related protein 3; phospho-count=1	Dpysl3	Y	612	ARRKMADLHAVPRGMYDGPVFDLTTTPKGGT	0	0	0
E9PWE8	Dihydropyrimidinase-related protein 3; phospho-count=2	Dpysl3	Y	612	ARRKMADLHAVPRGMYDGPVFDLTTTPKGGT	0	4971400	0
E9PWE8	Dihydropyrimidinase-related protein 3; phospho-count=3	Dpysl3	Y	612	ARRKMADLHAVPRGMYDGPVFDLTTTPKGGT	0	0	0

2.2.3 Data processing

2.2.3.1 Normalization

Median normalization and quantile normalization is offered in the pipeline. Median normalization centers all sample medians to the global median of the data by first obtaining the global median of the entire dataset, then determining the difference between the global median and the sample median and apply the difference to the respective sample data distribution [150]. Quantile normalization assigns identical quantiles to each sample, or column, of the data. Columns are sorted separately in numerical order, then averages are calculated for each row of the sorted dataset. Each element in the row is then replaced with the numerical value of the average. The data is then put back in the original order to complete the quantile normalization [151]. Both normalizations are performed with the "proBatch" package in R [150].

2.2.3.2 Imputation

Aside from normalization, "MSImpute" function is utilized in the pipeline for missing value imputation. The imputation algorithm is low-rank approximation via alternating least squares [131]. A dataset with n samples and m features can be approximated and reconstructed by a set of linear combination of features where the size of the set is less or equal to the minimum of n and m. The said algorithm calculates two low rank matrices, and takes their product to reconstruct the original matrix, with missing values estimated [131]. The two low rank matrices would have dimensions n x r and m x r, where r is less or equal to the minimum of n and m. the matrices are calculated from the following minimizing problem:

$$\underset{A,B}{minimize} \|P_{\Omega}(X - AB^{T})\|_F^2 + \frac{\lambda}{2}\left(\|A\|_F^2 + \|B\|_F^2\right) \quad (1)$$

Where A and B are the two matrices, P_{Ω} is the subset of data where missing values are removed, $\|_F^2$ is the nuclear norm which encourages low rank solutions, and λ is the shrinkage operator [145]. To solve this function, two least square problems needed to be solved in alternation. They are:

$$\underset{B}{minimize} \|P_{\Omega}(X - AB^{T})\|_F^2 + \lambda \|B\|_F^2 \quad (2)$$

And

$$\underset{A}{minimize} \|P_{\Omega}(X - AB^{T})\|_F^2 + \lambda \|A\|_F^2 \quad (3)$$

These steps are repeated until consecutive iterations produce converging results. The λ variable is data-driven, the optimal λ value is calculated through the "msImpute" function, within "MSImpute" R package [131].

2.2.4 Overview figures

Five different types of figures are plotted to show the overall distribution of the data, histogram, boxplot, heatmap, 3D PCA, and PCA with k-means clusters. Figure 21 illustrates four of them, histogram, boxplot, heatmap and PCA with k-means clustering, 3D PCA is plotted as an animation and hence is not included in the figure.

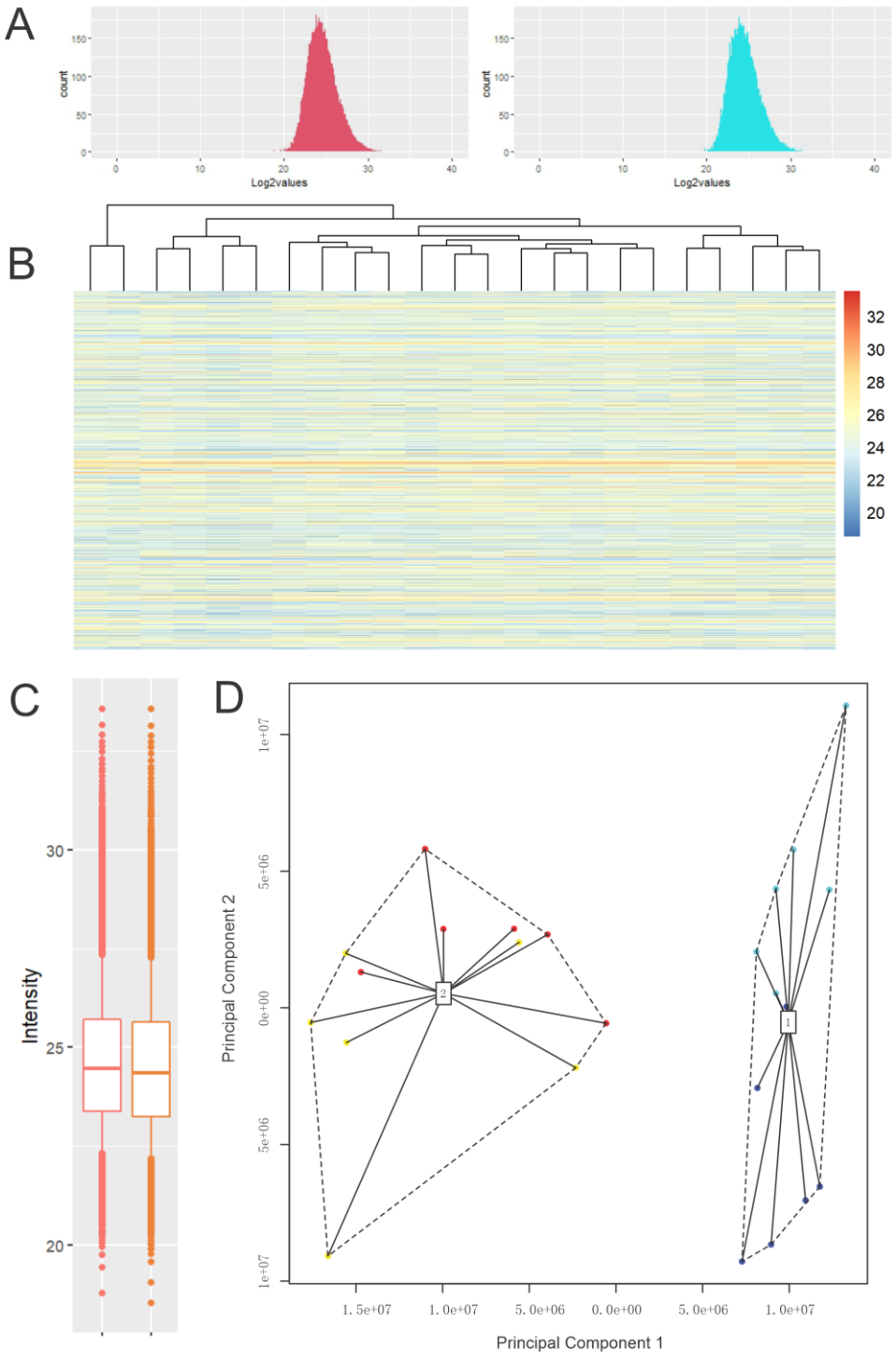

Figure 21. Illustration of histogram (A), heatmap (B), boxplot (C) and PCA with K-means clustering (D) figure formats.

Histogram partitions the data into different value ranges and displays the frequency of data points in each value range. Therefore, x-axis of histogram displays intensity ranges, and y-axis of histogram displays frequency of data points belonging to each range.

A boxplot comprises a center box and two whiskers, one on each side. It describes the numerical data distribution. A line inside the box indicates the median, and the box itself represents the range of the interquartile region, where values between 25^{th} percentile to 75^{th} percentile are included. The whiskers extend from the 25^{th} or 75^{th} percentile, to include values that are either 1.5 times the interquartile range smaller than the 25^{th} percentile, or larger than the 75^{th} percentile. Any points not represented by the whiskers are plotted as dots at their respective values, indicating outliers.

Heatmap changes numerical value into a corresponding color gradient assigned to the data range, it is coupled with a color key to show the gradient scale, and it is often shown with dendrograms clustering samples or features by hierarchical clustering.

PCA plots take the first two (2D figure) or three (3D figure) principal components of the principal component analysis, and represent the data, with multiple feature dimensions, in a reduced dimensional format on a scatterplot, where each dot represent an entity sample or feature, and x- and y-, (and z- for 3D figures) axis correspond to first and second (and third) principal components of the entity. Principal component analysis refers to the calculation of n principal components from a dataset with n dimensions for each of its variables. Principal components are projections of the original individuals from the data onto a subspace where the variance of the data is maximally kept. Each n components retain a portion of the original variance, and they are ordered in such way that the first component preserves the most variance, and each component after it preserves a progressively decreased amount of variance. Hence it is commonly utilized as a dimensional reduction method considering the first two or three components would hold a high percentage of the total variance. To calculate the subspaces which retain maximum variance through projection, the mean of each dimension, and the covariance matrix of the dataset is calculated. Eigenvalues of the covariance matrix are determined by setting the determinant of the difference between covariance matrix and eigenvalue identify matrix to 0, then solving the equation for the eigenvalues. Eigenvectors, or the dimensions for the subspaces, is subsequently calculated for each of the real eigenvalue solutions. The data can then be projected onto the subspaces via matrix multiplication [152].

K-means clustering is performed on the 2D PCA. This cluster method belongs to the category of clustering by partitioning. The algorithm separates datapoints into groups, where the best placement for each point is determined by minimizing the

sum of squared distance between the point and the center position (centroid) of the closest group [153]. Initially, n+1 centroids are randomly assigned in the data space, where n is the user defined number of sample groups. Data points are assigned to the nearest centroid, after which the next iteration of finetuning begins. Each iteration recalculates the centroids positions based on the data point locations included in each group. From the new centroid position, data points are then reassigned following the same distance criteria. Finetuning of the group assignment ends when distance between data points and the assigned centroids reaches a local minimum. The starting positions of the centroids are crucial to achieve the best grouping, where a global minimum is reached instead of hitting a local minimum. Due to the position randomization in the beginning, rerunning the algorithm more than once is advised [153].

Histogram and boxplot are plotted with R package "ggplot2" [154], heatmap is plotted with R package "pheatmap" [155], 3D PCA and 2D PCA with k-means clusters are created with the support of R packages "fingerprint" [156], "vegan" [157], "rgl" [158], "FactoMineR" [159], "factoextra" [160], "plot3D" [161], and "magick" [162].

2.2.5 Annotation

For each entry, or row, in the dataset, a collection of information is mined, including various ID symbols, sequence information, taxonomy, location information, PTM information, interactions, pathology, related publications and more. The information sources are Ensembl database [133] and UniProt database [134], and "biomaRt" [163], "protr" [164] and "UniprotR" [165] R packages are utilized to obtain the annotations.

For any nonhuman dataset, an option to align the proteins to human orthologs is offered. Pairwise sequence alignment of the target protein and its ortholog protein is performed in this case. The alignment algorithm applies a scoring system to penalize mismatches and gaps, and aim to transpire the best alignment by optimizing the final summed score. Not all mismatches should have the same score considering in an actual protein the likelihood of different amino acids becoming a substitute for the target amino acid varies greatly. For this reason, the block amino acid substitution matrices (BLOSUM) are employed to assign mismatch scores. BLOSUM have specific score tables with different scoring systems for a range of sequence similarity (Figure 22) to explicitly accommodate alignment of divergent organisms. These tables are denoted BLOSUMN, where N indicates the similarity percentage threshold for the two sequences being assigned, and a greater than threshold similarity is preferred to optimize the alignment [166]. Here, BLOSUM100, BLOSUM80, BLOSUM45 and BLOSUM62 are utilized for sequence similarity of

greater than 90%, greater than 80%, less than 45%, and all in between, respectively. BLOSUM62 is a special case and tested to be working well with a wide range of similarities. Ortholog similarity is acquired from the Ensembl database through information mining. Gap penalty scores are assigned with the default scores of "pairwiseAlignment" function from "Biostrings" R package [167], where gap opening penalty is ten, and incremental gap extension penalty is four. The alignment itself is performed with the same "pairwiseAlignment" function.

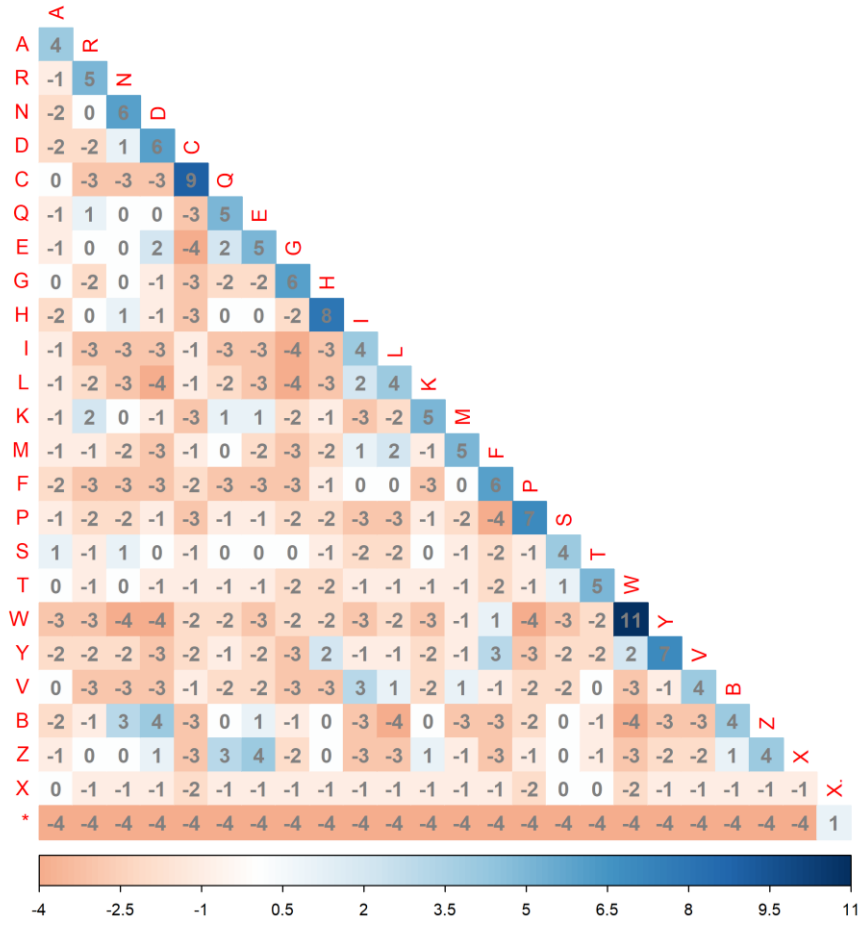

Figure 22. An illustration of substitution score table of amino acids from BLOSUM sequence similarity matrix. This specific matrix is from BLOSUM62. The matrix values are obtained from the National Library of Medicine repository [168].

2.2.6 Statistical tests

Four different types of statistical tests are offered for differential expression analysis, they are T-test, Wilcoxon rank sum test, ROTS test, and rank product test.

T-test compares the mean of two groups in a two-sample t-test. The t-value is calculated by

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{S^2(\frac{1}{n_1} + \frac{1}{n_2})}} \quad (4)$$

Where t is the t-value, $\bar{x}_1$ and $\bar{x}_2$ are the group means, S^2 is the total sample variance, and n_1 and n_2 are the number of elements included in group one and group two, respectively [169]. The calculated t-value is inspected on a t-distribution with the corresponding degree of freedoms from the two groups. The area of the t-distribution with more extreme absolute values are calculated, which in turn gives the p-value that is used to determine whether the two distributions being compared are significantly different [169].

Wilcoxon rank-sum test combines samples from both groups, and rank from lowest to highest, assigning the smallest value a rank of 1 and the largest value a rank of n, where n is the total sample size. The sum of the ranks for the group being compared is calculated as the test statistic. The test statistic is then compared to the expected rank-sum distribution under the assumption there is no difference between the two groups. The area of the rank-sum distribution with more extreme absolute values are calculated to give the p-value for the test [170].

ROTS test optimizes a modified t-type statistic for the input data. The test will try to maximize the reproducibility of

$$\frac{|\bar{x}_1 - \bar{x}_2|}{\alpha_1 + \alpha_2 s} \quad (5)$$

Where $\bar{x}_1$ and $\bar{x}_2$ are mean of group one and two, respectively, s is pooled standard error, α_1 and α_2 are non-negative parameters to be optimized [171]. There are two special cases of ROTS, when α_1 and α_2 is optimized to be zero and one, respectively, it is an ordinary t-statistic, and when α_1 and α_2 is optimized to be one and zero, respectively, it is a signal log-ratio [171].

Rank Product utilizes a ranking algorithm rather than t-statistic. Rank Product assumes a non-significant expression pattern will result in random ordering among repeats of the same condition; significant differences, on the contrary, will always fall in the top ranks. Hence, the final observed rank placement probability assuming no differential expression takes place for each protein is related to the value

$$\prod_{i=1}^{k} \frac{r_{i,p}}{n_i} \quad (6)$$

where $r_{i,p}$ is the rank of protein p in ith replicate of k replicates, n_i is the total number of proteins in ith replicate, and the product of all replicates will determine the significance of final ranking placement for each protein [172].

T-test and Wilcoxon rank-sum test are performed with base R, ROTS is performed with "ROTS" R package [171], and rank product is performed with "RankProd" R package [173].

2.2.7 Enrichment

Gene ontology (GO) enrichment, Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment, cell marker enrichment, and disease association enrichment are performed with the "clusterProfiler" R package [174]. All of which belong to the category of over representation analysis, which employs a hypergeometric test, and requires a significant list, or a small set of significant proteins, as input.

PTM-SEA is a "gene set" enrichment analysis performed with the ssGSEA2.0 tool, which employs the PTMsigDB manually curated library [175]. The library stores signature sets which group phosphopeptides by a common functional connection, for example, the mTOR kinase set groups together phosphopeptides that is directly affected by mTOR kinase activities. PTM-SEA predicts a signature's activity change from the enrichment analysis. Using the previous example, mTOR's activity change is predicted based on its substrates' enrichment as well as the substrates relation with mTOR activity, i.e. whether each individual substrate is inhibited or activated by mTOR activity.

2.2.8 Kinase analysis

Kinase analysis is carried out with "KinSwingR" R package [176]. Several steps are performed (Figure 23), yielding the final result where kinases, as well as its activity alteration, are predicted based on phopshopeptide intensity derived phosphorylation changes. The first step of the analysis is to define kinase motifs from reference libraries. Kinase library from PhosphoSitePlus is utilized to provide kinase identities and their reference substrates. For each kinase, a log likelihood ratio matrix is calculated. The 20 amino acids are represented in matrix rows, and the substrate sequence (15 in length from the reference library) is represented in the matrix columns. The likelihood of amino acid, a, at sequence position, p, in a substrate of kinase, k, is determined and represented in the matrix at row a, and column p [176]. Once the motif is solved, the next step of the analysis predicts kinase-substrate match from the user dataset. For each phosphopeptide entry from the dataset, where the phosphosite is centered on the sequence, probability scores are calculated, one for

each kinase, based on the likelihood values of that kinase. The score sums the corresponding likelihood values from the likelihood matrix for each amino acid in each sequence position of the phosphopeptide entry. After which, 1000 random sequences of the same length are generated, with their likelihood scores calculated as a background distribution. Only when the likelihood score of the phosphopeptide entry is significantly larger than the background likelihood score, would the phosphopeptide count as a substrate match for this particular kinase. Not all substrate matches are included in the prediction calculation of kinase activities [176]. Step three of the analysis inspects the fold change and p-value of the substrate matches and keeps only the significantly changing substrates. The directionality of the change is also preserved, although not the scale of the change [176]. Kinase activity is then calculated from the significant substrates in step four of the analysis. The raw swing score is calculated as a ratio of positively and negatively changing substrates, while taking into account the number of significant substrates, and the number of total matched substrates. The raw score is then transformed into a weighted z-score, where the swing score distribution mean is centered at zero, with standard deviation of one [176]. The direction and scale of the kinase activity change is made apparent by the weighted swing score.

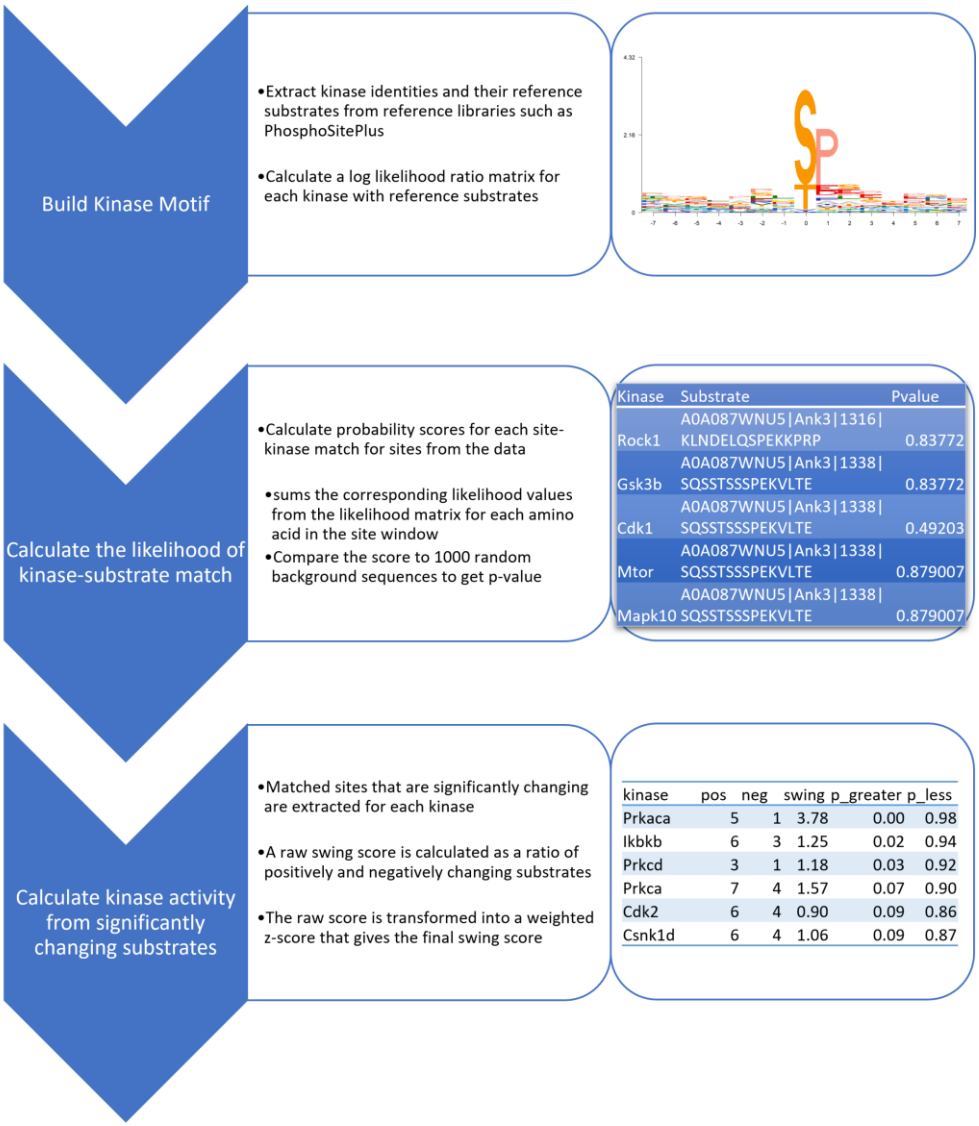

Figure 23. The workflow of KingSwingR package to predict kinases and their activity changes for a phosphoproteomics dataset. Example output from each step are shown on the right side of the figure.

2.2.9 Network

Kinase network connects kinases to their respective substrates utilizing information from PhosphoSitePlus kinase database [177], which is visualized with a chord diagram plotted utilizing the "circlize" R package [178].

Protein interaction network utilizes STRING database [179] (Figure 24) and extracts all proteins that interact with the query proteins with a confidence score of

0.4 or higher. For protein interaction networks, hub significance can be calculated with the following methodology: for each hub protein in the query network, 1000 background networks are created from randomly chosen proteins that are present in the dataset, and the hub protein itself. The size of the background networks matches the query network as network size influences interaction magnitude. The hub protein's interaction counts in background networks formulate a reference distribution, which enables the calculation of a p-value and FDR value to determine whether the interaction magnitude of the target hub is significant in the query network.

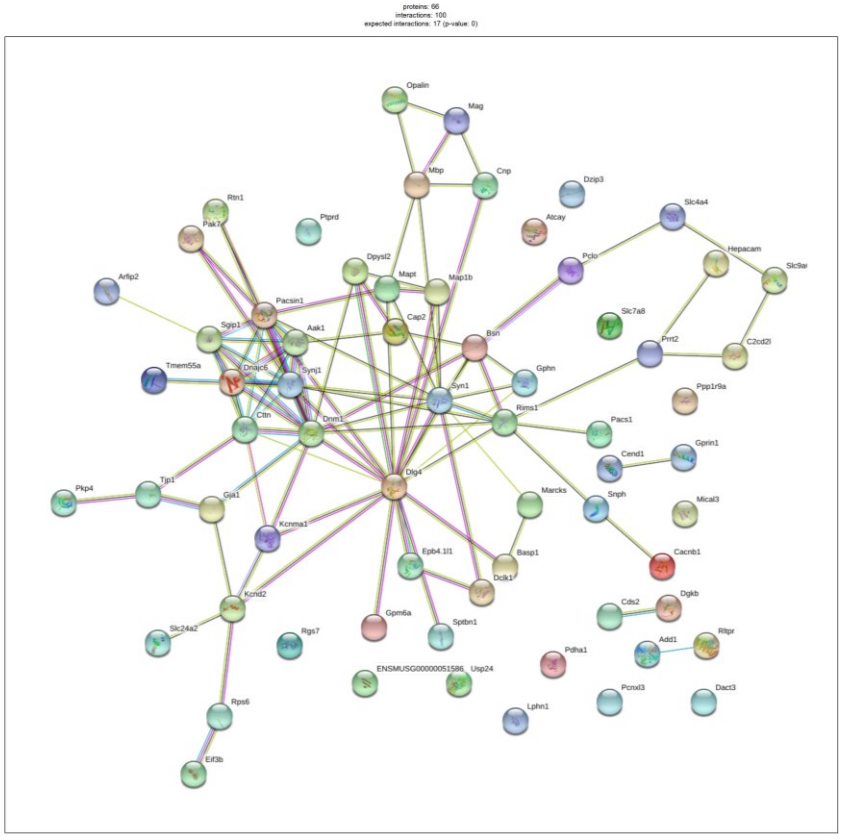

Figure 24. An example STRING network generated from the PhosPiR analysis of brain synaptoneurosomes in sleep-deprived mice from Brüning et al.'s study.

2.3 Additional analysis methods for MS intensity data (outside of PhosPiR methods)

2.3.1 Fisher's exact test

Fisher's exact test is utilized in Study IV. Fisher's exact test determines whether there was a significant association between two or more categories of variables by comparing their frequencies. The frequencies of the variables are recorded in a two by two or larger contingency table, and a p-value is calculated from the frequencies indicating the likelihood of observing the recorded frequencies under the assumption that there is no association between the tested variables [180]. Fisher's exact test p-value for two variables is calculated with the formula

$$\frac{r!\,(n_1 - r)!\,(n_2 - r)!\,(N - n_1 - n_2 + r)!}{n_1!\,n_2!\,(N + 1)!} \tag{7}$$

where r is the number of observations in the two groups that have the variable of interest, n_1 is the total number of observations in the first group, n_2 is the total number of observations in the second group, and N is the total number of observations in both groups combined. Rather than using a normal approximation, the formula calculates the exact probability of observing the recorded frequencies and more extreme frequencies based on a hypergeometric distribution [180]. Fisher's exact test is performed with a base R function.

2.3.2 MetaCore enrichment analysis

MetaCore is a commercial bioinformatics software platform developed by Clarivate Analytics that provides a suite of tools for pathway analysis, network building, and functional annotation of genomic and proteomic data. MetaCore enrichment analysis was performed in Study IV on the MetaCore software platform. Lists of significant proteins were uploaded, and enrichment analysis was performed using a hypergeometric test to determine the statistical significance of the overlap between the input lists and each pathway or process in the MetaCore curated database. The p-values were corrected for multiple testing using the Benjamini-Hochberg method [181].

2.3.3 Cytoscape network analysis

Cytoscape is an open-source software platform for visualizing and analyzing molecular interaction networks [181]. Networks were built with GeneMANIA plugin in Cytoscape software in Study IV. A To build a network with GeneMANIA, set of input proteins is provided and interaction types are selected from co-

expression, co-localization, genetic interactions, physical interactions, and pathway relationships. For each input, GeneMANIA calculates a score for each of the selected interaction types based on the strength of the evidence linking the input to other proteins in the network. These scores are combined using a weighted sum to generate a final score for each protein-protein relationship in the network [182].

3 Results

3.1 Section Content

In this result section, the outcomes of four studies included in this thesis is presented. Study I provided an overview of PhosPiR, an automated R pipeline. The methodologies included in PhosPiR is introduced in the method section. The results of the pipeline analysis, including graphical and table outputs, are shown in this section, along with a summary of the key findings from PhosPiR analysis of brain synaptoneurosomes in sleep-deprived mice. Studies II to IV focused on exploring kinase-associated pathologic mechanisms of Parkinson's disease and schizophrenia using various methods, which are briefly described before presenting the results. As my role in these studies is to analyze MS intensity data, any methods not introduced in the method section were not performed by myself, but they are still reported since they are part of the studies.

3.2 Functionalities and generated output of PhosPiR (Study I)

The study of phosphoproteome in conjunction with proteome is essential for mechanistic investigations of brain-related diseases. To aid in the study of phosphoproteome and proteome, an automated pipeline in R called PhosPiR has been developed, that automates the analysis workflow starting from data preprocessing, and offering a range of analysis methods, making it a versatile and efficient tool for analyzing for phosphoproteomic and proteomic datasets (Figure 25). The analysis methods included in PhosPiR are described in the Methods section. The results of these analysis methods are presented in various formats, examples of output results that are not illustrated in the Methods section are presented here.

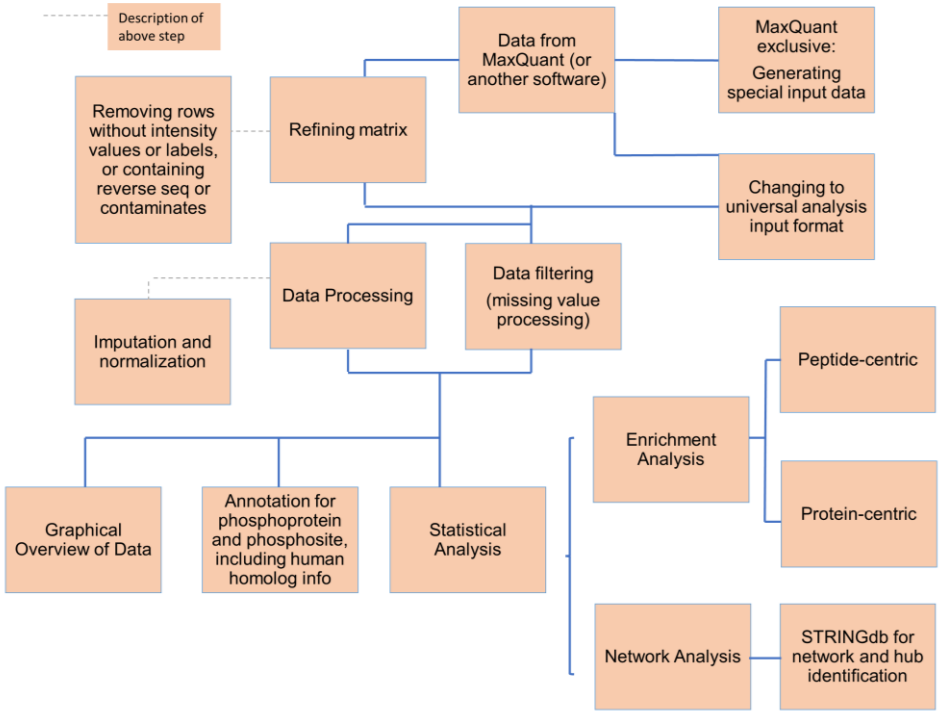

Figure 25. Overview of PhosPiR workflow

To showcase PhosPiR output, we have utilized data from brain synaptoneurosomes in sleep-deprived mice from Brüning et al.'s study [183]. Data overview figures are shown in Figure 21, which visualizes the data values after PhosPiR pre-processing steps such as normalization and imputation have been applied. To identify differentially expressed proteins, statistical analysis is performed, and volcano plots are employed to visualize the results (Figure 26). The resulting protein lists of interest undergo enrichment analysis, PTM-SEA, kinase activity prediction, and network analysis by PhosPiR. Dotplots are used to represent the results of enrichment analysis (Figure 27), and PTM-SEA results are represented in rank plots (Figure 28). To display the networks created from protein lists of interest, STRING is utilized, and the corresponding network image is depicted in Figure 24. The hub genes from these networks are analyzed for significance against background networks, and the results are represented in boxplots (Figure 29). Using the kinase prediction results, kinase-substrate networks is constructed then illustrated using circos plots. Additionally, annotations are extracted from databases such as UniProt and PhosphoSitePlus and presented in various tables, and an example of the information obtained from the annotation is shown in Table 2. In case the dataset organism is not human, PhosPiR performs pairwise alignment for each protein to its

human protein homolog, and the output from pairwise alignment result is depicted in Figure 30.

PhosPiR analysis of brain synaptoneurosomes in sleep-deprived mice has revealed several important biological implications. The dopaminergic synapse pathway was significantly enriched, with significantly altered phosphorylation during wake and sleep time. Phosphosite-centric enrichment analysis showed a downregulation of the "rapamycin" signature set by 40% and an upregulation of the "mTOR" signature set by 14%, consistent with known negative regulation of mTOR by rapamycin. Through kinase-substrate analysis, NEFM, with the most significant decrease in phosphorylation, was shown to be regulated by SRC, ADRBK1, CSNK1D, and PRKCD during wake hours in sleep-deprived mice. RPS6KA1 showed the most increased activity among kinases based on motif phosphorylation, while PRKCZ showed the largest decrease in activity during wake hours in sleep-deprived mice. The hub analysis of protein phosphorylation identified GRIN2B, SHANK3, and SYN1 as highly significant signaling hubs. MAPT phosphorylation was also shown to increase upon sleep deprivation stress. These results confirm previous findings and provide novel insights into the molecular mechanisms underlying sleep deprivation and its effect on neurological disorders, highlighting the utility of the PhosPiR pipeline.

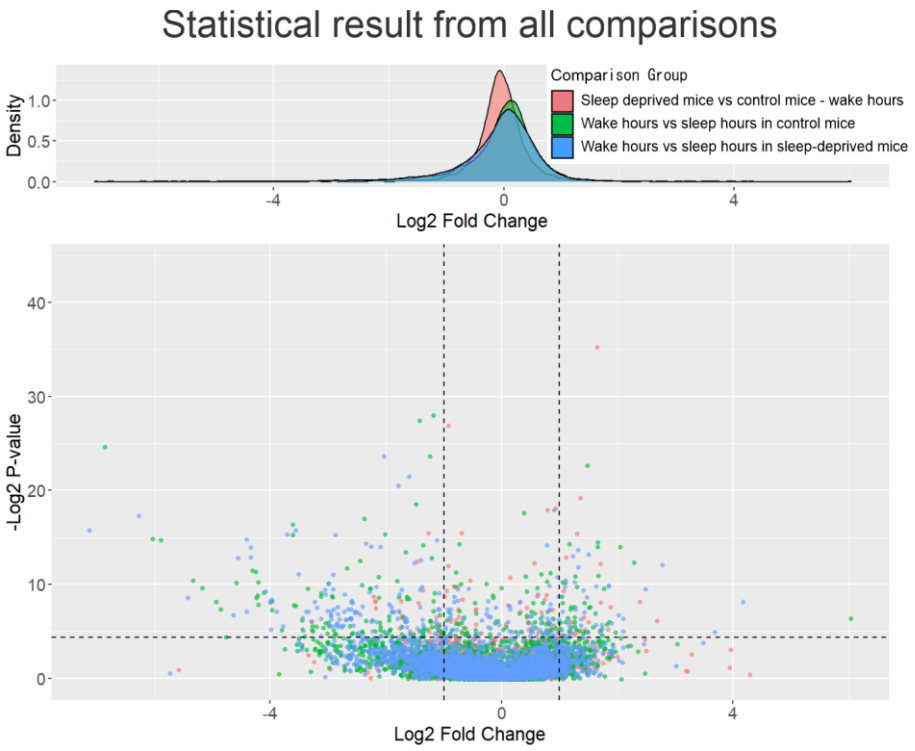

Figure 26. Example figure of a volcano plot showing fold change (X-axis) and statistical results (Y-axis) [184]. A density plot is accompanied at the top showing the distribution of the fold changes. This volcano plot include comparison results from control verses sleep-deprived, and wake period verses sleep period with or without sleep-deprivation. A total of 367 significantly phosphorylated peptides were identified.

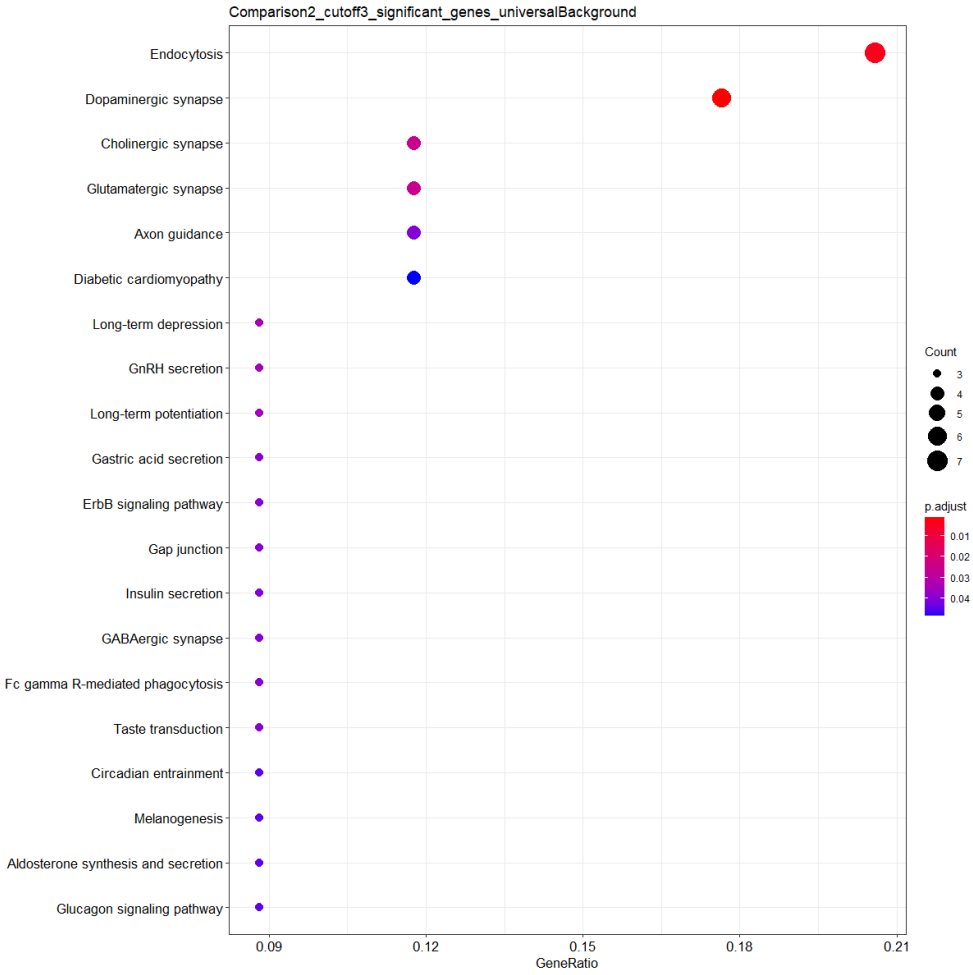

Figure 27. Example figure of a dotplot showing pathway enrichment result [184]. Endocytosis and dopaminergic synapse pathways are shown to be significantly enriched for the protein list with significantly altered phosphorylation between wake and sleep.

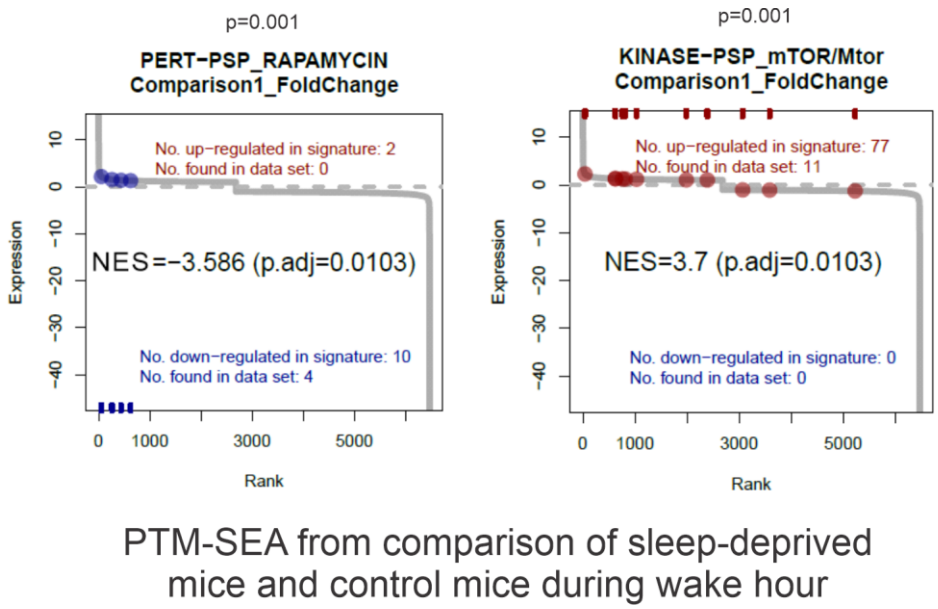

Figure 28. Example figures of rank plot showing PTM-SEA enrichment results with adjusted pvalues labeled [184]. Results shown here came from comparison of synaptoneurosomes between sleep-deprived and control mice during wake hours, which yielded a significant peptide list that was enriched with both rapamycin and mTOR activity changes.

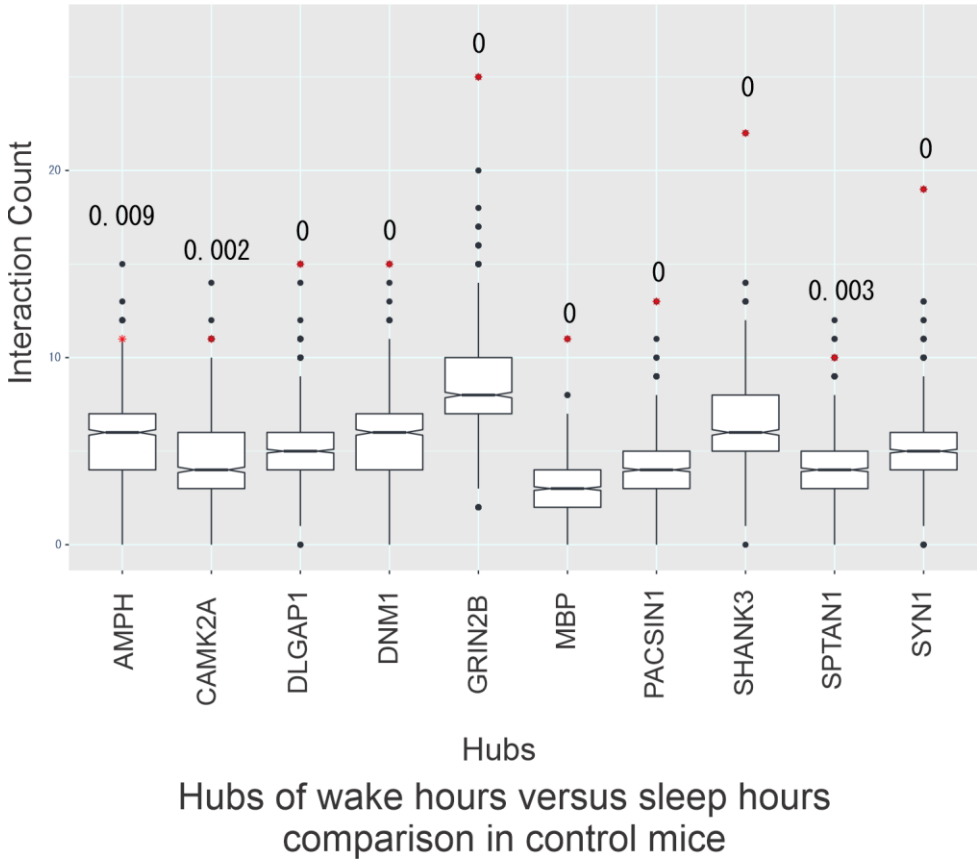

Figure 29. Example figure of hub significance plotted in boxplots [184]. The boxplots represent the background connectivity distribution, and the red dot in each boxplot represent the query network's connectivity count. Adjusted pvalue is labeled for each tested hub. This figure shows the hub analysis result from comparing synaptoneurosomes during wake time verses sleep time.

Table 1. Selected annotation output examples highlighting interesting annoation categories from the annotation extraction results.

ID				Taxonomy		Interaction	Pathology	Publication
P14873	MAP1B_MOUSE	Map1b Mtap1b Mtap5	Map1b	Mus musculus (Mouse)	UP000000589: Chromosome 13	Q61166; Q9WTU3...	NA	2480963; 19468303; 12147674; 12807913...
P16054	KPCE_MOUSE	Prkce Pkce Pkcea	Prkce	Mus musculus (Mouse)	UP000000589: Chromosome 17	P23242; Q9CQV8	NA	2917656; 9467942; 11746497; 12407104...
P20444	KPCA_MOUSE	Prkca Pkca	Prkca	Mus musculus (Mouse)	UP000000589: Unplaced	O08785	DISEASE: Note=Expression of the mutant form UV25 causes malignant transformation of cells.	2469625; 2601739; 8321321; 7844141...
P28867	KPCD_MOUSE	Prkcd Pkcd	Prkcd	Mus musculus (Mouse)	UP000000589: Chromosome 14	P23242	NA	1868068; 1765103; 11558579; 9054438...
P31750	AKT1_MOUSE	Akt1 Akt Rac	Akt1	Mus musculus (Mouse)	UP000000589: Chromosome 12	Q9Z2V5; P07901...	NA	8437858; 12783884; 16141072; 19468303...
P39053	DYN1_MOUSE	Dnm1 Dnm Kiaa4093	Dnm1	Mus musculus (Mouse)	UP000000589: Chromosome 2	Q7TQF7; Q9JJV2-1...	NA	9143510; 16141072; 19468303; 15489334...
P39688	FYN_MOUSE	Fyn	Fyn	Mus musculus (Mouse)	UP000000589: Chromosome 10	P22682; P51807...	NA	2488273; 9895129; 16141072; 15489334...

```
#######################################
# Program: Biostrings (version 2.58.0), a Bioconductor package
# Rundate: Thu Jul 01 13:01:54 2021
#######################################
#=====================================
#
# Aligned_sequences: 2
# 1: P1
# 2: S1
# Matrix: NA
# Gap_penalty: 14.0
# Extend_penalty: 4.0
#
# Length: 480
# Identity:     472/480 (98.3%)
# Similarity:    NA/480 (NA%)
# Gaps:           0/480 (0.0%)
# Score: 4688
#
#
#=====================================

P1                 1 MNDVAIVKEGWLHKRGEYIKTWRPRYFLLKNDGTFIGYKERPQDVDQRES     50
                     | ||||||||||||||||||||||||||||||||||||||||||||||||||
S1                 1 MSDVAIVKEGWLHKRGEYIKTWRPRYFLLKNDGTFIGYKERPQDVDQREA     50

P1                51 PLNNFSVAQCQLMKTERPRPNTFIIRCLQWTTVIERTFHVETPEEREEWA    100
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
S1                51 PLNNFSVAQCQLMKTERPRPNTFIIRCLQWTTVIERTFHVETPEEREEWT    100

P1               101 TAIQTVADGLKRQEEETMDFRSGSPSDNSGAEEMEVSLAKPKHRVTMNEF    150
                     |||||||||||| |||| ||||||||||||||||||||||||||||||||||
S1               101 TAIQTVADGLKKQEEEEMDFRSGSPSDNSGAEEMEVSLAKPKHRVTMNEF    150

P1               151 EYLKLLGKGTFGKVILVKEKATGRYYAMKILKKEVIVAKDEVAHTLTENR    200
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
S1               151 EYLKLLGKGTFGKVILVKEKATGRYYAMKILKKEVIVAKDEVAHTLTENR    200

P1               201 VLQNSRHPFLTALKYSFQTHDRLCFVMEYANGGELFFHLSRERVFSEDRA    250
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
S1               201 VLQNSRHPFLTALKYSFQTHDRLCFVMEYANGGELFFHLSRERVFSEDRA    250

P1               251 RFYGAEIVSALDYLHSEKNVVYRDLKLENLMLDKDGHIKITDFGLCKEGI    300
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
S1               251 RFYGAEIVSALDYLHSEKNVVYRDLKLENLMLDKDGHIKITDFGLCKEGI    300

P1               301 KDGATMKTFCGTPEYLAPEVLEDNDYGRAVDWWGLGVVMYEMMCGRLPFY    350
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
S1               301 KDGATMKTFCGTPEYLAPEVLEDNDYGRAVDWWGLGVVMYEMMCGRLPFY    350

P1               351 NQDHEKLFELILMEEIRFPRTLGPEAKSLLSGLLKKDPTQRLGGGSEDAK    400
                     |||||||||||||||||||||||||||||||||||| ||||||||||||||
S1               351 NQDHEKLFELILMEEIRFPRTLGPEAKSLLSGLLKKDPKQRLGGGSEDAK    400

P1               401 EIMQHRFFANIVWQDVYEKKLSPPFKPQVTSETDTRYFDEEFTAQMITIT    450
                     ||||||||| |||| ||||||||||||||||||||||||||||||||||||
S1               401 EIMQHRFFAGIVWQHVYEKKLSPPFKPQVTSETDTRYFDEEFTAQMITIT    450

P1               451 PPDQDDSMECVDSERRPHFPQFSYSASGTA    480
                     ||||||||||||||||||||||||||||||
S1               451 PPDQDDSMECVDSERRPHFPQFSYSASGTA    480
```

Figure 30. Example output of pairwise alignment to human homolog showing stats of the alignment and sequece to sequence aligning patterns. This particular alignment is performed for protein AKT1, comparin seqences between rat and human.

3.3 Parkinson's disease and LRRK2

3.3.1 Study of protein synthesis in sporadic and familial Parkinson's disease by LRRK2 (Study II)

3.3.1.1 Connecting LRRK2 activity with RNA translation

Meta analysis of Parkinson's disease data has shown that leucine rich-repeat kinase 2 (LRRK2)-G2019S mutation is one of the most common mutations associated with late onset Parkinson's disease. Comprehending the role of LRRK2 and the effect of the G2019S mutation is expected to be beneficial in elucidating the pathological mechanism of Parkinson's disease. With the aforementioned goal, we first separated the rat brain into fractions, and phosphorylated each fraction in vitro with purified LRRK2-G2019S utilizing a kinase substrate identification assay developed in the lab [185], to identify in which region of the brain LRRK2-G2019S function is most active. Shown in Figure 31, ribosome-enriched fractions were preferentially phosphorylated. Further resolving the ribosomal fractions showed the LRRK2 is localized to the small 40S ribosomal subunit. This led us to test whether LRRK2 activity regulates RNA translation. We applied three different LRRK2 inhibitors (IN1, GSK-2578215A, and MLi-2) separately to cultured dopaminergic and hippocampal neurons and checked de novo protein synthesis one hour following treatment with both AHA and S-methionine labeling. Protein synthesis was increased by 16% and 50% from S-methionine and AHA labeling, respectively, while LRRK2 protein levels remained the same, leading us to conclude that LRRK2 activity suppresses RNA translation.

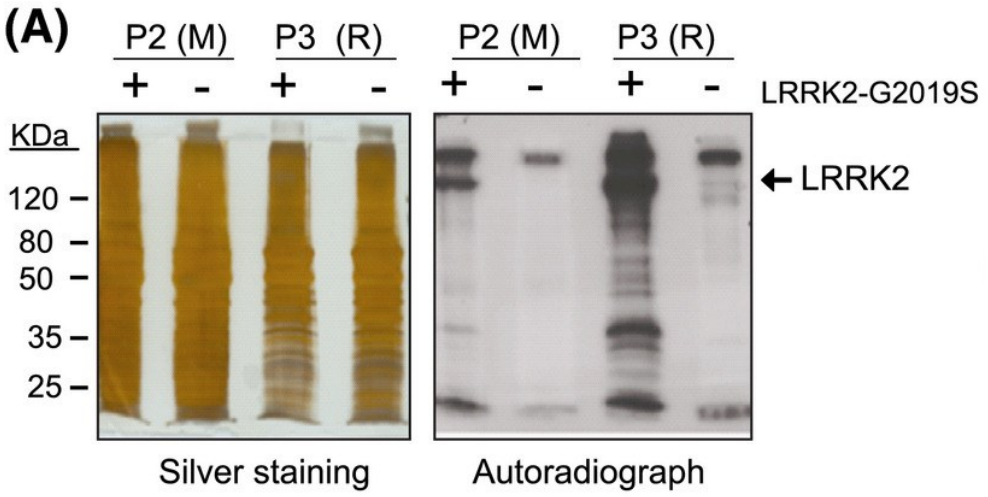

Figure 31. Mitochondrial (P2 (M)) and ribosomal (P3 (R)) fractions of rat brain were phosphorylated with and without LRRK2-G2019S in the presence of [γ-32P]-ATP. A silver-stained gel (left) and a corresponding autoradiograph (right) is depicting total protein in each lane [186].

3.3.1.2 Cellular model validation

Further wet lab tests were conducted to validate our findings and to characterize LRRK2's activity on translation in greater detail. We isolated neurons from wild type and *Lrrk2-/-* mice and utilized the same approach to quantify protein synthesis, and found it significantly increased in *Lrrk2-/-* compared to wild type mice. MLi-2 no longer had effect on protein synthesis in *Lrrk2-/-*, which validated its effect was mediated by LRRK2 inhibition. We also quantified protein synthesis in *Lrrk2* knockdown hippocampal neurons, and as expected, protein synthesis was increased. To check if LRRK2-G2019S inhibits translation by acting directly on ribosomes, we performed in-vitro translation with purified ribosomal machinery and found adding LRRK2-G2019S reduced translation by 40%, hence confirming that LRRK2-G2019S interacts with the translational machinery to inhibit translation.

As LRRK2's effect on translation appeared clear, we looked at cellular models of Parkinson's disease to inspect whether translation was affected in the disease phenotype. Utilizing the rotenone model [187], we found increased LRRK2 activity in rotenone treated mid brain cultures, and 40% reduced translation in dopaminergic neurons. This reduction was prevented by adding MLi-2. Our models indicated that LRRK2-dependent translational reduction takes place in the cellular model of Parkinson's disease. We next checked whether LRRK2 activity contributed to neurite atrophy in our rotenone model and indeed rotenone induced die back of neurites, and LRRK2 inhibitors prevented this effect, thus indicating LRRK2

activity's involvement in neurite atrophy. We also checked whether LRRK2 action on translation and atrophy was due to ATP depletion in our rotenone model. Rotenone reduced ATP level in neurons, however, this reduction was not prevented with LRRK2 inhibitor treatment, prompting us to conclude that LRRK2 affects translation and atrophy either downstream of mitochondrial dysfunction, or independent of it.

3.3.1.3 Animal model validation

Next an animal model was utilized to investigate LRRK2 and translation in vivo. Rotenone treated rat brain was fractionated and LRRK2 was found to be enriched in the ribosomal fractions. Translation was repressed in the rotenone treated brain, indicated by increased expression of translation repressor 4E-BP1. Mass spectrometry was finally performed in this study on substantia nigra and striatum of the control and rotenone-treated rats (Figure 32), the resulting protein phosphorylation intensity data was analyzed with fold change and statistical calculations. The two regions of the brain were chosen because Parkinson's disease is characterized by the loss of the dopamine producing nerve cells in the midbrain, which encompasses substantia nigra and striatum.

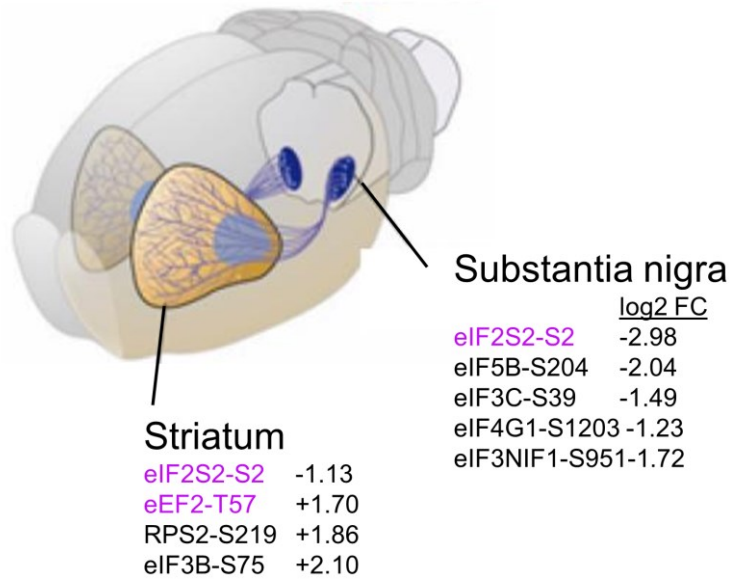

Figure 32. An illustration of the substantia nigra and the striatum in the rat brain, along with example fold change outputs [186].

The result of the analysis showed significantly changing phosphorylation in the rotenone model. Among the altered proteins are quite a few translation initiation and elongation regulators. A list is shown in Table 3. Among them, eIF2s2 phosphorylation was decreased on S2 in both regions. eIF2s2 is a rate-limiting translation initiation factor and the loss of phosphorylation on eIF2s2 prevents translation [188]. Another rate limiting protein, eEF2, underwent increased phosphorylation on site T57, which inactivated this elongation factor and repressed translation [189]–[192].

Table 3. A list of significantly differentially phosphorylated translation initiation and elongation regulators from the rotenone model.

UniProt ID	Sequence	Score	E-value	Phosphorylation residue	
Eukaryotic translation initiation factors				Rat	Human
Q64350	eIF2B 5 AGSPQLDDIR	41.16	0.0045	S539	S544
D4A554	Eif4g3 TSSPTTLPPLAR	31.52	0.0025	S305	S267
Q4G061	eIF3b AEEEGGSDGSAAEAEPR	72.82	1.50E-06	S114	N131
Q4G061	eIF3b AKPAAQSEEETAASPAASPTPQSAQEPSA PGK	67.22	4.10E-06	S75/S79	S81/S85
40s Ribosomal proteins					
P62754	RpS6 RLSSLRASTSK	14.31	0.067	S240	S240
P62908	RpS3 DEILPTTPISEQK	13.55	0.0069	S224	S224
P15880	RpS2 GGFGSGLR	32	5.00E-03	S31	S31
P15880	RpS2 GTGTVSAPVPK	38	4.90E-03	T202	T202

We examined whether these translation regulators were LRRK2 dependent by adding MLi-2 to rotenone treated midbrain culture and employing phospho-specific antibody to measure eEF2-T57 and eIF2alpha-S52 phosphorylation changes. In both cases, adding MLi-2 prevented the phosphorylation alteration from rotenone treatment, leading us to conclude that LRRK2 is actively involved in alternating the phosphorylation of translation regulators, which led to protein synthesis arrest.

3.3.1.4 Patient sample examination

In addition to cellular and animal models, we also examined patient fibroblast samples as LRRK2 expression is not limited to the brain. Sporadic and G2019S Parkinson's patient data was obtained from the National Institute of Neurological Disorders (NINDS) repository and Telethon Network of Genetics Biobanks (TNGB). We found global protein synthesis reduced by >40% in both sporadic and G2019S patients, and this decrease was reversed by treatment with LRRK2 inhibitors. To validate that LRRK2 not only reduces translation in G2019S cases but also reduces translation in sporadic cases, we further collected skin punches from 13 sporadic Parkinson's patients from Turku University Hospital (TUH) and matched them with seven controls of corresponding age. Upon calculation, the global protein synthesis of the TUH patients was significantly reduced. (Figure 33)

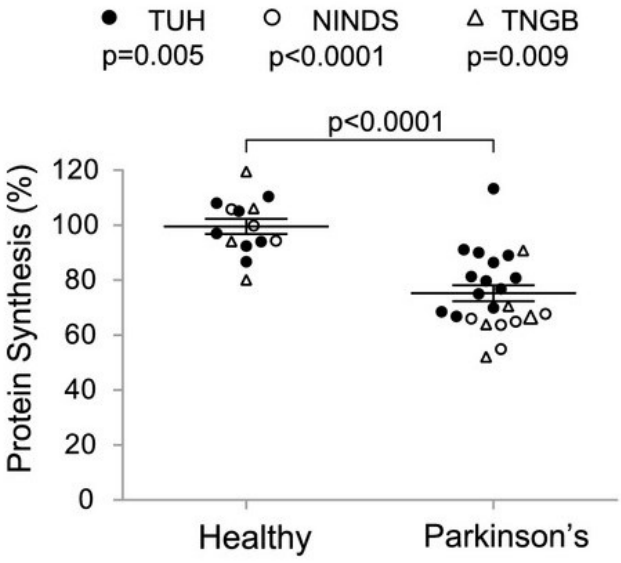

Figure 33. This plot shows the statistical results of protein synthesis analysis in cells isolated from skin biopsies of early stage Parkinson's patients and healthy volunteers [186]. The graph displays protein synthesis level of each individual, and the statistical significance between healthy and Parkinson's patients is determined using a Student's t-test. The results indicate that global protein synthesis was reduced in the patient group when analyzed alone (P = 0.005, size effect = 1.56, power = 0.88; black circles) or when combined with NINDS and TNGB cohorts (P < 0.0001, size effect = 1.91, power = 0.99).

From these results we believe repressed protein synthesis serves as a good biomarker of Parkinson's disease, even in early stage, as many TUH patients were not fully diagnosed at the time when we took the skin punch samples. It is also quite specific to Parkinson's disease. Atypical Parkinsonian disorders such as multiple system atrophy (MSA) or progressive supranuclear palsy did not show repressed protein synthesis from our examination. Additional analysis enabled us to characterize repressed translation further. We have established a correlating relationship between repressed translation and LRRK2-S935 phosphorylation, and a negative correlating relationship with age, but only for patients older than 60 years. This is consistent with LRRK2-G2019S action in late onset Parkinson's disease, which accounts for most cases.

3.3.2 Follow up study with fibroblasts from patients with sporadic and LRRK2-G2019S Parkinson's disease (Study III)

3.3.2.1 Patient fibroblast study introduction

Comprehensive testing from our study has established that de novo protein synthesis is repressed in both sporadic and LRRK2-G2019S Parkinson's patients, and this is detectable in fibroblast tissues. The result of this study piqued our interest on the subject matter, and we conducted a second study to answer more specific questions concerning reduced translation in fibroblast issues of Parkinson's patients, such as whether there are individual proteins with significantly altered overall translation pattern, and whether the pattern remains the same or differs in sporadic and LRRK2-G2019S Parkinson's.

3.3.2.2 MS study of de novo synthesis alterations in sporadic and LRRK2-G2019S Parkinson's patients

First, we labeled cultured cells from sporadic and LRRK2-G2019S Parkinson's patients and healthy controls utilizing the FUNCAT method so that de novo synthesized proteins are marked by fluorescence and can be quantified based on intensity. Comparing patient groups to control group, bulk de novo synthesis was reduced in both sporadic and LRRK2-G2019S Parkinson's patients. We then aimed to identify individual proteins that were differentially regulated at the level of translation in patient groups. We took skin punch samples from ten sporadic Parkinson's patients and six healthy donors from TUH, and five LRRK2-G2019S Parkinson's patients and six healthy donors from NINDS and TNGB, labeled newly translated proteins with bio-orthogonal non-canonical amino acid tagging (BONCAT) method, and isolated these proteins for mass spectrometry analysis.

Following MaxQuant spectral analysis, the data was analyzed through PhosPiR, where statistical tests and enrichment analysis were performed. We identified 33 and 30 nascent proteins with reduced synthesis in sporadic and LRRK2-G2019S Parkinson's cases, respectively (Figure 34). 65% of the significantly differentially synthesized proteins overlap between sporadic and LRRK2-G2019S Parkinson's. The enrichment result of the significantly differentially synthesized proteins revealed that the biological process "cytosolic signal recognition particle (SRP)-dependent co-translational protein targeting to membrane" was functionally significantly affected in both sporadic and LRRK2-G2019S Parkinson's. This process regulates the translation of secretory pathway proteins and their translocation to the endoplasmic reticulum (ER). On the other hand, "Tubulin/FTsz C-terminal

domain superfamily network" was only significantly enriched in LRRK2-G2019S Parkinson's, which supports LRRK2's well known association with microtubules.

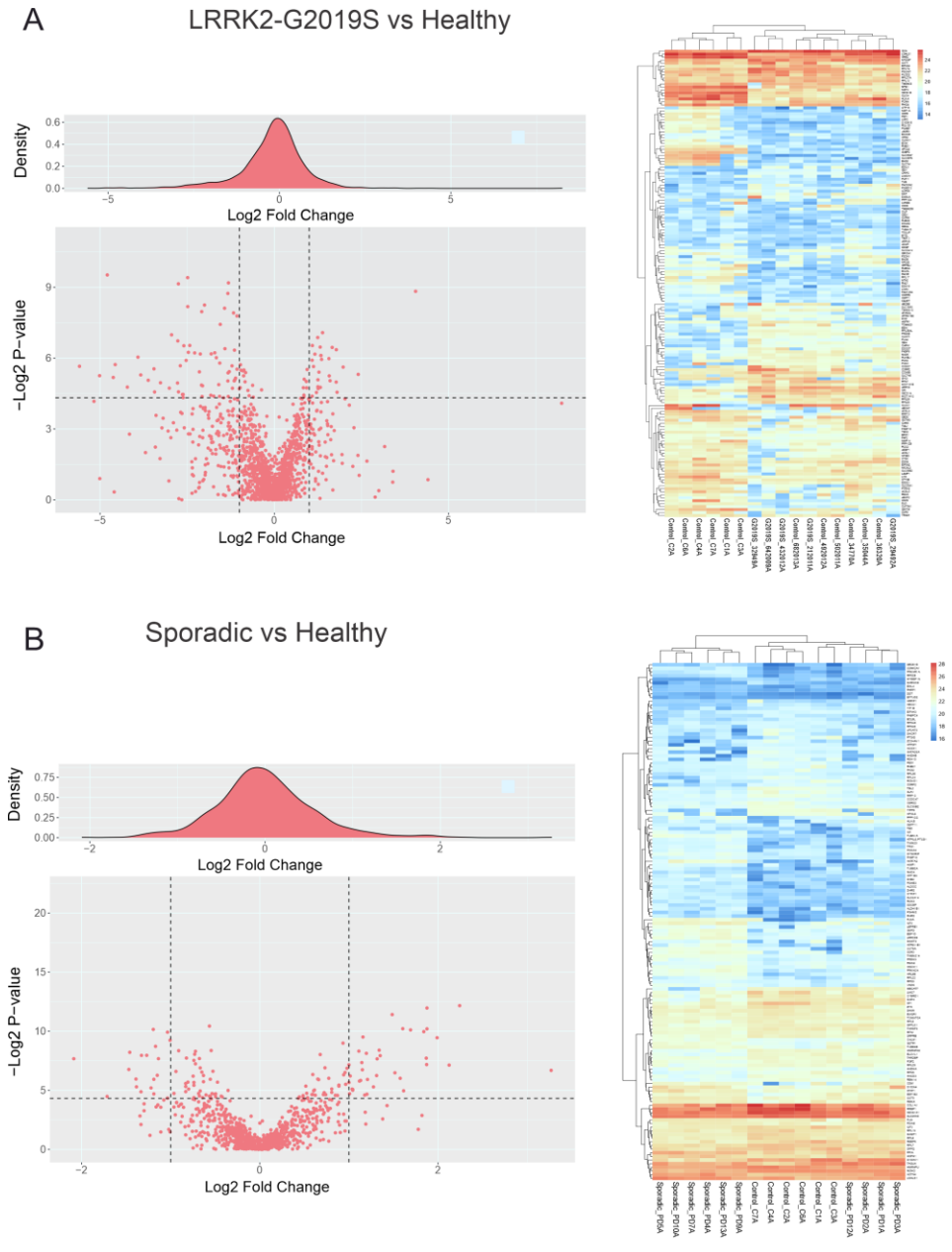

Figure 34. This figure shows the results of comparing AHA-labelled protein intensities between LRKK2-G2019S Parkinson's and healthy individuals, as well as sporadic Parkinson's and healthy individuals [193]. A shows a volcano plot of all AHA-labelled protein intensities for LRKK2-G2019S versus healthy. The heat map with hierarchical clustering depicts the regulation of nascent protein levels in fibroblasts from LRRK2-G2019S patients compared to healthy individuals using the union of ROTS and t-test with a p-value <0.05. B shows a volcano plot of all AHA-labelled protein intensities for sporadic cases versus healthy using the ROTS statistical test, and the heat map with hierarchical clustering depicts the nascent proteins in the same format as A.

3.3.2.3 Total lysate validation of significantly altered protein expressions

The mass spectrometry analysis of de novo synthesis identified a list of significantly differentially translated proteins. To further examine whether these proteins were homeostatically disturbed in total cell lysate, we employed targeted proteomics to measure changes specifically from these proteins. We incorporated significant proteins from two statistical tests, Student's T-test and ROTS, with each test evaluating two sets of input data, protein intensity data normalized with MaxQaunt LFQ method, and with or without imputation. A total of 247 proteins each from sporadic and LRRK2-G2019S Parkinson's and healthy control were measured with PRM-analysis. Statistical tests and enrichment analysis were performed again. In LRRK2-G2019S Parkinson's cases, all targeted proteins showed lower level of expression from total cell lysate, without exception (Figure 35). In sporadic Parkinson's cases, almost all targeted proteins showed decreased expression, however, to a lesser extent compared to LRRK2-G2019S Parkinson's cases. We compared the proteins significantly reduced in expression from sporadic and LRRK2-S2019S Parkinson's cases and found out the majority from both groups overlap with each other, as shown in Figure 36's venn diagram. From enrichment results, "mRNA splicing" and "pre-ribosome and ribosome biogenesis" are the most enriched functions in LRRK2-S2019S Parkinson's, while "viral mRNA translation", "peptide chain elongation" and "ribosome KEGG pathway" are the most enriched associations in sporadic Parkinson's.

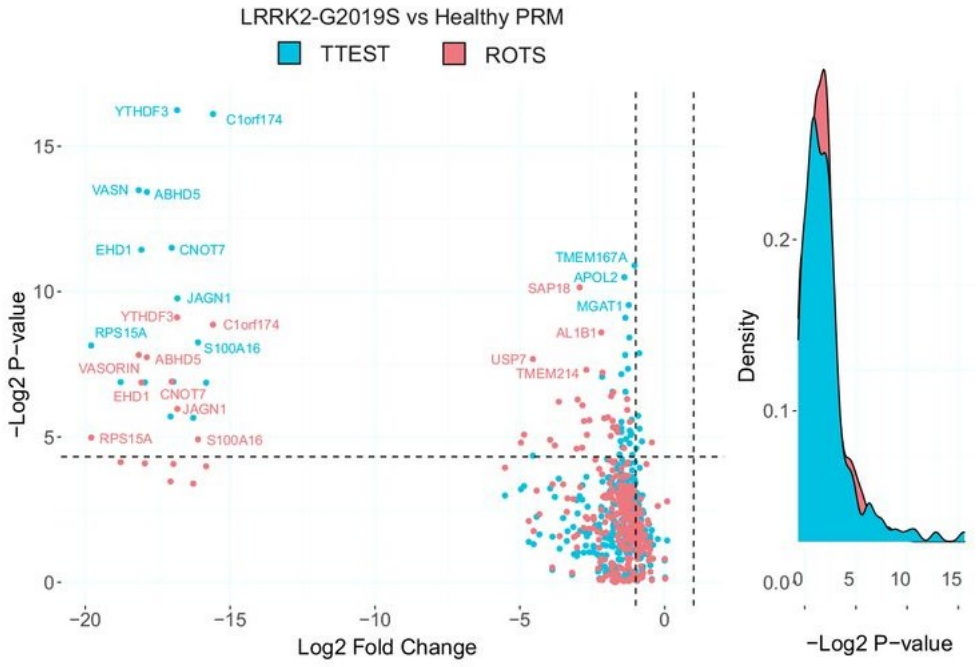

Figure 35. Volcano plot showing the fold change and p-value of comparing LRRK2-G2019S Parkinson's cases to healthy controls for the targeted total cell lysate MS data [193]. All fold changes are negative without exception.

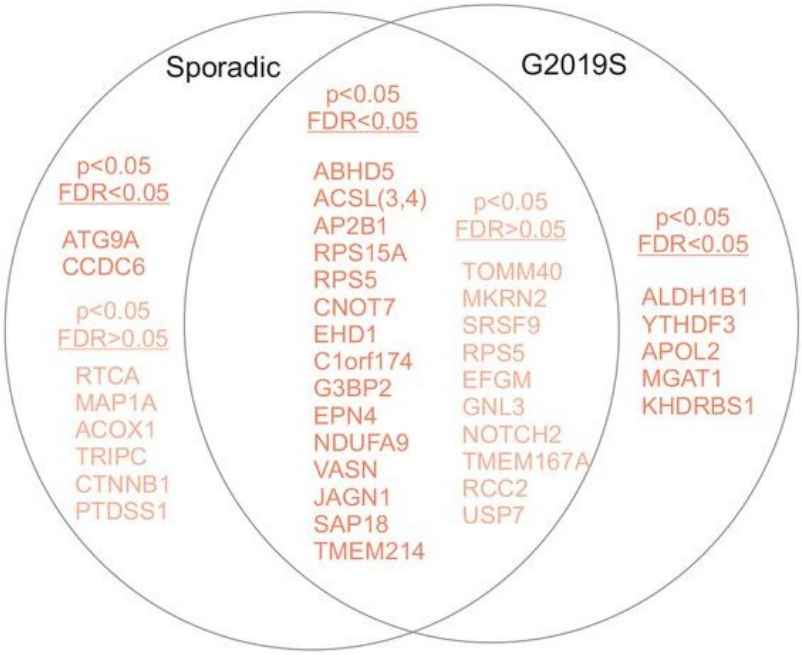

Figure 36. Venn diagram showing overlapping significantly reduced proteins between sporadic and LRRK2-S2019S Parkinson's cases [193].

3.3.2.4 mRNA level inspection of altered protein expressions

To confirm that the protein expression changes in patient cells were not due to altered mRNA levels, but rather from a post transcriptional step, we performed quantitative PCR on the mRNA of the significantly differentially expressed proteins. The result showed no significant changes in the mRNA levels, and thus validated our hypothesis.

3.4 JNK and schizophrenia (Study IV)

3.4.1 Phosphoproteomics study of *Jnk1-/-* mice brain

3.4.1.1 A brief method overview

Past studies have drawn relevance between schizophrenia and JNK function and proteins from the JNK signaling transduction cascade [65], [194], [195]. To further explore schizophrenia mechanistic insight through JNK activities, we prepared wild type and *Jnk1-/-* mice from four age groups, embryonic day 15 (E15), post-natal day zero (P0), post-natal day 21 (P21) and eight months (Adult), each with three

replicates, and we performed whole brain LC-MS/MS analysis to obtain the phosphorylation profile of *Jnk1-/-* and wild type mice brain.

The raw spectral data was analyzed with Progenesis software to identify and quantify phosphopeptides found from our brain samples. Output data from Progenesis was pre-processed in the following way: i) data from various gel slices were merged into one dataset, ii) entries with the same peptide sequence, UniProt ID, and number of phosphorylations were merged into one entry by taking the sum of intensities of all entries, and iii) for any age group, if less than one missing value were present between three replicates, the entire phosphopeptide entry was removed. The preprocessed data was then analyzed with plots from "Overview figures" section in the method description, rank product statistical test in R, and enrichment analysis utilizing MetaCore. We compared our significant results to MetaCore and SCHEMA lists for schizophrenia associated genes and generated two lists of phosphoproteins that were significantly differentially phosphorylated in *Jnk1-/-* brain and overlapping with MetaCore or SCHEMA lists. The two lists went through network analysis utilizing GeneMANIA database, then Fisher's exact test was performed to check for association between increased network connectivity and schizophrenia linked genes according to SCHEMA. Cellular and behavioral validation experiments were subsequently performed to better apprehend findings from the bioinformatics analysis.

3.4.1.2 Analysis result summary

An overview of the phosphoproteome data revealed ten percent of the detected phosphorylations were significantly altered in *Jnk1-/-* brain, indicating hub effect from JNK1 (Figure 37) network pic of dark background). Between the four age groups, there were both overlapping and unique alterations; heatmap showed a visible divergence in phosphorylation between the developing brain and the mature brain (Figure 38). Enrichment analysis of significantly differentially phosphorylated proteins revealed the cellular processes "cell adhesion" and "synaptic contact" are enriched, and in disease biomarker category, schizophrenia was highly enriched, followed by amyotrophic lateral sclerosis (ALS; motor neuron disease) and Parkinson's disease.

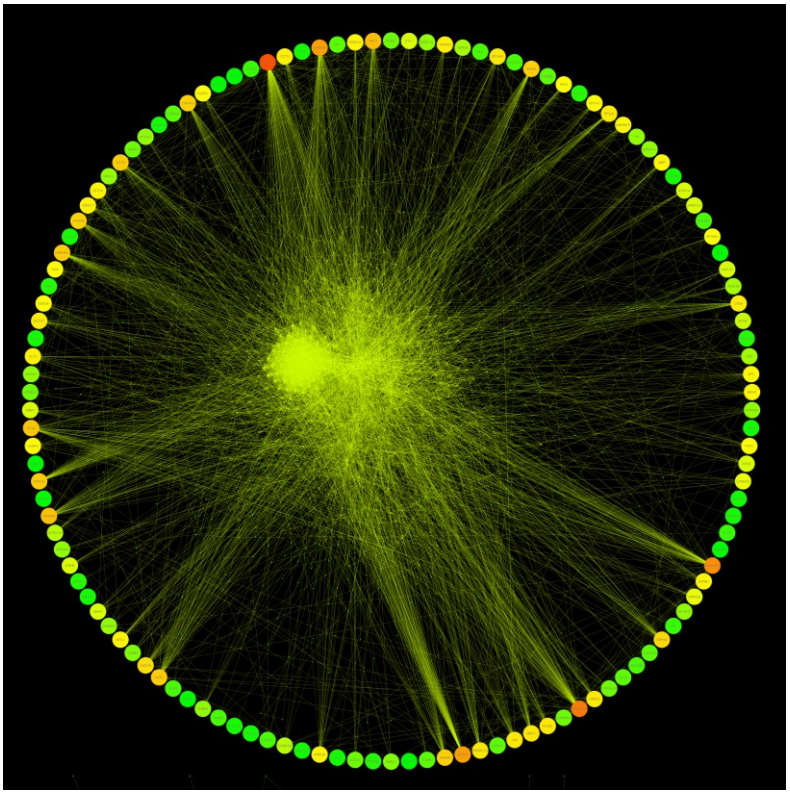

Figure 37. A network figure where the outer circle is showing significantly changing phosphoproteins from *Jnk1-/-* versus wild type mouse brain, and the proteins interacting directly and indirectly with them are shown inside the circle. The extensive network of interactions indicates JNK's effectiveness as a hub.

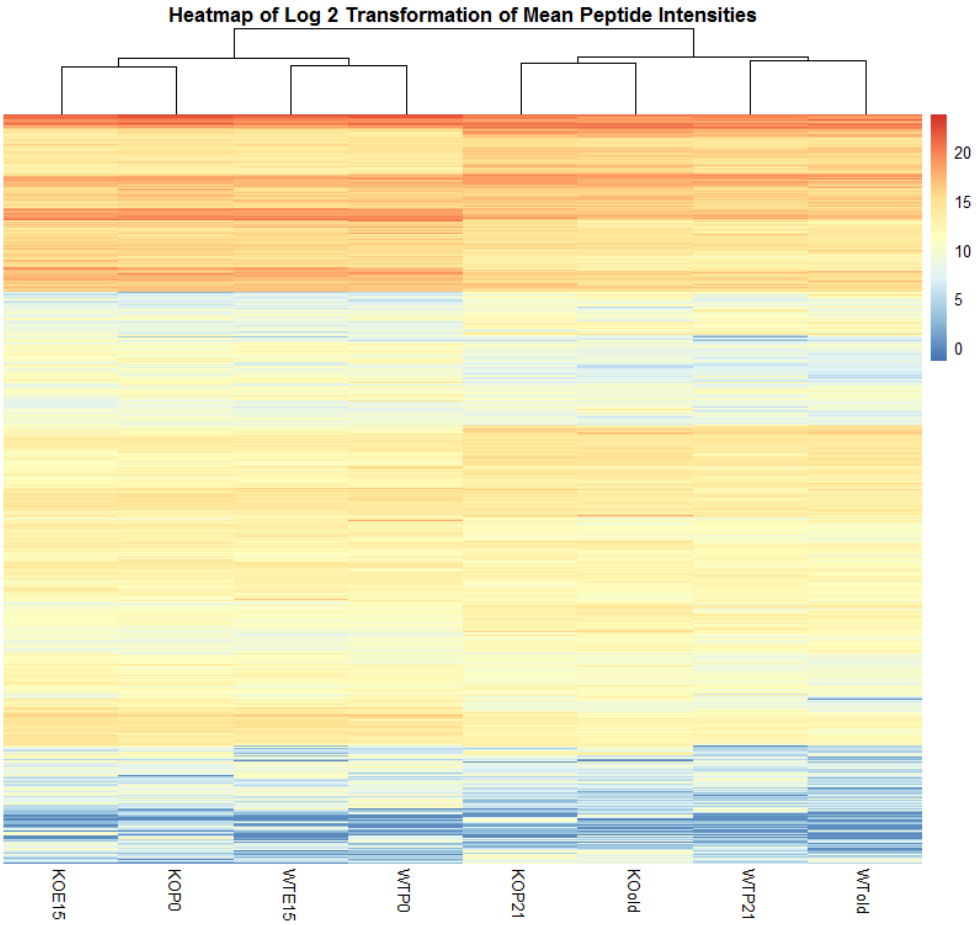

Figure 38. Heatmap of *Jnk1-/-* and wildtype mice MS intensity data. A clear visual divergence can be seen between young (embronic day 15 and post-natal day 0) and mature (post-natal day 21 and 7 months) brain.

126 significant entries overlapped with MetaCore schizophrenia gene list, most were kinases and phosphatases, and cytoskeletal proteins. Known JNK substrates such as microtubule-associated proteins (MAPs) were significantly differentially phosphorylated in all age groups, while calcium channel proteins only had phosphorylation altered in younger mice. Other prominent schizophrenia associated proteins in the list include CAMK2B, CAMKKI, NMDA receptor subunit Grin2A, 14-3-3 scaffold proteins and Src. 113 significant entries overlapped with schizophrenia Exome Meta-Analysis (SCHEMA) consortium risk genes, among them are TRIO, RB1CC1 and GluN2A to name a few.

Protein interaction networks were built from GeneMANIA interaction database. The significant schizophrenia associated gene lists used as were input data. To check whether interaction between the genes from the list were significantly more than expected by chance, 10,000 background networks were generated and a p-value was calculated for each interaction network in accordance to the description in the Method section. We found that the physical interaction network and the colocalization network had significantly increased interactions, while the genetic interaction network did not. Individual gene's interactions were also examined with Fisher's exact test, which determines whether there is a significant difference between the number of interactions of a particular gene from the significant schizophrenia network, and the number of interactions this gene encounters in its general facility. The result identified 14 genes with significantly higher interactions associated with the schizophrenia risk gene network, they were HIVEP2, AKT1, GRIN2A, GRIN2B, EIF4G1, ATP1A1, SHANK1, HSP90AA1, SRPK1, DLGAP1, NLGN3, EIF4G3, HUWE1 and NBEA.

We then investigated *Jnk1-/-* functional changes with pathway enrichment analysis. "NMDAR trafficking" pathway was highly enriched. Significantly phosphorylated proteins from the pathway included NMDAR subunits GluN2A and GluN2B, metabotropic receptor (mGluR5) families, and MAGUK family proteins, which regulates NMDARs at the plasma membrane [196]. Various other proteins that regulate surface expression and endocytosis of receptors were also significantly phosphorylated. This finding was significant as "NMDAR hypofunction" is believed to contribute to psychosis and cognitive problems in schizophrenia. Among the NMDAR trafficking pathway proteins identified, PKCε piqued our interest as it had the largest fold change difference in phosphorylation between wildtype and *Jnk1-/-* brain. Other isoforms of this protein were also significantly differentially phosphorylated in *Jnk1-/-* brain, leading us to hypothesize that JNK1 regulates PKCε through the classical sequential activation of PKCs to exert control over the NMDAR trafficking pathway. We conducted wet lab experiments to confirm these bioinformatics results and test our hypothesis.

3.4.2 Wet lab validation of MS analysis results

3.4.2.1 Neuron surface staining of NMDAR and GABAA subunits

Surface staining of NMDAR subunits GluN2A and GluN2B in neurons showed that surface expression of both subunits in *Jnk1-/-* neurons was significantly decreased while the overall expression remained unchanged. After applying PKC inhibitor bisindoylmaleimide-1, the surface expression changes were reduced particularly in GluN2B. GABAergic neurotransmission was reported being disturbed in

schizophrenia [197], additionally, it was significantly enriched in the cellular process enrichment analysis of *Jnk1-/-* brain, we thus examined GABAA subunits $GABA_A\alpha1$ and $GABA_A\beta3$ for their surface expression in *Jnk1-/-* neurons. Surface accumulation of the subunits were detected, and following one hour treatment of bisindoylmaleidide-1, the accumulation was reversed. We then examined whether JNK inhibitor on wild type neurons would alter the surface expression of these biomarkers. While inhibitor DJNKI significantly increased $GABA_A\alpha1$ and $GABA_A\beta3$ surface expression, NMDAR subunits did not show altered surface expression compared to control, which suggests NMDAR surface expression in postnatal neurons could be dependent on prenatal JNK1 influence. Together the wet lab experiments confirmed that when the *Jnk1* pathway is disturbed during neurodevelopment, NMDAR subunits increase in expression levels while GABAA subunits decrease in expression levels at the neuron surface, and these alterations may be partly dependent on PKC regulations.

3.4.2.2 Animal model behavior profiling

To inspect the symptomatic effect of the NMDAR related alterations we discovered in *Jnk1-/-* brain, and whether they align with schizophrenia behavior profiles, we conducted behavioral tests with MK801 treatment. MK801 is a psychotomimetic drug that blocks NMDARs and known to model positive and cognitive symptoms in schizophrenia [198], [199]. In the open field test, *Jnk1-/-* mice displayed increased locomotion and increased frequency of entering the center of the arena. After injecting MK801, both wildtype and *Jnk1-/-* mice displayed increased traveling compared to before injection, with *Jnk1-/-* mice displaying substantially greater hyperactivity response than wildtype. In the Y-maze test, both wildtype and *Jnk1-/-* mice performed similarly for working memory before and after the injection, however, *Jnk1-/-* mice displayed significantly increased stereotypies than wildtype after the saline I.P. injection.

We next measured a feature called “sensory gating”. This is a process whereby irrelevant processes are separated from meaningful ones. A sensory gating deficit is often found in schizophrenia and post-traumatic stress disorder (PTSD). It can be tested in humans and in mice using the pre-pulse inhibition test. We therefore performed a paired pulse inhibition (PPI) test on *Jnk1-/-* mice. A pre-pulse inhibition of the startle reflex response occurs when a weak pre-stimulus is applied before a close follow up strong sensory stimulus, and therefore inhibits the response to the second stimulus. Such inhibition is impaired in schizophrenia, making it a favored test to determine if an animal model is relevant for schizophrenia. The result of PPI test indicated significant reduction in PPI after MK801 treatment in wildtype mice as expected, however, in *Jnk1-/-* mice, the baseline PPI level was already reduced,

and MK801 treatment did not yield significant reduction from baseline. We ruled out genotype specific sensitivity difference to acoustic startle as the source of the baseline reduction by testing a range of pulse intensities and compared startle responses between *Jnk1-/-* and wildtype mice, and no significant differences were found. We then performed PPI test on a separate D-amphetamine model, which impairs the PPI response through dopamine interaction. Once again, the treatment response was not significantly reduced compared to the baseline due to *Jnk1-/-* mice having a reduced baseline PPI. From these tests we concluded that *Jnk1-/-* mice display reduced PPI at the baseline level similar to wildtype mice following the MK801 treatment. Together these findings associate the *Jnk1-/-* molecular and behavioral phenotype with schizophrenia and neuropsychiatric disease.

4 Discussion

4.1 Combining phosphoproteome data and proteome data in the study

The studies included in my thesis focus on both proteome and phosphoproteome analysis to study kinases and their pathological impact in the brain and in patient samples. Both the JNK and LRRK2-G2019S Parkinson's studies revolve around kinase activity alterations that shifted the homeostasis of brain phosphoproteomes, resulting in both expression and activity variations of a series of proteins, and impacting the pathological mechanisms of the neurological disorders. Although the direct effect of kinase alteration is phosphorylation changes, these phosphorylation changes in turn influence protein expression, hence proteome data analysis is a necessary complement to the kinase study, especially when recent technology has enabled the quantification of both proteomic and phosphoproteomic entries in one MS run, making it easier to obtain and analyze both datasets. Phosphoproteome data allows the identification of protein activity and network regulation changes, while proteome data reflects the downstream expression alterations of the directly and indirectly regulated proteins. Together, comprehensive understanding of the full impact of the target kinase can be achieved.

4.2 Technological and methodological improvements over time

4.2.1 Study of *Jnk-/-* – the chronological first study

During the progression of the thesis studies, MS technology has been advancing rapidly, as have the analysis methods available. Chronologically, the *Jnk1-/-* brain dataset was the first to be analyzed for this thesis (referred to in the thesis as original manuscript IV). As the oldest dataset in the study sequence, it was generated (with HPLC system coupled to ThermoFisher Scientific LTQ-Orbitrap XL mass spectrometer operated) in data-dependent-acquisition (DDA) mode. At the time of the data generation, data-independent-acquisition (DIA) method was still premature in its development stage. Preprocessing work done was elementary compared to later

datasets, however, it was not necessarily inferior to later analysis. Data quality was first checked with boxplot and histogram, after confirming the sample distributions were closely aligned, normalization was skipped. A strict entry removal strategy upon encountering NAs have preserved the integrity of the non-NA intensity values with minimized influence from NA and calculations involving NA, even though NA imputation method employed was the simple system of replacing with one unanimously. In retrospect the preprocessing made sense, as utilizing an incompatible imputation method, for example, could have potentially yielded invalid results. Choosing simpler methods instead retains the original information from the data, and combined with a reliable NA removal strategy, the data quality can be decent. The only drawback would be losing a portion of the data from NA row-filtering. Thankfully from the results it seems that important mechanism relevant changes remained.

4.2.2 Initial LRRK2 and Parkinson's study – the next study in chronological order

The next datasets, following up from the previous study, were from the first LRRK2 and Parkinson's study (referred to in the thesis as original publication II). The study itself was heavy on experimental data with less focus on bioinformatics analysis. However, it was a crucial study that laid the groundwork for the follow up studies and patents on Parkinson's biomarker discovery and validation, though most of these studies that I have also participated extensively in are outside of the scope of this thesis unfortunately.

The dataset from this study contained protein phosphorylation intensities from substantia nigra and striatum of rotenone-treated rats and control. For this data, we applied median normalization, then imputed the missing values with Perseus imputation method, where random values were drawn from a normal distribution with a down shift. Median normalization is a rather reserved method with minor adjustment to the data. We preferred this method over more dramatic methods such as quantile normalization, to stay true to the original data. Unlike proteomics data, where the normal practice is to equalize the total protein inserted into each sample, and in turn expect the resulting intensity distribution to reflect this setup, phosphorylation intensities can vary in total amount between samples even with equalized total protein level in the samples. This means that the entire phosphorylation distribution can shift based on sample treatment. Quantile normalization, as an example, would mask such changes and possibly yield inaccurate results in this case. Perseus imputation method assumes that the data is normally distributed, and the missing values are localized to the lower abundance spectra [200]. Both assumptions are suitable for our dataset. We inspected the data

distribution with histogram and Q-Q plot, and it was normally distributed when excluding the missing values (data was log2 transformed during inspection and when applying the Perseus imputation). We have also examined missing value distributions from example phosphoproteome dataset and have found missing values to be associated more with lower abundance protein or peptide entries. Therefore, we could reasonably expect Perseus imputation to replace missing values with sensible intensity values.

4.2.3 The follow up Parkinson's study – the latest study

The most recent datasets analyzed in this thesis were fibroblast samples from LRRK2-G2019S and sporadic Parkinson's patients and healthy control (the study is referred to in the thesis as original publication III). For AHA labelled samples, the data was normalized with MaxLFQ method during the quantification stage, and the resulting dataset was imputed with Perseus imputation. The MaxLFQ method was designed to accurately determine the relative abundance of proteins in two or more samples based on the chromatographic ion intensities. The method recognizes biases introduced by sample fractionation and corrects it by applying an optimization algorithm on the total protein calculation equation where the normalization factors are set as variables [201], [202]. The MaxLFQ method assumes that most proteins exhibit minimal or no changes between conditions, and no more than one third of the proteome is altered [201]. Looking at the volcano plot of the Parkinson's versus control comparisons utilizing quantified intensity data, the number of significant changes did match the assumption of the MaxLFQ method. However, the PRM validation dataset resulted in a rather more significant outcome for the Parkinson's versus control comparisons, leaving us with a reasonable doubt that MaxLFQ method in this case might have masked weaker protein changes. Our result remains valid, however, as all significant results from the AHA labeling dataset continued to be significant in our PRM validation dataset. The PRM dataset employed targeted proteomics and is therefore without the drawbacks of the label-free method, hence we simply replaced NA by one for preprocessing.

4.3 Downstream bioinformatics methods discussion

Various post analyses were performed for the MS datasets in this thesis. Among them, statistics analysis followed by enrichment analysis were most useful for our study goals. Significantly altered protein expression or phosphorylation identified from statistics analysis directed us to the target group of our interest, and enrichment

analysis revealed localization and functional information of these targets, leading us to the mechanistic clues of the disease states.

There are numerous statistical methods which calculate the significance of variation between two or more distributions, and that number is ever increasing as the field advances. PhosPiR alone included four choices of statistical tests, choosing the most suitable one became an inevitable task in the analysis sequence. A good way to make a choice would be to look at benchmarking studies which focus solely on comparing a set of tools with the same purpose and operating on one or more types of data. Besides statistical tests, benchmarking studies are also good references for selecting normalization and imputation methods. The best performing tools in the studies are usually safe choices for the same type of data included in the study. A recent benchmarking study by Miao-Hsia Lin et al. examined methods specifically for differential expression, imputation, and quantification for proteomics data [203]. Among the best performing tools concluded from this study are MaxQuant LFQ for intensity generation, Perseus imputation for pre-processing, and ROTS for differential analysis. Interestingly, we have also selected these tools for our LRRK2 studies. The validation from this benchmarking study adds another layer of confidence to our analysis methods selection.

Besides benchmarking studies, the best method choices are made from an in-depth understanding of one's data. For example, if one of the distributions being compared have two extreme values that are a lot higher than all the other values from the same distribution, this would shift the mean of the distribution disproportionately to the right, and this bias would be incorporated into any statistical tests that employs the distribution mean. Hence in this case, it is better to choose a statistical test that does not include distribution mean as part of the calculation, such as ranking tests. Different statistical tests make different assumptions about the data, some assumes normal distribution, some assumes at least 50% of each distribution would be non-missing values. When the data at hand does not match the method's assumptions, the method would not be a good fit for the data.

Enrichment analysis links the data at hand to known knowledge through various databases. These databases on one hand provides very useful information for the analysis, on the other hand, however, can subject the data to a few drawbacks. One of them is Information source bias. As more information are generated from more poplar studies such as cancer research, general purpose databases are usually filled with information from these areas. Caner pathology of course cannot be applied to many other fields of studies; hence one needs to be careful of the information source of the databases utilized for each study. Another drawback of databases is the customary identification (ID) codes which only have meaning for a specific database rather than universal. The excessive number of customary IDs can create a barrier in uniting different type of knowledges, especially when converting one to another, it

is rarely a one-to-one match. MetaCore database, for example, can match their network object code to several proteins or isoforms, while a single protein could match to more than one network objects. It gets worse for cross species comparisons, multiple match or unavailability of information are common results. It would be a good idea to have studies delicate to improving inter-database linkage.

4.4 Thoughts on PhosPiR

4.4.1 Initial aspirations

Original publication I of this thesis describes a tool called PhosPiR, which automatically performs a range of proteomic and phosphoprotemic analyses. The PhosPiR tool was developed from the phosphoproteomics and proteomics analyses that were performed throughout my doctoral training.

Coffey group is more biological oriented; besides my own studies, one part of my task, which I enjoy, is to automate some of the tedious calculations or analyses my colleagues had to perform on their data. This manual work can consume considerable time, impinging on advancement of wet lab experiments. To replace repetitive manual work with automated calculations, we would schedule meetings to discuss in detail the type of data they work with, the desired data processing to be conducted, and the type of analysis that would be most suitable for their hypothesis. Then I would proceed to design and write a code that satisfied all the points from the discussion, with a simple graphical user interface (GUI) to guide the analysis steps. In the process of interacting with my colleagues and working together to achieve better efficiency through automation, I recognized there is an unfilled niche, where scientists without coding knowledge could benefit greatly from the vast range of tools that R or Python packages can offer, yet they lacked the means to access these tools.

Thus, PhosPiR was developed with the following goals 1) To implement proteomics and phosphoproteomics analysis methods that I have utilized to analyze various MS brain data; and 2) To make the workflow automated to a point where scientists without coding knowledge could also perform analysis with the R functions offered in the workflow.

4.4.2 Strengths and weaknesses

PhosPiR pipeline was created toward the end of my thesis studies, however, it has proven useful in the follow up studies for Parkinson's disease and other projects we work on. It is designed to be nonprogrammer friendly and proteomics beginner friendly. By following the GUI of PhosPiR, a series of useful analyses for proteomic

and phosphoproteomic data are performed even if the user is not familiar with some of the analyses. This design distinguishes PhosPiR from peer analysis tools, and is recommended for anyone working with a proteomics or phosphoproteomics dataset due to the low technical threshold.

Ming-Xiao Zhao et al. have included PhosPiR in their review of phosphorylation database and prediction tools alongside other phosphorylation predictors such as NetPhos, however, the evaluation did not adequately represent PhosPiR's capabilities, as only the KinSwingR tool was assessed as a phosphorylation predictor, while the PTM-SEA and kinase network components were not included [204]. It could be difficult to classify PhosPiR's functionalities into a single category due to it being an integrated pipeline and encompassing a diverse array of features derived from multiple stages of data analysis. Comparing to other R based tools, however, PhosPiR's inclusion of both data preprocessing tools and downstream analysis tools in one pipeline is novel and provides more convenience for the user by offering all analysis steps in one go. Comparing to software-based tools, PhosPiR brings to the table a means for non-programmers to utilize excellent analysis tools from R. Analysis methods such as ROTS statistical test, rank product statistical test, PTM-SEA, and more are only available as function implementations in R to date.

Despite its benefits, PhosPiR also have some limitations that should be considered. Firstly, as an automated pipeline, the focus is on simplifying the analysis process and extensive customization options are not implemented. This means that the preprocessing power of PhosPiR is weaker compared to other analysis tools, as only general normalization and imputation methods are included. This may be insufficient for some datasets and users are encouraged to perform tailored pre-processing before inputting the data into PhosPiR. Additionally, if an error occurs during analysis, the user must restart the entire analysis from the beginning, which can be time-consuming. Although efforts have been made to lessen the setup work required for a rerun, such as recording group and group comparison setups and allowing the user to select the recordings in a new run, improvements are still needed in this aspect. Currently efforts are being made to improve PhosPiR's functionality with an improved rerun feature, where previous PhosPiR results can be selected during a rerun to skip any combinations of analysis already performed and included in the previous results.

Even though there are limitations, we believe PhosPiR remains a valuable tool for the analysis of phosphoproteomic and proteomic data, with its automated pipeline simplifying the analysis process and offering a range of analysis methods. To promote the use of PhosPiR, we have undertaken various advertising efforts. We have used social media platforms to share information about PhosPiR and its functionalities, and we plan to attend conferences to showcase its capabilities in poster presentations. This thesis serves as another platform to advertise PhosPiR.

However, we strongly believe that the best advertisement is providing high-quality service to our users. We have a reputation of providing immediate responses to user queries and actively assisting them in resolving any issues that may arise during their PhosPiR runs. We plan to continue this service and build strong relationships with our users. We feel a solid reputation for excellent service is the best way to gain more users for PhosPiR.

5 Summary/Conclusions

Phosphorylation is a highly prevalent and essential post-translational modification that plays a critical role in various biological processes. Dysregulation of phosphorylation signaling has been implicated in the pathogenesis of various neurological disorders, including chronic depression, Alzheimer's disease, Parkinson's disease, and schizophrenia. By Investigating the interplay between phosphorylation and changes in protein expression, underlying disease mechanisms can be elucidated and potential drug targets can be identified. The present thesis focuses on characterizing alterations in the phosphoproteome and protein abundance associated with two such disorders, schizophrenia and Parkinson's disease, with the aim of uncovering disease mechanisms and associated regulatory networks and pathways. To streamline the analysis process, an automated R pipeline was developed, integrating various analysis methods utilized from the previous studies as well as additional useful phosphoproteomics analysis methods, allowing users to save weeks of analysis work without requiring coding knowledge.

Prior studies have suggested an association between c-Jun N-terminal Kinase (JNK) and schizophrenia, but the underlying mechanism remains unclear. We have conducted a study (Study IV in the present thesis) which aimed to investigate the role of JNK1 in schizophrenia by analyzing the phosphorylation profile of wild type and *Jnk1-/-* mice from four age groups using LC-MS/MS analysis. The data was pre-processed, statistically analyzed and subjected to network analysis to identify significant differentially phosphorylated proteins associated with schizophrenia. Enrichment analysis revealed that cell adhesion and synaptic contact processes were enriched, and schizophrenia was highly enriched in the disease biomarker category. 126 proteins which were associated with schizophrenia overlapped with the significantly differentially phosphorylated proteins in *Jnk1-/-* mice brain, including kinases, phosphatases, cytoskeletal proteins, CAMK2B, CAMKKI, NMDA receptor subunit Grin2A, 14-3-3 scaffold proteins, Src, TRIO, RB1CC1, and GluN2A. Protein interaction networks were built from GeneMANIA interaction database to identify significant schizophrenia-associated entries from these phosphoproteins. While physical interaction network and colocalization network have significantly higher interactions compared to expected, genetic interaction network did not.

Fisher's exact test identified 14 genes with significantly higher interactions associated with the schizophrenia risk gene network, including HIVEP2, AKT1, GRIN2A, GRIN2B, EIF4G1, ATP1A1, SHANK1, HSP90AA1, SRPK1, DLGAP1, NLGN3, EIF4G3, HUWE1 and NBEA. Pathway enrichment identified the NMDAR trafficking pathway to be highly enriched, and surface staining of NMDAR subunits in neurons showed that surface expression of both subunits in *Jnk1-/-* neurons was significantly decreased. GABAergic neurotransmission was also significantly enriched in the cellular process enrichment analysis of *Jnk1-/-* brain, and the wet lab experiments confirmed that when the *Jnk1* pathway is disturbed during neurodevelopment, NMDAR subunits increase in expression levels while GABAA subunits decrease in expression levels at the neuron surface, and these alterations may be partly dependent on PKC regulations. Behavioral tests were conducted with MK801 treatment to investigate the symptomatic effect of the NMDAR related alterations and whether they align with schizophrenia behavior profiles. *Jnk1-/-* mice displayed increased locomotion and increased frequency of entering the center of the arena. In the Y-maze test, *Jnk1-/-* mice displayed significantly increased stereotypies than wildtype after the saline I.P. injection. The paired pulse inhibition (PPI) test indicated significant reduction in PPI after MK801 treatment in wildtype mice as expected, however, in *Jnk1-/-* mice, the baseline PPI level was already reduced, and MK801 treatment did not yield significant reduction from baseline. From these tests, it is concluded that *Jnk1-/-* mice display reduced PPI at the baseline level similar to wildtype mice following the MK801 treatment, associating the *Jnk1-/-* molecular and behavioral phenotype with schizophrenia and neuropsychiatric disease. The results of our study contribute to a better understanding of the molecular mechanisms underlying schizophrenia and provide novel insights into potential targets for future research. Our identification of protein targets and pathways that contribute to schizophrenia phenotypic symptoms suggests that disruption in JNK regulation may play a role in the symptomatic progression of the disorder. These findings could help the development of more effective therapies for schizophrenia.

The LRRK2-G2019S mutation is one of the most frequent genetic causes of late onset Parkinson's disease. We have conducted two studies to understand its role and the effect of the G2019S mutation, which could aid in uncovering the disease's pathological mechanism. In the first study (Study II in the present thesis), rat brain was separated into fractions and each fraction was phosphorylated *in vitro* with purified LRRK2-G2019S to identify in which region of the brain LRRK2-G2019S function is most active. It was found that LRRK2 is localized to the small 40S ribosomal subunit and that LRRK2 activity suppresses RNA translation. Further tests were conducted to validate the findings, and they demonstrated that LRRK2-dependent translational reduction takes place in all tested models of Parkinson's disease, including Parkinson's patient fibroblast samples. The second study (Study

III in the present thesis) was conducted with the aim to identify individual proteins with significantly altered overall translation patterns in sporadic and LRRK2-G2019S Parkinson's patients. Newly translated proteins were labeled with bio-orthogonal non-canonical amino acid tagging and isolated for mass spectrometry analysis. The data was analyzed to identify 33 and 30 nascent proteins with reduced synthesis in sporadic and LRRK2-G2019S Parkinson's cases, respectively. The analysis revealed that the biological process "cytosolic signal recognition particle (SRP)-dependent co-translational protein targeting to membrane" was functionally significantly affected in both sporadic and LRRK2-G2019S Parkinson's, while "Tubulin/FTsz C-terminal domain superfamily network" was only significantly enriched in LRRK2-G2019S Parkinson's. The study also used targeted proteomics to measure changes in protein expression from total cell lysate, which showed lower levels of expression in both LRRK2-G2019S Parkinson's cases and sporadic Parkinson's cases. The identification of reduced protein synthesis in sporadic and LRRK2-G2019S Parkinson's patients, as well as the protein targets associated with this reduction provided crucial groundwork for our subsequent studies which aim at identifying diagnostic and prognostic biomarkers for Parkinson's disease. Building upon these findings, our current study has expanded data cohorts and explored new signature types to develop a biomarker panel with a high predictive rate for Parkinson's diagnosis. These efforts aim to contribute to the development of more accurate and sensitive diagnostic tools for Parkinson's disease, to achieve earlier intervention and better management of this debilitating disorder.

The studies discussed in this thesis utilized mass spectrometry (MS) technology, which identifies and quantifies thousands of proteins, as well as post-translational modifications such as phosphorylation. An automated R pipeline called PhosPiR was developed in Study I of the present thesis, which integrates the various layers of MS data analysis, offering multi-level functional analyses and supporting 18 different organisms. This pipeline saves time and effort in analyzing proteomics and phosphoproteomics datasets, and provides a user-friendly means for non-programmers to access analysis tools such as ROTS and rank product statistical tests, PTM-SEA, and more, that are only available as R packages to date.

Overall, our studies highlight the importance of incorporating proteomics and phosphoproteomics data to gain a comprehensive understanding of the complex biological processes involved in psychiatric and neurodegenerative disorders. The application of analytical tools such as PhosPiR and GeneMANIA can aid in the identification of key pathways involved in disease pathology. Our studies made significant contributions to the field of neural research and provide a foundation for further investigation into the molecular mechanisms of schizophrenia and Parkinson's disease.

List of References

[1] S. Ramazi and J. Zahiri, "Post-translational modifications in proteins: Resources, tools and prediction methods," *Database*, vol. 2021. Oxford University Press, 2021. doi: 10.1093/database/baab012.

[2] G. Duan and D. Walther, "The Roles of Post-translational Modifications in the Context of Protein Interaction Networks," *PLoS Comput Biol*, vol. 11, no. 2, 2015, doi: 10.1371/journal.pcbi.1004049.

[3] F. Ardito, M. Giuliani, D. Perrone, G. Troiano, and L. lo Muzio, "The crucial role of protein phosphorylation in cell signalingand its use as targeted therapy (Review)," *International Journal of Molecular Medicine*, vol. 40, no. 2. Spandidos Publications, pp. 271–280, Aug. 01, 2017. doi: 10.3892/ijmm.2017.3036.

[4] E. J. Needham, B. L. Parker, T. Burykin, D. E. James, and S. J. Humphrey, "Illuminating the dark phosphoproteome," 2019. [Online]. Available: www.chemicalprobes.org/

[5] H. Nishi, A. Shaytan, and A. R. Panchenko, "Physicochemical mechanisms of protein regulation by phosphorylation," *Frontiers in Genetics*, vol. 5, no. AUG. Frontiers Research Foundation, 2014. doi: 10.3389/fgene.2014.00270.

[6] R. Roskoski, "ERK1/2 MAP kinases: Structure, function, and regulation," *Pharmacological Research*, vol. 66, no. 2. pp. 105–143, Aug. 2012. doi: 10.1016/j.phrs.2012.04.005.

[7] M. L. Miller *et al.*, "Linear motif atlas for phosphorylation-dependent signaling," *Sci Signal*, vol. 1, no. 35, Sep. 2008, doi: 10.1126/scisignal.1159433.

[8] T. Hunter, "Tyrosine phosphorylation: thirty years and counting," *Current Opinion in Cell Biology*, vol. 21, no. 2. pp. 140–146, Apr. 2009. doi: 10.1016/j.ceb.2009.01.028.

[9] G. A. Wayman, H. Tokumitsu, M. A. Davare, and T. R. Soderling, "Analysis of CaM-kinase signaling in cells," *Cell Calcium*, vol. 50, no. 1. Elsevier Ltd, pp. 1–8, 2011. doi: 10.1016/j.ceca.2011.02.007.

[10] E. J. Eide and D. M. Virshup, “CASEIN KINASE I: ANOTHER COG IN THE CIRCADIAN CLOCKWORKS,” *Chronobiol Int*, vol. 18, no. 3, pp. 389–398, Jan. 2001, doi: 10.1081/CBI-100103963.

[11] J. Jin and T. Pawson, “Modular evolution of phosphorylation-based signalling systems,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 367, no. 1602. Royal Society, pp. 2540–2555, 2012. doi: 10.1098/rstb.2012.0106.

[12] A. Müller-Taubenberger, H. C. Ishikawa-Ankerhold, P. M. Kastner, E. Burghardt, and G. Gerisch, “The STE group kinase SepA controls cleavage furrow formation in dictyostelium,” in *Cell Motility and the Cytoskeleton*, Nov. 2009, pp. 929–939. doi: 10.1002/cm.20386.

[13] L. R. Pearce, D. Komander, and D. R. Alessi, “The nuts and bolts of AGC protein kinases,” *Nature Reviews Molecular Cell Biology*, vol. 11, no. 1. pp. 9–22, Jan. 2010. doi: 10.1038/nrm2822.

[14] A. I. Abdi, T. G. Carvalho, J. M. Wilkes, and C. Doerig, “A secreted Plasmodium falciparum kinase reveals a signature motif for classification of tyrosine kinase-like kinases,” *Microbiology (United Kingdom)*, vol. 159, no. PART 12, pp. 2533–2547, 2013, doi: 10.1099/mic.0.070409-0.

[15] M. v. Sundaram, “RTK/Ras/MAPK signaling.,” *WormBook : the online review of C. elegans biology*. pp. 1–19, 2006. doi: 10.1895/wormbook.1.80.1.

[16] D. Barford, “PROTEIN PHOSPHATASES Molecular mechanisms of the protein serine/threonine phosphatases,” 1996.

[17] F. M. Moeslein, M. P. Myers, and G. E. Landreth, “The CLK Family Kinases, CLK1 and CLK2, Phosphorylate and Activate the Tyrosine Phosphatase, PTP-1B*,” 1999. [Online]. Available: http://www.jbc.org

[18] B. ’Johnson, A. ’Lewis, J. ’Raff, M. ’Roberts, K. ’Walter, P. ’Alberts, *Molecular Biology of the Cell*, 5th ed. New York, NY: Garland Science, 2007.

[19] K. ’Huether, S. ’McCance, *Pathophysiology: The Biologic Basis for Disease in Adults and Children.*, 7th ed. Elsevier, 2014.

[20] S. J. Humphrey, D. E. James, and M. Mann, “Protein Phosphorylation: A Major Switch Mechanism for Metabolic Regulation,” *Trends in Endocrinology & Metabolism*, vol. 26, no. 12, pp. 676–687, 2015, doi: https://doi.org/10.1016/j.tem.2015.09.013.

[21] N. Case *et al.*, “Mechanical regulation of glycogen synthase kinase 3β (GSK3β) in mesenchymal stem cells is dependent on Akt protein serine 473 phosphorylation via mTORC2 protein,” *Journal of Biological Chemistry*, vol. 286, no. 45, pp. 39450–39456, Nov. 2011, doi: 10.1074/jbc.M111.265330.

[22] P. A. Cole, K. Shen, Y. Qiao, and D. Wang, “Protein tyrosine kinases Src and Csk: A tail’s tale,” *Current Opinion in Chemical Biology*, vol. 7, no. 5. Elsevier Ltd, pp. 580–585, 2003. doi: 10.1016/j.cbpa.2003.08.009.

[23] H. Nishi, J. H. Fong, C. Chang, S. A. Teichmann, and A. R. Panchenko, “Regulation of protein-protein binding by coupling between phosphorylation and intrinsic disorder: Analysis of human protein complexes,” *Mol Biosyst*, vol. 9, no. 7, pp. 1620–1626, 2013, doi: 10.1039/c3mb25514j.

[24] H. Kuwahara, M. Nishizaki, and H. Kanazawa, “Nuclear localization signal and phosphorylation of serine350 specify intracellular localization of DRAK2,” *J Biochem*, vol. 143, no. 3, pp. 349–358, Mar. 2008, doi: 10.1093/jb/mvm236.

[25] Y. Fukami and F. Lipmann, “Reversal of Rous sarcoma-specific immunoglobulin phosphorylation on tyrosine (ADP as phosphate acceptor) catalyzed by the src gene kinase (tumor-bearing rabbit serum/protein-bound tyrosine 0-phosphate-ADP equilibrium constant/AG0’ of hydrolysis),” 1983.

[26] X. Xu *et al.*, “The CUL7 E3 Ubiquitin Ligase Targets Insulin Receptor Substrate 1 for Ubiquitin-Dependent Degradation,” *Mol Cell*, vol. 30, no. 4, pp. 403–414, May 2008, doi: 10.1016/j.molcel.2008.03.009.

[27] Z. Liu, Y. Wang, and Y. Xue, “Phosphoproteomics-based network medicine,” *FEBS Journal*, vol. 280, no. 22. pp. 5696–5704, Nov. 2013. doi: 10.1111/febs.12380.

[28] G. E. Lienhard, “Non-functional phosphorylations?” *Trends in Biochemical Sciences*, vol. 33, no. 8. pp. 351–352, Aug. 2008. doi: 10.1016/j.tibs.2008.05.004.

[29] E. D. Levy, S. W. Michnick, and C. R. Landry, “Protein abundance is key to distinguish promiscuous from functional phosphorylation based on evolutionary information,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 367, no. 1602. Royal Society, pp. 2594–2606, 2012. doi: 10.1098/rstb.2012.0078.

[30] J. v. Olsen *et al.*, “Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks,” *Cell*, vol. 127, no. 3, pp. 635–648, Nov. 2006, doi: 10.1016/j.cell.2006.09.026.

[31] P. Blume-Jensen and T. Hunter, “Oncogenic kinase signalling,” *Nature*, vol. 411, no. 6835, pp. 355–365, 2001, doi: 10.1038/35077225.

[32] H. K. Kwon, H. Choi, S. G. Park, W. J. Park, D. H. Kim, and Z. Y. Park, “Integrated quantitative phosphoproteomics and cell-based functional screening reveals specific pathological cardiac hypertrophy-related phosphorylation sites,” *Mol Cells*, vol. 44, no. 7, pp. 500–516, 2021, doi: 10.14348/molcells.2021.4002.

[33] J. v. Arrington, C. C. Hsu, S. G. Elder, and W. Andy Tao, “Recent advances in phosphoproteomics and application to neurological diseases,” *Analyst*, vol. 142, no. 23. Royal Society of Chemistry, pp. 4373–4387, Dec. 07, 2017. doi: 10.1039/c7an00985b.

[34] M. Bouhaddou *et al.*, “The Global Phosphorylation Landscape of SARS-CoV-2 Infection,” *Cell*, vol. 182, no. 3, pp. 685-712.e19, Aug. 2020, doi: 10.1016/j.cell.2020.06.034.

[35] P. A. Jones and S. B. Baylin, “The fundamental role of epigenetic events in cancer,” *Nature Reviews Genetics*, vol. 3, no. 6. pp. 415–428, 2002. doi: 10.1038/nrg816.

[36] V. Singh, M. Ram, R. Kumar, R. Prasad, B. K. Roy, and K. K. Singh, “Phosphorylation: Implications in Cancer,” *Protein Journal*, vol. 36, no. 1. Springer New York LLC, Feb. 01, 2017. doi: 10.1007/s10930-017-9696-z.

[37] T. Hunter, “Signaling-2000 and Beyond Review,” *Cell*, 7;100(1):113-27, Jan. 2000. doi: 10.1016/s0092-8674(00)81688-8. PMID: 10647936.

[38] D. Stehelin, G. E. Harold Var, and D. J. Michael, “Purification of DNA Complementary to Nucleotide Sequences Required for Neoplastic Transformation of Fibroblasts by Avian Sarcoma Viruses,” 1976.

[39] N. Spector, W. Xia, I. El-Hariry, Y. Yarden, and S. Bacus, “HER2 therapy. Small molecule HER-2 tyrosine kinase inhibitors,” *Breast Cancer Research*, vol. 9, no. 2. Mar. 02, 2007. doi: 10.1186/bcr1652.

[40] Wyeth Ayerst Research, “Enhanced sensitivity of PTEN-deficient tumors to inhibition of FRAPmTOR.” [Online]. Available: www.pnas.orgcgidoi10.1073pnas.171076798

[41] P. J. Roberts and C. J. Der, “Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer,” *Oncogene*, vol. 26, no. 22. pp. 3291–3310, May 14, 2007. doi: 10.1038/sj.onc.1210422.

[42] J. M. Shields, K. Pruitt, A. McFall, A. Shaub, and C. J. Der, “Understanding Ras: ‘it ain’t over ’til it’s over’,” *Trends Cell Biol*, vol. 10, no. 4, pp. 147–154, 2000, doi: https://doi.org/10.1016/S0962-8924(00)01740-2.

[43] S. Benzeno *et al.*, “Identification of mutations that disrupt phosphorylation-dependent nuclear export of cyclin D1,” *Oncogene*, vol. 25, no. 47, pp. 6291–6303, Oct. 2006, doi: 10.1038/sj.onc.1209644.

[44] D. Hanahan and R. A. Weinberg, “The Hallmarks of Cancer Review evolve progressively from normalcy via a series of pre,” 2000.

[45] M. B. Datto, P. P.-C. Hu, T. F. Kowalik, Jonathan Yingling, and X.-F. Wang, “The Viral Oncoprotein E1A Blocks Transforming Growth Factor-Mediated Induction of p21/WAF1/Cip1 and p15/INK4B,” *Mol Cell Biol*, 17(4):2030-7, Apr. 1997, doi: 10.1128/MCB.17.4.2030.

[46] L. Zuo *et al.*, "Germline mutations in the p16INK4a binding domain of CDK4 in familial melanoma," *Nat Genet*, vol. 12, no. 1, pp. 97–99, 1996, doi: 10.1038/ng0196-97.

[47] L. Pontano and J. Diehl, "Speeding through cell cycle roadblocks: Nuclear cyclin D1-dependent kinase and neoplastic transformation," *Cell Div*, vol. 3, p. 12, Oct. 2008, doi: 10.1186/1747-1028-3-12.

[48] M. F. Lindberg and L. Meijer, "Dual-specificity, tyrosine phosphorylation-regulated kinases (Dyrks) and cdc2-like kinases (clks) in human disease, an overview," *International Journal of Molecular Sciences*, vol. 22, no. 11. MDPI, Jun. 01, 2021. doi: 10.3390/ijms22116047.

[49] X. Xiao *et al.*, "Association of Genes Involved in the Metabolic Pathways of Amyloid-β and Tau Proteins With Sporadic Late-Onset Alzheimer's Disease in the Southern Han Chinese Population," *Front Aging Neurosci*, vol. 12, 2020, [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnagi.2020.584801

[50] M. M. Attwood, D. Fabbro, A. v. Sokolov, S. Knapp, and H. B. Schiöth, "Trends in kinase drug discovery: targets, indications and inhibitor design," *Nature Reviews Drug Discovery*, vol. 20, no. 11. Nature Research, pp. 839–861, Nov. 01, 2021. doi: 10.1038/s41573-021-00252-y.

[51] E.J. Brown *et al.*, "A mammalian protein targeted by Gl-arresting rapamycin-receptor complex," *Nature*, 30;369(6483):756-8, doi: 10.1038/369756a0, Jun 1994.

[52] B.J. Druker *et al.*, "Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells," *Nat Med.* May 1996. [Online]. Available: http://www.nature.com/naturemedicine

[53] H. K. Agop Antarjian *et al.*, "HEMATOLOGIC AND CYTOGENETIC RESPONSES TO IMATINIB MESYLATE IN CHRONIC MYELOGENOUS LEUKEMIA," *N Engl J Med*, Feb. 2002. [Online]. Available: www.nejm.org

[54] W. Vainchenker, A. R. Green, J. Robyn, and S. N. Constantinescu, "A Unique Activating Mutation in JAK2 (V617F) Is at the Origin of Polycythemia Vera and Allows a New Classification of Myeloproliferative Diseases," *Hematology Am Soc Hematol Educ Program*, 2005. [Online]. Available: http://ashpublications.org/hematology/article-pdf/2005/1/195/645191/195.pdf

[55] FDA, "FDA approves everolimus for tuberous sclerosis complex- associated partial- onset seizures.," Apr. 11, 2018. [Online]. Available: https://www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-everolimus-tuberous-sclerosis-complex-associated-partial-onset-

seizures#:~:text=On%20April%2010%2C%202018%2C%20the,)%2Dassociated%20partial%2Donset%20seizures.

[56] P. L. McCormack, “Nintedanib: first global approval,” *Drugs*, vol. 75, no. 1, pp. 129–139, Jan. 2015, doi: 10.1007/s40265-014-0335-0.

[57] Y. Y. Zhou, Y. Li, W. Q. Jiang, and L. F. Zhou, “MAPK/JNK signalling: A potential autophagy regulation pathway,” *Biosci Rep*, vol. 35, no. 3, pp. 1–10, 2015, doi: 10.1042/BSR20140141.

[58] Z. Xia, M. Dickens, J. Raingeaud, R. J. Davis, and M. E. Greenberg, “Opposing Effects of ERK and JNK-p38 MAP Kinases on Apoptosis,” *Science*, 270,1326-1331, 1995, DOI:10.1126/science.270.5240.1326. [Online]. Available: https://www.science.org

[59] A. Zeke, M. Misheva, A. Reményi, and M. A. Bogoyevitch, “JNK Signaling: Regulation and Functions Based on Complex Protein-Protein Partnerships,” *Microbiology and Molecular Biology Reviews*, vol. 80, no. 3, pp. 793–835, Sep. 2016, doi: 10.1128/mmbr.00043-14.

[60] J. Ha, E. Kang, J. Seo, and S. Cho, “Phosphorylation dynamics of jnk signaling: Effects of dual-specificity phosphatases (dusps) on the jnk pathway,” *International Journal of Molecular Sciences*, vol. 20, no. 24. MDPI AG, Dec. 02, 2019. doi: 10.3390/ijms20246157.

[61] J. Cicenas, E. Zalyte, A. Rimkus, D. Dapkus, R. Noreika, and S. Urbonavicius, “JNK, p38, ERK, and SGK1 Inhibitors in Cancer,” *Cancers (Basel)*, vol. 10, no. 1, p. 1, Dec. 2017, doi: 10.3390/cancers10010001.

[62] Y. Mizukami, K. Yoshioka, S. Morimotoi, and K. Yoshida, “A Novel Mechanism of JNK1 Activation NUCLEAR TRANSLOCATION AND ACTIVATION OF JNK1 DURING ISCHEMIA AND REPERFUSION,” *J Biol Chem*, 27;272(26):16657-62, Jun 1997, doi: 10.1074/jbc.272.26.16657. [Online]. Available: http://www.jbc.org

[63] S. Morton, R. J. Davis, A. McLaren, and P. Cohen, “A reinvestigation of the multisite phosphorylation of the transcription factor c-Jun,” *EMBO J*, vol. 22, no. 15, pp. 3876–3886, Aug. 2003, doi: 10.1093/emboj/cdg388.

[64] M. Treier, L. M. Staszewski, and D. Bohmann, “Ubiquitin-dependent c-Jun degradation in vivo is mediated by the δ domain,” *Cell*, vol. 78, no. 5, pp. 787–798, 1994, doi: https://doi.org/10.1016/S0092-8674(94)90502-9.

[65] E. Komulainen *et al.*, “JNK1 controls dendritic field size in L2/3 and l5 of the motor cortex, constrains soma size, and influences fine motor coordination,” *Front Cell Neurosci*, vol. 8, Sep. 2014, doi: 10.3389/fncel.2014.00272.

[66] R. L. Openshaw *et al.*, “Map2k7 Haploinsufficiency Induces Brain Imaging Endophenotypes and Behavioral Phenotypes Relevant to Schizophrenia,”

Schizophr Bull, vol. 46, no. 1, pp. 211–223, Jan. 2020, doi: 10.1093/schbul/sbz044.

[67] S. H. Ansarey, "Inflammation and JNK's Role in Niacin-GPR109A Diminished Flushed Effect in Microglial and Neuronal Cells With Relevance to Schizophrenia," *Frontiers in Psychiatry*, vol. 12. Frontiers Media S.A., Nov. 30, 2021. doi: 10.3389/fpsyt.2021.771144.

[68] G. B. of D. S. 2013 Collaborators, "Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013," *Lancet*, vol. 386, no. 9995, pp. 743–800, Aug. 2015, doi: 10.1016/S0140-6736(15)60692-4.

[69] H. Y. Chong, S. L. Teoh, D. B. C. Wu, S. Kotirum, C. F. Chiou, and N. Chaiyakunapruk, "Global economic burden of schizophrenia: A systematic review," *Neuropsychiatric Disease and Treatment*, vol. 12. Dove Medical Press Ltd, pp. 357–373, Feb. 16, 2016. doi: 10.2147/NDT.S96649.

[70] I. R. Winship *et al.*, "An Overview of Animal Models Related to Schizophrenia," *Canadian Journal of Psychiatry*, vol. 64, no. 1. SAGE Publications Inc., pp. 5–17, Jan. 01, 2019. doi: 10.1177/0706743718773728.

[71] P. F. Buckley, "Neuroinflammation and Schizophrenia," *Current Psychiatry Reports*, vol. 21, no. 8. Current Medicine Group LLC 1, Aug. 01, 2019. doi: 10.1007/s11920-019-1050-z.

[72] C. Dong *et al.*, "JNK is required for effector T-cell function but not for T-cell activation," *Nature*, vol. 405, no. 6782, pp. 91–94, 2000, doi: 10.1038/35011091.

[73] A. Kadakia *et al.*, "The economic burden of schizophrenia in the United States," *Journal of Clinical Psychiatry*, vol. 83, no. 6, p. 22m14458, 2022.

[74] L. K. Nguyen, D. Matallanas, D. R. Croucher, A. von Kriegsheim, and B. N. Kholodenko, "Signalling by protein phosphatases and drug development: A systems-centred view," *FEBS Journal*, vol. 280, no. 2. pp. 751–765, Jan. 2013. doi: 10.1111/j.1742-4658.2012.08522.x.

[75] U. S. Bhalla, P. T. Ram, and R. Iyengar, "MAP Kinase Phosphatase As a Locus of Flexibility in a Mitogen-Activated Protein Kinase Signaling Network," *Science (1979)*, vol. 297, no. 5583, pp. 1018–1023, Aug. 2002, doi: 10.1126/science.1068873.

[76] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam, "The Protein Kinase Complement of the Human Genome," *Science (1979)*, vol. 298, no. 5600, pp. 1912–1934, Dec. 2002, doi: 10.1126/science.1075762.

[77] S. Gupta1 *et al.*, "Selective interaction of JNK protein kinase isoforms with transcription factors," *EMBO J.*, 3;15(11):2760-70, Jun 199, PMID: 8654373 1996.

[78] C.-Y. Kuan, D. D. Yang, D. R. S. Roy, R. J. Davis, P. Rakic, and R. A. Flavell, "The Jnk1 and Jnk2 Protein Kinases Are Required for Regional Specific Apoptosis during Early Brain Development," *Neuron*, vol. 22, no. 4, pp. 667–676, 1999, doi: https://doi.org/10.1016/S0896-6273(00)80727-8.

[79] L. Chang, Y. Jones, M. H. Ellisman, L. S. B. Goldstein, and M. Karin, "JNK1 Is Required for Maintenance of Neuronal Microtubules and Controls Phosphorylation of Microtubule-Associated Proteins," *Dev Cell*, vol. 4, no. 4, pp. 521–533, 2003, doi: https://doi.org/10.1016/S1534-5807(03)00094-7.

[80] D. D. Yang *et al.*, "Differentiation of CD4+ T Cells to Th1 Cells Requires MAP Kinase JNK2," *Immunity*, vol. 9, no. 4, pp. 575–585, 1998, doi: https://doi.org/10.1016/S1074-7613(00)80640-8.

[81] C. R. Amura, L. Marek, R. A. Winn, and L. E. Heasley, "Inhibited neurogenesis in JNK1-deficient embryonic stem cells," *Mol Cell Biol*, vol. 25, no. 24, pp. 10791–10802, Dec. 2005, doi: 10.1128/MCB.25.24.10791-10802.2005.

[82] S. G. Reich and J. M. Savitt, "Parkinson's Disease," *Medical Clinics of North America*, vol. 103, no. 2. W.B. Saunders, pp. 337–350, Mar. 01, 2019. doi: 10.1016/j.mcna.2018.10.014.

[83] M. T. Hayes, "Parkinson's Disease and Parkinsonism," *American Journal of Medicine*, vol. 132, no. 7. Elsevier Inc., pp. 802–807, Jul. 01, 2019. doi: 10.1016/j.amjmed.2019.03.001.

[84] A. J. Jagadeesan *et al.*, "Current trends in etiology, prognosis and therapeutic aspects of Parkinson's disease: a review," *Acta Biomed*, vol. 88, pp. 249–262, 2017, doi: 10.23750/abm.v%vi%i.6063.

[85] S. Lotankar, K. S. Prabhavalkar, and L. K. Bhatt, "Biomarkers for Parkinson's Disease: Recent Advancement," *Neurosci Bull*, vol. 33, no. 5, pp. 585–597, Oct. 2017, doi: 10.1007/s12264-017-0183-5.

[86] Z. Ou *et al.*, "Global Trends in the Incidence, Prevalence, and Years Lived With Disability of Parkinson's Disease in 204 Countries/Territories From 1990 to 2019," *Front Public Health*, vol. 9, 2021, [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpubh.2021.776847

[87] R. Aebersold and M. Mann, "Mass-spectrometric exploration of proteome structure and function," *Nature*, vol. 537, no. 7620. Nature Publishing Group, pp. 347–355, Sep. 14, 2016. doi: 10.1038/nature19949.

[88] N. Pappireddi, L. Martin, and M. Wühr, "A Review on Quantitative Multiplexed Proteomics," *ChemBioChem*, vol. 20, no. 10. Wiley-VCH Verlag, pp. 1210–1224, May 15, 2019. doi: 10.1002/cbic.201800650.

[89] M. Mardamshina and T. Geiger, "Next-Generation Proteomics and Its Application to Clinical Breast Cancer Research," *Am J Pathol*, vol. 187, no. 10, pp. 2175–2184, 2017, doi: https://doi.org/10.1016/j.ajpath.2017.07.003.

[90] A. Montoya, L. Beltran, P. Casado, J.-C. Rodríguez-Prados, and P. R. Cutillas, "Characterization of a TiO2 enrichment method for label-free quantitative phosphoproteomics," *Methods*, vol. 54, no. 4, pp. 370–378, 2011, doi: https://doi.org/10.1016/j.ymeth.2011.02.004.

[91] L. Krasny and P. H. Huang, "Data-independent acquisition mass spectrometry (DIA-MS) for proteomic applications in oncology," *Mol Omics*, vol. 17, no. 1, pp. 29–42, 2021, doi: 10.1039/D0MO00072H.

[92] Y. Ma and J. R. Yates 3rd, "Proteomics and pulse azidohomoalanine labeling of newly synthesized proteins: what are the potential applications?," *Expert Rev Proteomics*, vol. 15, no. 7, pp. 545–554, Jul. 2018, doi: 10.1080/14789450.2018.1500902.

[93] J. Zhang, J. Wang, S. Ng, Q. Lin, and H.-M. Shen, "Development of a novel method for quantification of autophagic protein degradation by AHA labeling," *Autophagy*, vol. 10, no. 5, pp. 901–912, May 2014, doi: 10.4161/auto.28267.

[94] N. Rauniyar, "Parallel Reaction Monitoring: A Targeted Experiment Performed Using High Resolution and High Mass Accuracy Mass Spectrometry," *Int J Mol Sci*, vol. 16, no. 12, pp. 28566–28581, Dec. 2015, doi: 10.3390/ijms161226120.

[95] A. F. M. Altelaar, J. Munoz, and A. J. R. Heck, "Next-generation proteomics: towards an integrative view of proteome dynamics," *Nat Rev Genet*, vol. 14, no. 1, pp. 35–48, 2013, doi: 10.1038/nrg3356.

[96] S. Tyanova, T. Temu, and J. Cox, "The MaxQuant computational platform for mass spectrometry-based shotgun proteomics," *Nat Protoc*, vol. 11, no. 12, pp. 2301–2319, 2016, doi: 10.1038/nprot.2016.136.

[97] T. Välikangas, T. Suomi, and L. L. Elo, "A systematic evaluation of normalization methods in quantitative label-free proteomics," *Brief Bioinform*, vol. 19, no. 1, pp. 1–11, Jan. 2018, doi: 10.1093/bib/bbw095.

[98] Y. v. Karpievitch, A. R. Dabney, and R. D. Smith, "Normalization and missing value imputation for label-free LC-MS analysis.," *BMC Bioinformatics*, vol. 13 Suppl 16, 2012, doi: 10.1186/1471-2105-13-S16-S5.

[99] B. M. Bolstad, R. A. Irizarry, M. ˚ Astrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, 22;19(2):185-93, Jan. 2003, doi: 10.1093/bioinformatics/19.2.185. [Online] Available: http://www.bioconductor.org.

[100] O. Kauko *et al.*, "Label-free quantitative phosphoproteomics with novel pairwise abundance normalization reveals synergistic RAS and CIP2A signaling," *Sci Rep*, vol. 5, Aug. 2015, doi: 10.1038/srep13099.

[101] S. Saraei, T. Suomi, O. Kauko, and L. L. Elo, “Phosphonormalizer: An R package for normalization of MS-based label-free phosphoproteomics,” *Bioinformatics*, vol. 34, no. 4, pp. 693–694, Feb. 2018, doi: 10.1093/bioinformatics/btx573.

[102] J. T. Leek *et al.*, “Tackling the widespread and critical impact of batch effects in high-throughput data,” *Nature Reviews Genetics*, vol. 11, no. 10. pp. 733–739, Oct. 14, 2010. doi: 10.1038/nrg2825.

[103] W. W. bin Goh, W. Wang, and L. Wong, “Why Batch Effects Matter in Omics Data, and How to Avoid Them,” *Trends in Biotechnology*, vol. 35, no. 6. Elsevier Ltd, pp. 498–507, Jun. 01, 2017. doi: 10.1016/j.tibtech.2017.02.012.

[104] H. Zhang *et al.*, “Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer,” *Cell*, vol. 166, no. 3, pp. 755–765, Jul. 2016, doi: 10.1016/j.cell.2016.05.069.

[105] T. Sajic *et al.*, “Similarities and Differences of Blood N-Glycoproteins in Five Solid Carcinomas at Localized Clinical Stage Analyzed by SWATH-MS,” *Cell Rep*, vol. 23, no. 9, pp. 2819-2831.e5, May 2018, doi: 10.1016/j.celrep.2018.04.114.

[106] B. C. Collins *et al.*, “Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry,” *Nat Commun*, vol. 8, no. 1, Dec. 2017, doi: 10.1038/s41467-017-00249-5.

[107] L. Zhou, A. Chi-Hau Sue, and W. W. bin Goh, “Examining the practical limits of batch effect-correction algorithms: When should you care about batch effects?,” *Journal of Genetics and Genomics*, vol. 46, no. 9, pp. 433–443, Sep. 2019, doi: 10.1016/j.jgg.2019.08.002.

[108] J. Čuklina *et al.*, “Diagnostics and correction of batch effects in large-scale proteomic studies: a tutorial,” *Mol Syst Biol*, vol. 17, no. 8, Aug. 2021, doi: 10.15252/msb.202110240.

[109] B. M. Bolstad, R. A. Irizarry, M. ˚ Astrand, and T. P. Speed, “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias,” *Bioinformatics*, 22;19(2):185-93, Jan. 2003, doi: 10.1093/bioinformatics/19.2.185. [Online] Available: http://www.bioconductor.org.

[110] W. E. Johnson, C. Li, and A. Rabinovic, “Adjusting batch effects in microarray expression data using empirical Bayes methods,” *Biostatistics*, vol. 8, no. 1, pp. 118–127, Jan. 2007, doi: 10.1093/biostatistics/kxj037.

[111] C. Lazar *et al.*, “Batch effect removal methods for microarray gene expression data integration: A survey,” *Brief Bioinform*, vol. 14, no. 4, pp. 469–490, Jul. 2013, doi: 10.1093/bib/bbs037.

[112] B. J. M. Webb-Robertson *et al.*, "Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics," *Journal of Proteome Research*, vol. 14, no. 5. American Chemical Society, pp. 1993–2001, May 01, 2015. doi: 10.1021/pr501138h.

[113] W. W. B. Goh, Y. H. Lee, M. Chung, and L. Wong, "How advancement in biological network analysis methods empowers proteomics," *Proteomics*, vol. 12, no. 4–5. pp. 550–563, Feb. 2012. doi: 10.1002/pmic.201100321.

[114] W. W. bin Goh, M. J. Sergot, J. C. Sng, and L. Wong, "Comparative network-based recovery analysis and proteomic profiling of neurological changes in valproic acid-treated mice," *J Proteome Res*, vol. 12, no. 5, pp. 2116–2127, May 2013, doi: 10.1021/pr301127f.

[115] W. Zhang, F. Li, L. Nie, G. Wu, and J. Qiao, "Prediction and characterization of missing proteomic data in desulfovibrio vulgaris," *Comp Funct Genomics*, vol. 2011, 2011, doi: 10.1155/2011/780973.

[116] L. Jin *et al.*, "A comparative study of evaluating missing value imputation methods in label-free proteomics," *Sci Rep*, vol. 11, no. 1, p. 1760, 2021, doi: 10.1038/s41598-021-81279-4.

[117] R. Wei *et al.*, "Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data," *Sci Rep*, vol. 8, no. 1, Dec. 2018, doi: 10.1038/s41598-017-19120-0.

[118] O. Troyanskaya *et al.*, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520-525, Jun. 2001. [Online]. Available: http://smi-web.

[119] T. Clough, S. Thaminy, S. Ragg, R. Aebersold, and O. Vitek, "Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs.," *BMC Bioinformatics*, vol. 13 Suppl 16, 2012, doi: 10.1186/1471-2105-13-S16-S6.

[120] X. Wang, G. A. Anderson, R. D. Smith, and A. R. Dabney, "A hybrid approach to protein differential expression in mass spectrometry-based proteomics," *Bioinformatics*, vol. 28, no. 12, pp. 1586–1591, Jun. 2012, doi: 10.1093/bioinformatics/bts193.

[121] S. J. Deeb, R. C. J. D'Souza, J. Cox, M. Schmidt-Supprian, and M. Mann, "Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles," in *Molecular and Cellular Proteomics*, May 2012, pp. 77–89. doi: 10.1074/mcp.M111.015362.

[122] N. C. Hubner *et al.*, "Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions," *Journal of Cell Biology*, vol. 189, no. 4, pp. 739–754, May 2010, doi: 10.1083/jcb.200911091.

[123] S. Oh, D. D. Kang, G. N. Brock, and G. C. Tseng, “Biological impact of missing-value imputation on downstream analyses of gene expression profiles,” *Bioinformatics*, vol. 27, no. 1, pp. 78–86, Jan. 2011, doi: 10.1093/bioinformatics/btq613.

[124] K. M. Fouad, M. M. Ismail, A. T. Azar, and M. M. Arafa, “Advanced methods for missing values imputation based on similarity learning,” *PeerJ Comput Sci*, vol. 7, pp. 1–38, 2021, doi: 10.7717/PEERJ-CS.619.

[125] H. Kim, G. H. Golub, and H. Park, “Missing value estimation for DNA microarray gene expression data: Local least squares imputation,” *Bioinformatics*, vol. 21, no. 2, pp. 187–198, Jan. 2005, doi: 10.1093/bioinformatics/bth499.

[126] T. H. Bø, B. Dysvik, and I. Jonassen, “LSimpute: accurate estimation of missing values in microarray data with least squares methods.,” *Nucleic Acids Res*, vol. 32, no. 3, 2004, doi: 10.1093/nar/gnh026.

[127] T. Schneider, “Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values,” *J. Climate*, 2001.

[128] Y. Karpievitch *et al.*, “A statistical framework for protein quantitation in bottom-up MS-based proteomics,” *Bioinformatics*, vol. 25, no. 16, pp. 2028–2034, Aug. 2009, doi: 10.1093/bioinformatics/btp362.

[129] M. E. Tipping and C. M. Bishop, “Mixtures of Probabilistic Principal Component Analyzers,” *Neural Comput*, vol. 11, no. 2, pp. 443–482, Feb. 1999, doi: 10.1162/089976699300016728.

[130] H. Hegde, N. Shimpi, A. Panny, I. Glurich, P. Christie, and A. Acharya, “MICE vs PPCA: Missing data imputation in healthcare,” *Inform Med Unlocked*, vol. 17, Jan. 2019, doi: 10.1016/j.imu.2019.100275.

[131] S. Hediyeh-Zadeh, A. I. Webb, and M. J. Davis, “MSImpute: Imputation of label-free mass spectrometry peptides by low-rank approximation”, *bioRxiv*, Aug. 2020. doi: 10.1101/2020.08.12.248963.

[132] M. Shen *et al.*, “Comparative assessment and novel strategy on methods for imputing proteomics data,” *Sci Rep*, vol. 12, no. 1, p. 1067, 2022, doi: 10.1038/s41598-022-04938-0.

[133] K. L. Howe *et al.*, “Ensembl 2021,” *Nucleic Acids Res*, vol. 49, no. D1, pp. D884–D891, Jan. 2021, doi: 10.1093/nar/gkaa942.

[134] A. Bateman *et al.*, “UniProt: the universal protein knowledgebase in 2021,” *Nucleic Acids Res*, vol. 49, no. D1, pp. D480–D489, Jan. 2021, doi: 10.1093/nar/gkaa1100.

[135] S. Adhikari *et al.*, “A high-stringency blueprint of the human proteome,” *Nature Communications*, vol. 11, no. 1. Nature Research, Dec. 01, 2020. doi: 10.1038/s41467-020-19045-9.

[136] Q. Wang *et al.*, "The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas," *Cell*, vol. 181, no. 4, pp. 936-953.e20, May 2020, doi: 10.1016/j.cell.2020.04.007.

[137] P. v. Hornbeck *et al.*, "15 years of PhosphoSitePlus ®: Integrating post-translationally modified sites, disease variants and isoforms," *Nucleic Acids Res*, vol. 47, no. D1, pp. D433–D441, Jan. 2019, doi: 10.1093/nar/gky1159.

[138] T. K. Kim, "T test as a parametric statistic," *Korean J Anesthesiol*, vol. 68, no. 6, pp. 540–546, Dec. 2015, doi: 10.4097/kjae.2015.68.6.540.

[139] D. J. Sheskin, "Parametric Versus Nonparametric Tests," in *International Encyclopedia of Statistical Science*, M. Lovric, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1051–1052. doi: 10.1007/978-3-642-04898-2_440.

[140] K. Kammers, R. N. Cole, C. Tiengwe, and I. Ruczinski, "Detecting Significant Changes in Protein Abundance," *EuPA Open Proteom*, vol. 7, pp. 11–19, Jun. 2015, doi: 10.1016/j.euprot.2015.02.002.

[141] K. Zhao and S. Y. Rhee, "Interpreting omics data with pathway enrichment analysis," *Trends in Genetics*, 2023, doi: https://doi.org/10.1016/j.tig.2023.01.003.

[142] J. Reimand *et al.*, "Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap," *Nat Protoc*, vol. 14, no. 2, pp. 482–517, Feb. 2019, doi: 10.1038/s41596-018-0103-9.

[143] A. Subramanian *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc Natl Acad Sci U S A*, 25;102(43):15545-50, Oct. 2005, doi: 10.1073/pnas.0506580102 2005. [Online]. Available: www.pnas.orgcgidoi10.1073pnas.0506580102

[144] C. Hernandez-Armenta, D. Ochoa, E. Gonçalves, J. Saez-Rodriguez, and P. Beltrao, "Benchmarking substrate-based kinase activity inference using phosphoproteomic data," *Bioinformatics*, vol. 33, no. 12, pp. 1845–1851, Jun. 2017, doi: 10.1093/bioinformatics/btx082.

[145] S. Yılmaz, M. Ayati, D. Schlatzer, A. E. Çiçek, M. R. Chance, and M. Koyutürk, "Robust inference of kinase activity using functional networks," *Nat Commun*, vol. 12, no. 1, Dec. 2021, doi: 10.1038/s41467-021-21211-6.

[146] D. Bradley, C. Viéitez, V. Rajeeve, J. Selkrig, P. R. Cutillas, and P. Beltrao, "Sequence and Structure-Based Analysis of Specificity Determinants in Eukaryotic Protein Kinases," *Cell Rep*, vol. 34, no. 2, Jan. 2021, doi: 10.1016/j.celrep.2020.108602.

[147] D. Szklarczyk *et al.*, "STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Res*, vol. 47, no. D1, pp. D607–D613, Jan. 2019, doi: 10.1093/nar/gky1131.

[148] C. C. Garbutt, P. v Bangalore, P. Kannar, and M. S. Mukhtar, “Getting to the edge: protein dynamical networks as a new frontier in plant–microbe interactions,” *Front Plant Sci*, vol. 5, 2014, [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpls.2014.00312

[149] P. Grosjean, “SciViews-R,” 2020. Accessed: May 19, 2022. [Online]. Available: https://www.sciviews.org/SciViews-R/

[150] J. Cuklina *et al.*, *Computational challenges in biomarker discovery from high-throughput proteomic data*, 2018, doi: 10.3929/ethz-b-000307772

[151] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed, “A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias.” [Online]. Available: http://www.bioconductor.org.http://www.stat.berkeley.edu/˜bolstad/normalize/index.html

[152] I. T. Jollife and J. Cadima, “Principal component analysis: A review and recent developments,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065. Royal Society of London, Apr. 13, 2016. doi: 10.1098/rsta.2015.0202.

[153] P. Fränti and S. Sieranoja, “K-means properties on six clustering benchmark datasets,” *Applied Intelligence*, vol. 48, no. 12, pp. 4743–4759, Dec. 2018, doi: 10.1007/s10489-018-1238-7.

[154] H. Wickham, “ggplot2: Elegant Graphics for Data Analysis,” Springer-Verlag New York, 2016.

[155] R. Kolde, “pheatmap: Pretty Heatmaps,” 2019, R package version 1.0.12, https://CRAN.R-project.org/package=pheatmap.

[156] R. Guha, “fingerprint: Functions to Operate on Binary Fingerprint Data,” 2018, R package version 3.5.7, https://CRAN.R-project.org/package=fingerprint.

[157] J. Oksanen *et al.*, “vegan: Community Ecology Package,” 2022, R package version 2.6-4, https://CRAN.R-project.org/package=vegan.

[158] D. Adler, D. Murdoch, “rgl: 3D Visualization Using OpenGL,” 2022, R package version 0.110.2, https://CRAN.R-project.org/package=rgl

[159] S. Lê, J. Josse, and F. Husson, “FactoMineR: A Package for Multivariate Analysis,” *J Stat Softw*, vol. 25, pp. 1–18, 2008.

[160] A. Kassambara and F. Mundt, “factoextra: Extract and Visualize the Results of Multivariate Data Analyses,” 2020, R package version 1.0.7, https://CRAN.R-project.org/package=factoextra.

[161] K. Soetaert, “plot3D: Plotting Multi-Dimensional Data,” 2021, package version 1.4, https://CRAN.R-project.org/package=plot3D.

[162] J. Ooms, “magick: Advanced Graphics and Image-Processing in R,” 2021, R package version 2.7.3, https://CRAN.R-project.org/package=magick

[163] S. Durinck, P. T. Spellman, E. Birney, and W. Huber, “Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt,” *Nat.Protoc.*, vol. 4, pp. 1184–1191, 2009.

[164] N. Xiao, D. Cao, M. Zhu, and Q. Xu, “protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences,” *Bioinformatics*, vol. 31, pp. 1857–1859, 2015.

[165] M. Soudy *et al.*, “UniprotR: Retrieving and visualizing protein sequence and functional information from Universal Protein Resource (UniProt knowledgebase),” *J Proteomics*, vol. 213, 2020.

[166] R. Trivedi and H. A. Nagarajaram, “Substitution scoring matrices for proteins - An overview,” *Protein Science*, vol. 29, no. 11. Blackwell Publishing Ltd, pp. 2150–2163, Nov. 01, 2020. doi: 10.1002/pro.3954.

[167] H. Pagès, P. Aboyoun, R. Gentleman, and S. DebRoy, “Biostrings: Efficient manipulation of biological strings,” 2020, R package version 2.66.0, https://bioconductor.org/packages/Biostrings

[168] J. C. Setubal and R. Braeuning, “Similarity Search,” A. Gruber, A. M. Durham, C.-V. Huynh, C. Y. Kao, P. T. Law, G. Miranda, K. Pinheiro, Ö. Tastan Bishop, A. Wallqvist, A. Gruber, A. M. Durham, C.-V. Huynh, C. Y. Kao, P. T. Law, G. Miranda, K. Pinheiro, Ö. Tastan Bishop, and A. Wallqvist, Eds., Bethesda (MD): National Center for Biotechnology Information (US), 2008, p. A05. doi: 10.1007/978-0-387-76605-5_5.

[169] R. Bevans, “An Introduction to T-Tests | Definitions, Formula and Examples,” *Scribbr*, Jan. 31, 2020.

[170] C. Wild *et al.*, “The Wilcoxon Rank-Sum Test.” *CHANCE ENCOUNTERS: A First Course in Data Analysis and Inference*, chapter 10, Dec. 1999, ISBN: 978-0-471-32936-7

[171] T. Suomi, F. Seyednasrollah, M. Jaakkola, T. Faux, and L. Elo, “ROTS: An R package for reproducibility-optimized statistical testing,” *PLoS Comput Biol*, vol. 13, May 2017.

[172] R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk, “Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments,” *FEBS Lett*, vol. 573, no. 1, pp. 83–92, 2004, doi: https://doi.org/10.1016/j.febslet.2004.07.055.

[173] D. Carratore *et al.*, “RankProd 2.0: a refactored Bioconductor package for detecting differentially expressed features in molecular profiling datasets,” *Bioinformatics*, vol. 33, pp. 2774–2775, 2017.

[174] G. Yu, L. G. Wang, Y. Han, and Q. Y. He, “clusterProfiler: an R package for comparing biological themes among gene clusters,” *Omics*, vol. 16, pp. 284–287, May 2012.

[175] K. Krug *et al.*, "A Curated Resource for Phosphosite-specific Signature Analysis," *Molecular & Cellular Proteomics*, vol. 18, pp. 576–593, 2019.
[176] A. J. Waardenberg, "KinSwingR: KinSwingR: network-based kinase activity prediction," 2020, R package version 1.16.0.
[177] P. v Hornbeck *et al.*, "PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse," *Nucleic Acids Res*, vol. 40, pp. D261-70, Jan. 2012.
[178] Z. Gu, L. Gu, R. Eils, M. Schlesner, and B. Brors, "circlize implements and enhances circular visualization in R," *Bioinformatics*, vol. 30, pp. 2811–2812, 2014.
[179] D. L. A. J. S. W. J. H.-C. M. S. N. T. D. J. H. M. P. B. L. J. J. D. Szklarczyk A.L. Gable and C. v. Mering, "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Res.*, vol. 47, pp. D607–D613, 2018.
[180] J. I. E. Hoffman, "Chapter 13 - Hypergeometric Distribution," in *Basic Biostatistics for Medical and Biomedical Practitioners (Second Edition)*, J. I. E. Hoffman, Ed., Academic Press, 2019, pp. 193–195. doi: https://doi.org/10.1016/B978-0-12-817084-7.00013-9.
[181] S. Ekins *et al.*, "Pathway mapping tools for analysis of high content data," *Methods Mol Biol.* 2007;356:319-50. doi: 10.1385/1-59745-217-3:319.
[182] P. Shannon *et al.*, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Res*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003, doi: 10.1101/gr.1239303.
[183] S. Mostafavi, D. Ray, D. Warde-Farley, C. Grouios, and Q. Morris, "GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function," *Genome Biol*, vol. 9, no. 1, p. S4, 2008, doi: 10.1186/gb-2008-9-s1-s4.
[184] F. Brüning *et al.*, "Sleep-wake cycles drive daily dynamics of synaptic phosphorylation," *Science (1979)*, vol. 366, no. 6462, p. eaav3617, Oct. 2019, doi: 10.1126/science.aav3617.
[185] Y. Hong *et al.*, "PhosPiR: an automated phosphoproteomic pipeline in R," *Brief Bioinform*, vol. 23, no. 1, p. bbab510, Jan. 2022, doi: 10.1093/bib/bbab510.
[186] B. Björkblom *et al.*, "Constitutively Active Cytoplasmic c-Jun N-Terminal Kinase 1 Is a Dominant Regulator of Dendritic Architecture: Role of Microtubule-Associated Protein 2 as an Effector," *The Journal of Neuroscience*, vol. 25, no. 27, p. 6350, Jul. 2005, doi: 10.1523/JNEUROSCI.1517-05.2005.

[187] P. Deshpande *et al.*, “Protein synthesis is suppressed in sporadic and familial Parkinson’s disease by LRRK2,” *The FASEB Journal*, vol. 34, no. 11, pp. 14217–14233, Nov. 2020, doi: https://doi.org/10.1096/fj.202001046R.

[188] J. T. Greenamyre, R. Betarbet, and T. B. Sherer, “The rotenone model of Parkinson’s disease: genes, environment and mitochondria,” *Parkinsonism Relat Disord*, vol. 9, pp. 59–64, 2003, doi: https://doi.org/10.1016/S1353-8020(03)00023-3.

[189] F. Llorens, A. Duarri, E. Sarró, N. Roher, M. Plana, and E. Itarte, “The N-terminal domain of the human eIF2beta subunit and the CK2 phosphorylation sites are required for its function,” *Biochem J*, vol. 394, no. Pt 1, pp. 227–236, Feb. 2006, doi: 10.1042/BJ20050605.

[190] A. G. Ryazanov, E. A. Shestakova, and P. G. Natapov, “Phosphorylation of elongation factor 2 by EF-2 kinase affects rate of translation,” *Nature*, vol. 334, no. 6178, pp. 170–173, 1988, doi: 10.1038/334170a0.

[191] L. P. Ovchinnikov *et al.*, “Three phosphorylation sites in elongation factor 2,” *FEBS Lett*, vol. 275, no. 1–2, pp. 209–212, Nov. 1990, doi: https://doi.org/10.1016/0014-5793(90)81473-2.

[192] A. G. Ryazanov and E. K. Davydova, “Mechanism of elongation factor 2 (EF-2) inactivation upon phosphorylation Phosphorylated EF-2 is unable to catalyze translocation,” *FEBS Lett*, vol. 251, no. 1–2, pp. 187–190, Jul. 1989, doi: https://doi.org/10.1016/0014-5793(89)81452-8.

[193] N. T. Price, N. T. Redpath, K. v Severinov, D. G. Campbell, J. M. Russell, and C. G. Proud, “Identification of the phosphorylation sites in elongation factor-2 from rabbit reticulocytes,” *FEBS Lett*, vol. 282, no. 2, pp. 253–258, May 1991, doi: https://doi.org/10.1016/0014-5793(91)80489-P.

[194] D. Flinkman *et al.*, “Regulators of proteostasis are translationally repressed in fibroblasts from patients with sporadic and LRRK2-G2019S Parkinson’s disease,” *NPJ Parkinsons Dis*, vol. 9, no. 1, p. 20, 2023, doi: 10.1038/s41531-023-00460-w.

[195] S. H. Ansarey, “Inflammation and JNK’s Role in Niacin-GPR109A Diminished Flushed Effect in Microglial and Neuronal Cells With Relevance to Schizophrenia,” *Frontiers in Psychiatry*, vol. 12. Frontiers Media S.A., Nov. 30, 2021. doi: 10.3389/fpsyt.2021.771144.

[196] R. L. Openshaw *et al.*, “Map2k7 Haploinsufficiency Induces Brain Imaging Endophenotypes and Behavioral Phenotypes Relevant to Schizophrenia,” *Schizophr Bull*, vol. 46, no. 1, pp. 211–223, Jan. 2020, doi: 10.1093/schbul/sbz044.

[197] G. M. Elias and R. A. Nicoll, “Synaptic trafficking of glutamate receptors by MAGUK scaffolding proteins,” *Trends Cell Biol*, vol. 17, no. 7, pp. 343–352, 2007, doi: https://doi.org/10.1016/j.tcb.2007.07.005.

[198] J. C. de Jonge, C. H. Vinkers, H. E. Hulshoff Pol, and A. Marsman, "GABAergic Mechanisms in Schizophrenia: Linking Postmortem and In Vivo Studies," *Front Psychiatry*, vol. 8, p. 118, Aug. 2017, doi: 10.3389/fpsyt.2017.00118.

[199] J. P. Rung, A. Carlsson, K. Rydén Markinhuhta, and M. L. Carlsson, "(+)-MK-801 induced social withdrawal in rats; a model for negative symptoms of schizophrenia," *Prog Neuropsychopharmacol Biol Psychiatry*, vol. 29, no. 5, pp. 827–832, 2005, doi: https://doi.org/10.1016/j.pnpbp.2005.03.004.

[200] X. Song *et al.*, "Mechanism of NMDA receptor channel block by MK-801 and memantine," *Nature*, vol. 556, no. 7702, pp. 515–519, 2018, doi: 10.1038/s41586-018-0039-9.

[201] J. Cox *et al.*, "Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ," *Mol Cell Proteomics*, 13(9):2513-26, Sep. 2014, doi: 10.1074/mcp.M113.031591

[202] T. Hinzke, "MaxQuant-Information and Tutorial." [Online]. Available: http://www.coxdocs.org/doku.php?id=maxquant:common:download_and_installation

[203] A. R. Dinasarapu, "Quantitative proteomics: label-free quantitation of proteins," *Github*, Sep. 13, 2021.

[204] M.-H. Lin *et al.*, "Benchmarking differential expression, imputation and quantification methods for proteomics data," *Brief Bioinform*, vol. 23, no. 3, p. bbac138, May 2022, doi: 10.1093/bib/bbac138.

[205] M.-X. Zhao, Q. Chen, F. Li, S. Fu, B. Huang, and Y. Zhao, "Protein phosphorylation database and prediction tools," *Brief Bioinform*, p. bbad090, Mar. 2023, doi: 10.1093/bib/bbad090.