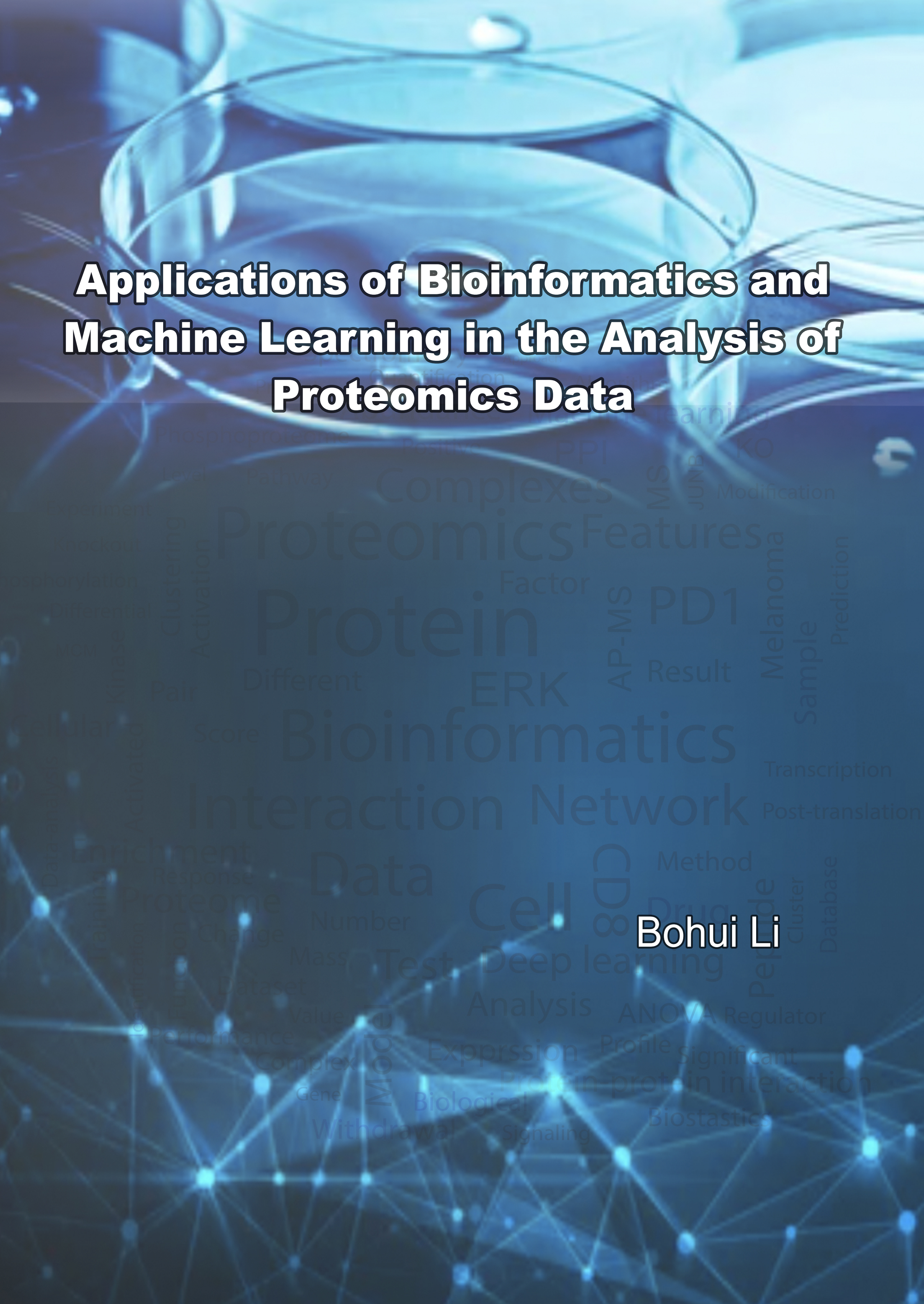
Applications of Bioinformatics and Machine Learning in the Analysis of Proteomics Data



Bohui Li

Applications of bioinformatics and machine learning in the analysis of proteomics data

Bohui Li
李伯会

ISBN: 978-94-6469-375-1

The research in this thesis was performed in the Biomolecular Mass Spectrometry and Proteomics Group, Utrecht University, Utrecht, The Netherlands

Applications of bioinformatics and machine learning in the analysis of proteomics data

Toepassingen van bioinformatica en machine learning bij de analyse van proteomics data

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 26 juni 2023 des middags te 2.15 uur

door

Bohui Li

geboren op 9 november 1986
te Guizhou, China

Promotor:

Prof. dr. A.F.M. Altelaar

Copromotor:

Dr. ir. B. van Breukelen

Contents

Chapter 1

General Introduction 7

Chapter 2

Proteomics and Phosphoproteomics Profiling of Drug-Addicted BRAFi-Resistant Melanoma Cells 45

Chapter 3

Exploring the role of PD-1 in CD8$^+$ T cell activation by Proteomic and phosphoproteomics profiling 73

Chapter 4

Identification of protein complexes by integrating protein abundance and interaction stoichiometries using a deep learning strategy 101

Chapter 5

Summary and future outlook 133

Chapter 1

General Introduction

1 General introduction

1.1 Proteomics: from genetic information to cellular function

The rapid development of high-throughput technologies has contributed to our understanding of the cell biology and complex diseases from a perspective of the molecular level. Three types of biomolecules, i.e., DNA, RNA, and proteins, are indispensable components in understanding cell activity and signaling (1). In the past decades, efforts have been made to build whole-genome-sequencing databases for numerous organisms, including human (2, 3), or to seek traits-associated genetic variations, such as disease-associated single-nucleotide polymorphisms (SNPs) (4, 5). The study of the whole genome within an organism is termed genomics (Figure 1) and is focused on DNA molecules. On the other hand, transcriptomics (Figure 1), is seeking to study the total RNA transcripts in the cells or tissues. Transcriptomics has become increasingly popular (6). Because gene transcription and subsequent RNA translation give rise to functional proteins, studying DNA and RNA expression would be expected to provide a good estimate for protein regulation. However, an increasing number of reports on mRNA and protein abundances find only a weak correlation between the respective abundances of RNA and proteins (7).

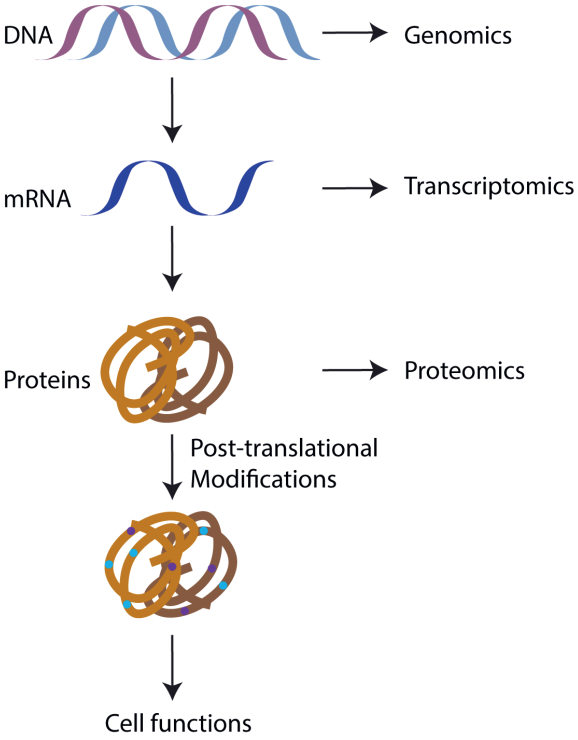


Figure 1. Biochemical context of genomics, transcriptomics, and proteomics.

Proteins are the executors of genetic information that is directly involved in signaling pathways, cellular growth, and maintenance. Moreover, the activity and function of proteins can be regulated by different expression levels, alternative splicing, and post-translational modifications (PTMs). Therefore, investigating the dynamics of proteins in cells and tissues, and unveiling the expression differences of proteins in complex diseases provide an intuitive reflection of cellular activities and disease pathologies. This brings the onset of proteomics (Figure 1), which is the study of identification and quantification of the proteins present in cells, or tissues (8). The following sections will give a brief overview of technologies in identifying and quantifying proteome, touching upon the subjects where bioinformatics methods play a role in the downstream data analysis and identification of protein complexes.

1.2 Mass spectrometry techniques

Proteomics is an approach focusing on the analysis of proteins or protein populations isolated from cells or tissues (9). Over the last century, mass spectrometry (MS) has emerged as an indispensable analytical technique for the characterization and quantification of the proteome. The basic MS system consists of several components, ion source (designed for the ionization of the target analytes), fragmentation (collision cell), mass analyzer (used to separate the gas phase ions by mass-to-charge ratio m/z), detector (designed for detecting the signals of ions and measuring their abundances) (10).

Generally, the high complexity of samples hinders the identification efficiency of the MS system, therefore, separation steps prior to admission into mass spectrometer are necessary. The introduction of pre-fractionation or separation into mass spectrometry using analytical techniques, including gas chromatography (GC) (11, 12), and liquid chromatography (LC) (13, 14) has dramatically reduced the complexity of the samples.

To detect and analyze the sample components based on their mass to charge ratios, ionization of the peptides or molecules is necessary. Several techniques have advanced the ionization of biomolecules, such as electrospray ionization (ESI) (15) and matrix-assisted laser desorption/ionization (MALDI) (16).

Fragmentation is a process of breaking the precursor ions retrieved from the ion sources into smaller product ions (fragment ions) (17). Different techniques are developed for the fragmentation of peptide molecules. The most widely used fragmentation technique for proteomics is collision-induced dissociation (CID) (18), where peptides are fragmented by energetic collisions with a neutral gas, often Helium, Nitrogen or Argon. Other recently developed fragmentation methods including electron transfer dissociation (ETD) (19), ultraviolet photodissociation (UVPD) fragmentation (20, 21), and dual fragmentation combining ETD and HCD (EThCD) (22) have improved the identification efficiency for the identification of post-translational modifications and intact proteins.

The mass analyzer is central to MS instrumentation. It performs the separation of ions based on their mass to charge ratio using electrical or magnetic fields. In the context of proteomics, its key parameters are sensitivity, resolution, mass accuracy and the ability to generate information-rich ion mass spectra from peptide fragments (tandem mass or MS/MS spectra) among the popular mass analyzers for proteomics research are the Orbitrap (23), ion trap (24) and time-of-flight (ToF) (25). These analyzers are designed with different performance, target molecules, resolution, and mass measurement accuracy, each with its own strengths and weaknesses.

Methods for protein quantification

Although recent developments in mass spectrometry allow the identification of thousands of peptides, proteins, and post-translational modifications (PTMs) from limited amounts of biological material, the quantification of differences between two or more physiological states of a biological system is among the most important but also most challenging technical tasks in proteomics (26). Quantifying protein or PTM abundances enables us to get insight into the changes of cells in response to stimuli, such as pathogen invasion and drug treatment. The quantification strategies can be classified into relative quantification and absolute quantification. The absolute quantification approach utilizes ultrapure synthesized peptides with incorporated stable isotopes and precisely determined absolute concentration as ideal internal standards to mimic native peptides formed by proteolysis (27). Such internal standard peptides are then used to measure the absolute expression of proteins and post-translationally modified proteins precisely and quantitatively after proteolysis by using a targeted MS analysis in a tandem mass spectrometer. The most used methods for relative quantification are: i) stable isotope labeling, either by metabolic or chemical labeling, and ii) label-free quantification (Figure 2) (26, 28, 29).

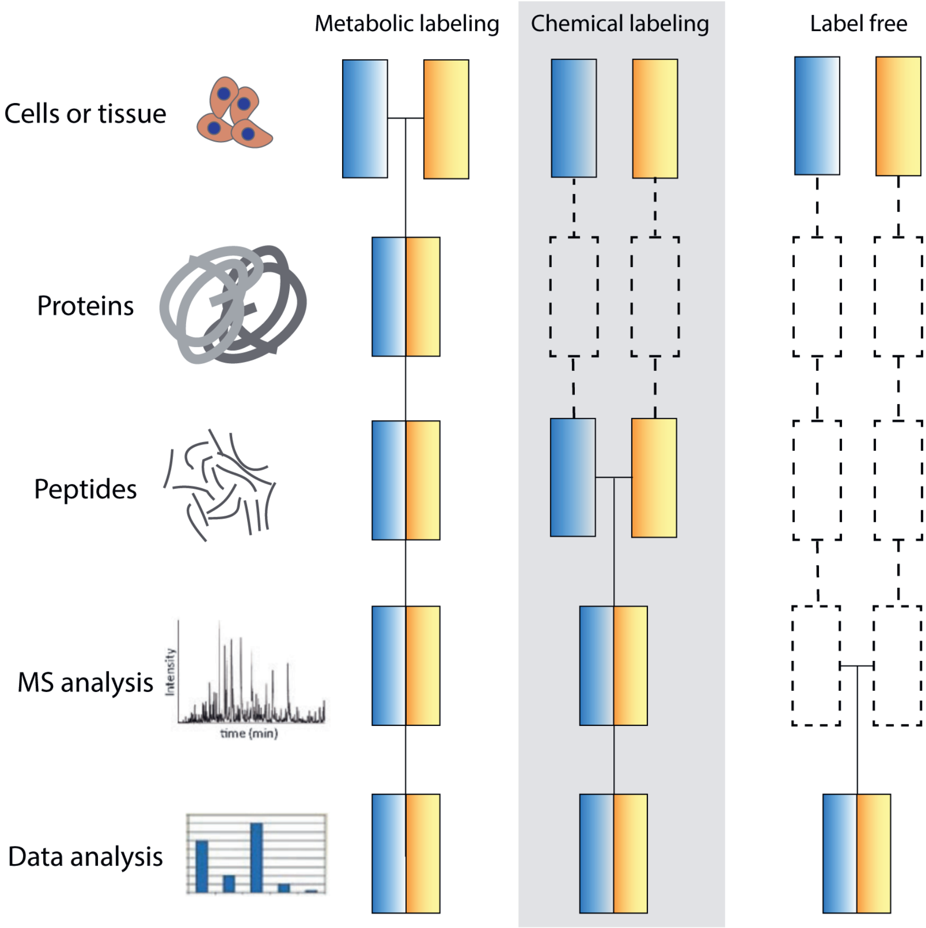


Figure 2. Schematic overview of alternative workflows which can be used in quantitative MS-based proteomics. Boxes in blue and yellow represent two experimental conditions. Horizontal lines indicate when samples are combined. Dashed lines depict steps of samples are treated in parallel, where quantification error can occur. Figure adapted from Ong et al (26) and Bantscheff et al (28).

Proteome quantification by stable isotope labeling

Stable isotope labeling was introduced into proteomics in 1999 (30, 31). This approach allows identification of equivalent peptides or proteins by utilizing the mass difference of the mass tags with stable isotopes. The common workflow is to tag protein or peptide with equivalent reagents, one of which includes a heavy mass tag and the other a light mass or no tag. The labeled samples can be recognized by a mass spectrometer and at the same time provide the basis for quantification. The basic assumption of isotope labeling approaches is that the labeling does not alter the physicochemical properties of a peptide or protein (32).

The stable isotopes can be introduced to proteomics samples via chemical labeling and metabolic labeling. The chemical labeling is a cell-free approach, examples include the isotope-coded affinity tag (ICAT) (33), 18O-labeling during proteolysis (34, 35), tandem mass tags (TMT) (36) or isobaric tags for relative and absolute quantification (iTRAQ) (37). On the contrary, the metabolic labeling is a cell dependent approach, where cells are cultured under heavy stable isotope enriched growth media. After a certain number of cell divisions or passages, the specific heavy isotopes will be incorporated in the newly synthesized proteins. One of the most well-established metabolic labeling methods is stable isotope labelling by amino acids in cell culture (SILAC) (38).

The differentially labeled samples are subsequently mixed in equal proportion for the mass spectrometry analysis. By utilizing the mass shift introduced by the isotopes with specific mass, the MS1 spectrum features (precursor ion signal) of the same peptide from different samples can be distinguished and quantified.

Label free quantification

Label-free quantification is a method in mass spectrometry that aims to determine the relative quantity of proteins in two or more biological samples. Unlike the isotope labeling methods for protein quantification, label-free quantification does not require additional chemistry or sample preparation steps. The label-free quantification can be achieved either by (a) measuring and comparing the signal intensity of peptide precursor ions or (b) counting and comparing the number of fragment spectra across different LC-MS/MS runs.

In the peak intensity-based method, label-free quantification is performed by computing extracted ion chromatograms (XICs) for all peptides over the whole LC-MS/MS run (39). Several parameters should be taken into consideration for complicated peptide mixtures in this peak intensity-based method. (i) To minimize the interfering signals of ions with similar mass, a mass spectrometer with higher resolution and mass accuracy is required. (ii) Specific methods have been developed to reduce the number of missing values, such as match between runs (40, 41). (iii) More sample replicates are advisable to increase reproducibility for ease finding corresponding peptides between different experiments. (iv) The right balance between the acquisition of survey and fragment spectra must be found. The spectral counting approach (42, 43) is based on the observation that the more of a particular protein is present in a sample, the more tandem MS spectra are collected for peptides of that protein. Therefore, relative quantification can be achieved by comparing the number of MS/MS spectra between a set of experiments. Compared to the peptide ion intensities strategy, the spectral counting approach benefits from extensive MS/MS data acquisition across the chromatographic time scale both for identification and quantification, while suffers from the inability to quantify low-abundant proteins with limited spectral counts.

1.3 Mass spectrometry data

A major aim of proteomics is to reveal the changes in protein expression at a global level, ideally monitoring all proteins present in cells or tissues. The shotgun proteomics approach is the most widely used method for the identification and quantification of proteins (14). With the recent advances in instrumentation, sample preparation, fractionation, and computational algorithms, it is quite routine to identify and quantify thousands of proteins in a single experiment (up to ten thousand). A classical bottom-up (shotgun) proteomics workflow is comprised of 4 major steps (Figure 3), i.e., sample preparation and digestion, fractionation or enrichment, MS data acquisition, and data analysis.

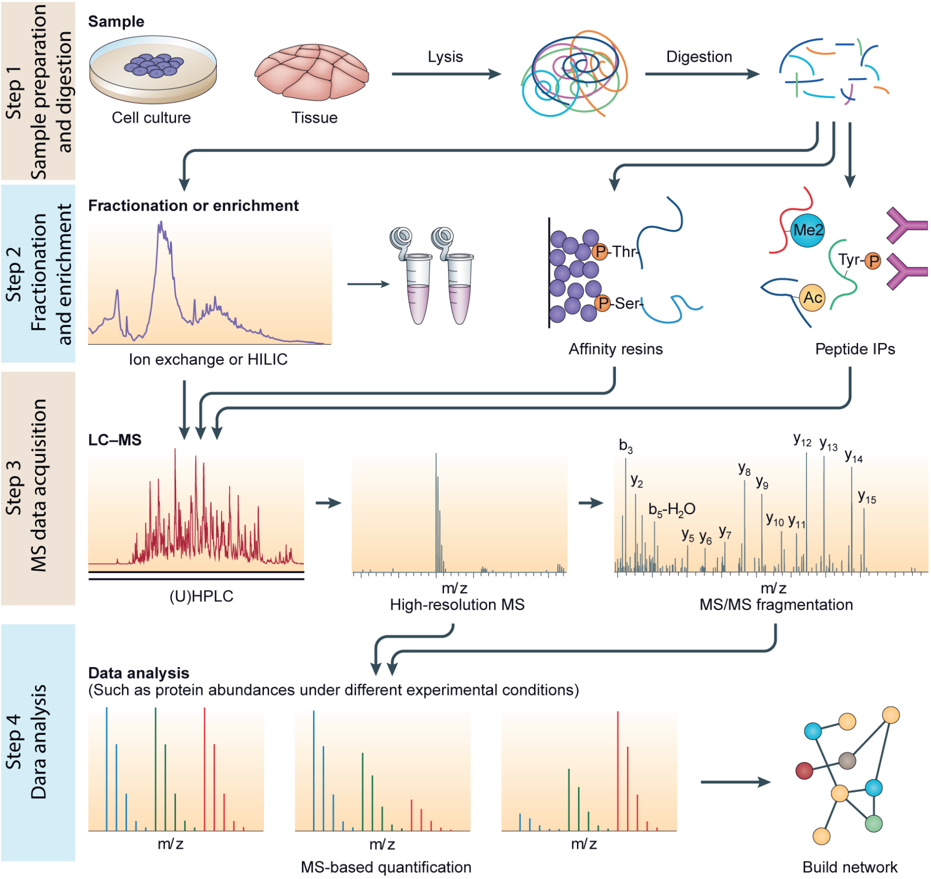


Figure 3. Overview of a generalized bottom-up MS-based proteomics workflow. Adapted from Altelaar et.al. (14). This generalized workflow consists of four steps: Step 1, Sample preparation and digestion, where the extracted proteins are digested into peptides; Step2, Fractionation and enrichment, in which the peptide population is subjected to liquid chromatography (LC) separation. Alternatively, specific subsets of the sample can be targeted through enrichment of peptides containing modifications (such as phosphorylation (P)) using affinity-based resins or antibody-based immunoprecipitation (IP); Step 3, MS data acquisition. The enriched samples are then introduced to the LC system for an additional separation, the eluted peptides from the LC are queried by the mass spectrometer to obtain their mass-to-charge (m/z) ratio; Step 4, Data analysis. The precursor m/z and its fragment ions are then matched to known peptide sequences using a search engine to obtain protein abundance. Subsequently, the protein abundance is analyzed and interpreted using specific computational methods to uncover the biological significance.

Sample preparation and digestion

Generally, the first step of sample preparation involves sample lysis and protein extraction from cells or tissues. The protein contents can be extracted either by physical disruption or reagent-based methods, where physical disruption methods uses lysis equipment such as a bead beater and a sonicator to disrupt tissues or cells, reagent-based methods utilize denaturants or detergents to lyse cells (44). In the digestion process, protease inhibitors and phosphatase inhibitors are often included to prevent nonspecific proteolysis and loss of phosphorylated peptides.

The protein mixture is usually subjected to specific proteases to digest proteins into smaller peptide before MS measurement. Trypsin is the most used protease for digestion as it recognizes and digests the carboxy-terminal at lysine (K) or arginine (R) residue (45), generating peptides with a desirable charge and length of approximately 6 to 25 amino acids that are usually unique to the protein and suitable for LC-MS identification. However, trypsin is less efficient in cleaving K-P or R-P bonds (46) and repeated basic residues (e.g. KK, RK), resulting in missed cleavages. As a result, alternative proteases with other cleavage specificities, such as ArgC, AspN, GluC, LysC, and LysN were designed for more comprehensive analysis of the proteome and improve sequence coverage (47-49). Therefore, a multiple enzyme strategy is recommended for the comprehensive analysis of complex proteomes.

Fractionation and enrichment

The digestion of proteins in the bottom-up proteomics analysis, on the one hand, enables the mass of peptides to fall within the mass range of the mass spectrometer, on the other hand, leads to the high complexity of the peptide mixture. Thus, appropriate separation or enrichment methods are essential to reduce the complexity before introducing the peptides into the mass spectrometer. The combination with nanoscale reversed-phase LC prior to MS analysis is still the dominating analytical technology for this purpose. In the recent years, serval additional fractionation strategies have become popular. Techniques such as strong cation exchange (SCX) separation, which can additionally enriching post-translation modified peptides in early fractions (50, 51), strong anion exchange (SAX) (52) and high-pH reversed phase chromatography (53) are increasingly used to boost the identification of peptides. Moreover, several technologies have been developed for the enrichment of phospho-peptides, such as titanium dioxide chromatography (TiO2) (54) and immobilized metal ion affinity chromatography (IMAC) (55).

MS data acquisition

Mass spectrometry is the crucial step for shotgun proteomics analysis. Ideally, all peptides eluted from the LC are identified by the mass spectrometer, however, many peptides elute simultaneously and compete for efficient ionization (i.e., high-abundant species can interfere with the identification of co-eluting less-abundant species, thus preventing the MS analysis of less-abundant species). Modern mass spectrometers collect three pieces of information from each peptide: its mass, its ion intensity and a list of its fragments (56). Upon ionization, the mass spectrometer first records the mass-to-charge (m/z) ratio, which is referred to as either MS, MS1 or survey spectrum. Subsequently, single peptides are selected and subjected to the collision cell for fragmentation, generating ‘b’ and ‘y’ fragmented ions (b-ions are amino-terminal fragment ions, y-ions are carboxy-terminal fragment ions) (14) and recorded in a second mass spectral scan that is referred to as MS2, MS/MS or tandem MS spectrum. The combination of precursor m/z and its fragment ions is then matched to peptide sequences from large protein databases using search algorithms, such as Mascot, SEQUEST, or MaxQuant.

Data analysis

Nowadays, advanced mass spectrometry allows rapidly, accurately, and sensitively identifying and quantifying thousands of proteins. Thus, proteomics technology has been successfully applied in biological research to depict protein-protein interactions, cellular signaling pathways to disease mechanisms, and identify biomarkers. In the meantime, this high-throughput method has also been applied to large-scale proteome and phosphoproteome profiling, generating a large amount of expression data from biological samples. Fortunately, the bioinformatics tools work as a crucial bridge connecting the identification generated from MS to the biological functions buried in the large-scale data. The following sections will detail the bioinformatics techniques in analyzing the MS data.

2 Proteomics data analysis

2.1 Analysis of quantitative proteomics data

Shotgun proteomics data are affected by a variety of known and unknown systematic biases as well as high proportions of missing values. Therefore, data normalization, is often the first step that needs to be taken and should be performed to remove systematic biases before statistical inference, sometimes followed by an imputation of missing values.

Data normalization

Systematic bias, alternatively defined as variation caused by nonbiological sources, is introduced by small variations in the experimental conditions in the course of carrying out the MS analysis (57). Bias may occur due to many factors including sample processing and handling conditions, instrument calibrations, LC columns, changes in temperature in the process of an experiment, etc. Normalization is the process that aims to account for the bias and make samples more comparable. Before normalization, log2-transformation of the recorded intensity is needed to reduce the dynamic range of intensity values (58). Such a step also converts the distribution of abundances of protein into a more symmetrical, almost normal distribution. In addition to the log transformation, most normalization methodologies were carried out by plotting data in a ratio versus intensity plot, or also commonly called as M versus

A (minus versus average) plot. Such ratio versus intensity plot enables an easier observation of linear or nonlinear trends resulting from biases. Several normalization strategies imported from microarray analysis have been successfully applied on proteomics data normalization (57, 59), such as linear regression normalization, local regression normalization, median normalization, and quantile normalization.

Linear regression normalization assumes that systematical bias is linearly dependent on the magnitude of peptide abundances (60). This method was performed by applying least squares regression to calculate a predicted peptide ratio that represents the deviation from the abscissa to the regression line. Local regression normalization assumes that systematic bias is nonlinearly dependent on the magnitude of peptide abundances (61). To remove such non-linear bias, linear regression analysis was performed on localized subdivisions of the peptide populations using the LOESS algorithm. The median normalization assumes that the samples of a data set are separated by a constant. It scales the samples so that they have the same median. The quantile normalization is based on the premise that the distribution of peptide abundances in different samples is expected to be similar. This normalization method can be achieved by replacing each point of a sample with the mean of the corresponding quantile (59).

Imputation of missing values

The widespread occurrence of missing values in MS quantitative proteomics is another challenge. There are several reasons that can lead to missing values and which can be caused by experimental conditions such as mis-cleavages, dynamic range issues, ionization competition, ion suppression, or as a result of data processing such as peptide mis-identification, ambiguous matching of the precursors in the quantitation step, etc. (62). Moreover, missing values occur when, for instance, the concentration of a peptide falls below the instrument detection limit; or lower abundant peptides that failed to pick (57). The common solution is to impute the missing values with the lowest sample values based on the hypothesis of detection limit or to impute them with the median/average value of that peptide/protein. Several other sophisticated statistical methods make use of an empirical distribution constructed from the quantified values to impute the missing values (58). For instance, a tail-based imputation strategy described by Kim et al. (63) imputed the missing values from the tail of the empirical normal distribution constructed from the quantified values across all proteins PTMs in a sample. In some cases, a protein or PTM contains a high percentage of missing values and should be removed before further processing.

Differential analysis

Basically, proteomics research aims to study the interrelation between protein expression and certain sample groups (e.g., distinct disease classes). One of the most common questions of this type of experiment is the comparison of protein expression profiles in two or more different types of biological samples, such as healthy and diseased tissues. Traditionally, fold change is commonly used in quantitative proteomic analysis where proteins differing by an arbitrary cut-off threshold in abundance are considered to be differentially expressed (64). Although it is a convenient and intuitive way to assess protein expression differences, fold change itself is not a statistical test that can indicate the level of confidence in differentially expressed proteins. Therefore, a statistical test is carried out on proteins or phosphorylation sites to assess a level of significance with specific p-values.

A statistical test consists of a null hypothesis (e.g., the mean of a population is equal to zero or a specific value) and an alternative hypothesis (e.g., the mean of a population is different from zero or a specific value). Based on the given data and certain assumptions on the probability distribution of the data, one decides to accept either the null or the alternative hypothesis. For instance, a classical one-sample t-test could be used to determine whether the expression of a protein in a group is significantly changed or not within biological or technical replicates. In this case, the abundance of proteins quantified in experiments is supposed to follow a normal distribution, which centers around the mean value (here is zero) and exhibits a width that depends on the variability. Alternatively, a two-sample t-test could be applied to experiments when trying to compare treated to untreated cells, wild-types to mutants, or samples from diseased to non-diseased subjects. In this case, the t-test evaluates whether the means of the two groups are statistically different from each other. T-statistic can be calculated according to a one-sample t-test or two-sample t-test formula (65). Given the t-statistic and degree of freedom for the test, the corresponding p-value can be mapped from the standard table of significance.

However, sample sizes are often small in biological research, which results in uncertainty in the sample variability estimates. Since these estimates are used in the test statistics to assess the statistical significance of the observed fold change, proteins exhibiting large fold change are often declared non-significant because of a large sample variance, while proteins with small fold changes might be declared statistically significant. Therefore, the LIMMA (linear models for microarray data) was introduced as an empirical Bayes approach that specifically allowed for a realistic distribution of biological variances (66). The statistical approach in LIMMA is to use the full data to shrink the observed sample variances towards a pooled estimate. This results in a more stable and powerful inference as compared to ordinary t-tests, particularly when the number of samples is small. To identify differentially expressed proteins among three or more conditions simultaneously, one might consider using the one-way analysis of variance (ANOVA). The ANOVA method assesses the relative size of variance among group means (between group variance) compared to the average variance within groups (within group variance). However, ANOVA analysis cannot provide detailed information on differences among the study groups. Therefore, further “multiple comparison analysis” tests are necessary to fully understand group differences in an ANOVA, which are post hoc tests.

Multiple testing correction

In large-scale proteomics data analysis, the statistical test is carried out on each protein separately, which may increase the number of false positives. This is because the repetition of the multiple tests may repeatedly add multiple chances of error, which may result in a larger α error level than the pre-set α level (67). Nowadays, proteomics analysis can routinely quantify thousands of proteins. That means, if testing a group of 5,000 proteins at a level of α of 0.05, a total of 250 of these false positives can be expected to be significant just by chance. Therefore, it is important to correct p-values in proteomics analysis, even in high-throughput experiments, as it helps to select the most significant changed proteins or PTMs for follow-up assays and avoid poor decisions (68).

The simplest and most widely used method of multiple testing correction is the Bonferroni adjustment (69). This method lowers the false discovery rate by dividing α by the number of tests. However, because of the conservativeness of the Bonferroni correction, this method could only be used in cases where the number of tests is relatively small (e.g, less than 5). An improved Bonferroni method called the Benjamini-Hochberg method (70) allows a fraction of the significant hits to be false although at the same time it helps to control the false discovery rate (FDR) for large-scale studies.

Functional annotation

Identification and quantification of proteins from a cellular proteome is most often not the purpose by itself. Actually, the interpretation of biological data and extraction of biological relevance from the vast amount of identified proteins are crucial to the understanding the complex mechanisms of biological systems. In the above sections, one can obtain the differentially expressed proteins within different groups (for example, healthy vs disease). The next step is to carry out the functional annotation, which seeks to uncover the biological relevance and to better understand the meaning of proteomics data. In order to interpret the proteomics data, many efforts have been devoted to developing biostatistics and bioinformatics tools (Table 1) (71). For instance, Ashburner et.al. developed a controlled vocabulary applicable to all eukaryotes, generating the Gene Ontology (GO) Consortium (72). Every gene or protein can thus be described by a finite number of vocabulary terms, which are classified into three GO categories or domains: biological process (BP), molecular function (MF), or cellular component (CC). This GO enrichment method employs a Fisher's exact test to result in a ranked list of GO terms, each term associated with a p-value.

The gene set enrichment analysis (GSEA) method is increasingly popular compared to GO analysis. GSEA defines a priori gene sets that have been grouped by their involvement in the same biological pathway, which can be found in the Molecular Signatures DataBase (MSigDB) (73). Instead of focusing on individual genes in a long list, GSEA is trying to identify classes of genes or proteins that are over-represented in a large set of genes or proteins and may have an association with disease phenotypes. With the application of the weighted

Kolmogorov–Smirnov-like statistic (74), the GSEA method calculates an enrichment score (ES) that reflects the degree to which a set S is overrepresented at the extremes (top or bottom) of the entire ranked list L (73).

Scientists around the world have different research questions, research backgrounds, and design assumptions. Therefore, researchers can choose one or more enrichment methods to execute the functional annotation, which allows the functional classification and the detection of the most represented biological terms of a gene/protein set. More enrichment methods and tools can be found in table 1.

Table 1. Computational tools for functional annotation (Adapted from Carnielli et.al. (75)).

	URL	Comments	References
Gene ontology			
GO	http://geneontology.org/	Functional annotation for genes or proteins	(72)
Matascape	https://metascape.org/gp/index.html#/main/step1	An online platform for hallmark enrichment analysis	(76)
BiNGO	http://www.psb.ugent.be/cbd/papers/BiNGO/Home.html	Tool for enrichment analysis on Cytoscape	(77)
DAVID	https://david.ncifcrf.gov/	Meta-tool for functional analysis of large gene lists	(78)
GeneMANIA	http://genemania.org/	Tool to identify the most related genes to a query gene	(79)
GSEA	https://www.gsea-msigdb.org/gsea/index.jsp	A method that determines whether a priori defined set of genes/proteins shows statistically significant	(73)
Pathway analysis			
KEGG	http://www.genome.jp/kegg/	Database resource for pathway analysis	(80)
Reactome	http://www.reactome.org/	Tool for pathway analysis	(81)
Interaction networks			
Cytoscape	http://www.cytoscape.org/	Open-source software for integration, visualization and analysis of biological networks	(82)
IIS — Integrated Interactome system	http://bioinfo03.ibi.unicamp.br/lnbio/IIS2/index.php	Integrative platform for the annotation, analysis and visualization of the interaction profiles of proteins/genes, metabolites and drugs of interest	(83)
STRING	http://string.embl.de	Database tool for direct or indirect protein interactions	(84)

Protein-protein interaction networks

Studying the biological systems from the network perspective is of great importance (85-87). A biomolecular network is a widely accepted form by which biomolecules interact with one another to perform and maintain their functions in cells, tissues, and organs (88). A typical network consists of two entities: nodes (also known as vertices) and edges (also known as links) (Figure 4A). According to the differences of edges, networks can be summarized into unweighted or weighted, and undirected or directed (Figure 4A). A directed network is connected by edges pointing in a direction, such as transcription factor (TF)–target network (89), drug-targets network (90). However, in real biological systems, a network can have multiple characteristics at the same time, for example it can be an undirected and weighted network, or a directed and weighted network.

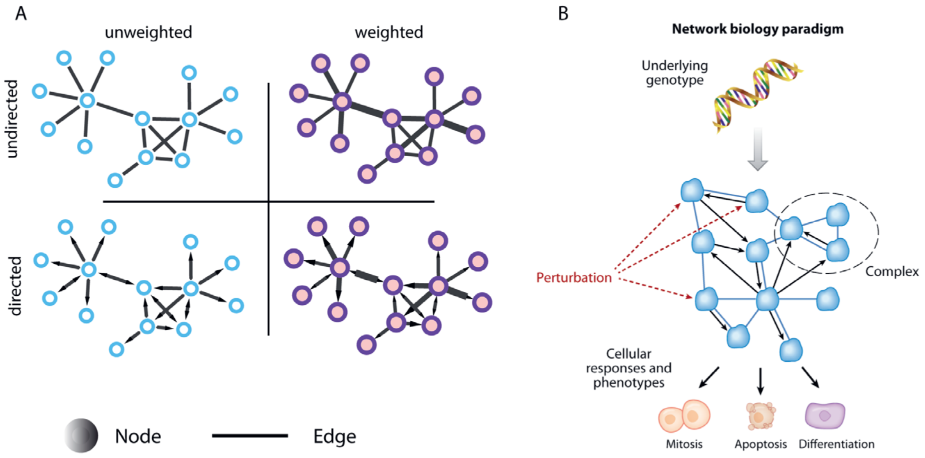


Figure 4. (A) Network types according to the properties of edges. (B) Network biology paradigm, networks of interacting molecules are placed. Nodes (represented in blue) represent molecule of interest, such as a genes, proteins, or metabolites. Adapted from Bensimon et al (91).

At the global level, the biological network comprises all manner of interactions among biological entities (Figure 4B). In this type of network, each node represents a molecule of interest, such as a gene, a protein, or smaller molecules, such as cofactors, metabolites, and messenger molecules. The edge between two nodes represents a direct or indirect relationship, such as an enzymatic reaction, a physical interaction, or a functional connection (91). Such a network paradigm is concerned with the network nodes and edges, placing networks of interacting molecules between genotype and phenotype (Figure 4B).

To simplify the analysis, biological networks are usually sub-categorized; the most common being protein-protein interaction (PPI) networks. The application of PPI networks translates expression maps into mathematical models for biomarker/drug-target discovery, phenotype correlation analysis (92), and complexes isolation and identification (93, 94). Over the past years, the number of reported protein-protein interactions has increased substantially and is still increasing. To provide researchers access to the PPI information, a number of publicly available databases have set out to collect and store protein-protein interaction data. Such as the Biological General Repository for Interaction Datasets (BioGRID) (95), the Molecular INTeraction database (MINT) (96), the Database of Interacting Proteins (DIP) (97), the IntAct molecular interaction database (IntAct) (98), the Human Protein Reference Database (HPRD) (99) and the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (84, 100) (see Table 2). These PPI databases catalog experimentally determined interactions between proteins from original publications.

Table 2. Public protein-protein interaction databases

Database	URL	Interactions	Publications	Organisms
BioGRID	https://thebiogrid.org/	1,677,595	77,459	71
MINT	http://mint.bio.uniroma2.it/mint/	235,000	4,750	30
DIP	https://dip.doe-mbi.ucla.edu/dip/Main.cgi	53,431	3,193	134
IntAct	https://www.ebi.ac.uk/intact/	1,139,018	22,368	131
HPRD	http://hprd.org/index_html	41,327	453,521	1
STRING	https://string-db.org/	296,567,750 (score >= 0.9)	NA	5,090

Among the above PPI databases, STRING is one of the most popular biological databases and web resources, which provides known and predicted protein-protein interaction (100-103). The STRING database is freely accessible, user friendly, and regularly updated. The resource also serves to highlight PPI networks and functional enrichments in user-provided lists of proteins, using a number of functional classification systems such as GO, Pfam and KEGG. Moreover, it also features a number of additional data access points, such as access through a Cytoscape app (http://apps.cytoscape.org/apps/stringapp), as well as download pages covering user-interested PPI networks/pathways, individual species networks and associated data. The STRING database allows scientists to obtain an intuitive view of protein-protein interactions and associated functions and pathways, which helps biologists understand cellular processes on the system level.

With the substantially increasing amount of the protein-protein interaction (PPI) data, more and more protein function prediction methods have been established. For instance, edge-betweenness clustering (104) method separates PPI into subgraphs of interconnected proteins, using the Girvan and Newman’s Edge-Betweenness algorithm to predict protein functions. The CFinder (105) performs a search for dense subgraphs (groups) of nodes in undirected networks, which allows users to predict the function of a single protein and to discover novel modules. Besides, the protein-protein interaction networks act as important role in identifying protein complexes (106). To this end, many successful methods have been developed based on different searching strategies, such as MCL (107) and RW (108) employed Flow Simulation, focusing on the imitation of ways in which information spreads through a network; ClusterONE (109) and SE-DMTG (110) employed the modularity, topological structure, and overlapping information to predict protein complexes.

Although proteins are key functional entities in the cell, the activity of proteins is regulated by different post-translational modifications (PTMs). The PTMs, such as acetylation, glycosylation, ubiquitylation, and phosphorylation, involve in modulating critical biological processes such as protein signaling, localization, and degradation and have been implicated in a wide variety of pathologies. In the next section, we will discuss how to analyze the post-translational modifications data (mostly phosphorylation data).

2.2 Post-translational modifications (PTMs)

Protein post-translational modifications (PTMs) increase the functional diversity of the proteome by the covalent addition of functional groups or proteins. The PTM modifications include phosphorylation, glycosylation, ubiquitination, acetylation, lipidation and proteolysis that regulate almost all aspects of normal cell biology and pathogenesis. As a result, it is estimated that the human genome (~20,000 protein-coding genes) may potentially produce on the order of 1.8 million different protein species by post-translation modification (111, 112) called proteoforms. Nowadays, more than 400 different PTM types are listed in the Uniprot database (113) and many more are still being discovered. Considering the complexity and diversity of PTMs, the number of potentially modified residues, the dynamic characteristic, and the often-low stoichiometry of these modifications, one realizes that it is a challenge to identify and localize the sites of these modifications. Advancements in proteomics methodologies have greatly improved the analysis of mapping and quantifying the sites of these PTMs. Among those different PTM types (such as acetylation (114), glycosylation (115), and ubiquitylation (116)), phosphorylation is the best characterized PTMs.

Nowadays, it is quite routine to identify and quantify tens of thousands of phosphorylation sites in a single experiment (117, 118). As the volume of “omics” data for phosphorylation continues to grow, exploring the biological functions and significance of such PTMs has become a major challenge. In the last decades, bioinformatics efforts have been intensively implemented to address these challenges. These can be roughly summarized into three categories: 1) gene-centric pathway analysis, 2) kinase activities inferring from phosphoproteomic data, and 3) site-centric signaling analysis.

Gene-centric pathway analysis

Gene-centric pathway analysis typically involves integrating phosphorylation changes of multiple sites on the same gene/protein by calculating the mean or median of the phosphorylation abundance (50, 119), or determine a specific deviation from the mean across a sample cohort (120), or rely on statistical testing based on replicate analyses to extract regulated phosphosites from a large-scale dataset (121). The resulting gene-centric expression matrix can then be queried against gene-centric pathway databases such as Reactome, KEGG, or MSigDB, in which each entry comprises a collection of genes in a biological pathway. However, this simple gene-centric pathway analysis method leaves out the multiplicity of phosphorylation sites on a single protein, different sites on protein isoforms, and most importantly the functional consequence of the phosphorylation event, which can be activating, inhibiting, leading to conformational changes, etc. This leads to the loss of critical information such as activating phosphorylation sites localized in the activation loop of kinases or inhibitory sites in close proximity.

Kinase activities inferring from phosphoproteomic data

Reversible phosphorylation plays a key role in nearly every cellular process by regulating the activity, localization, and interaction of proteins (122). Therefore, it is crucial to identify the kinase-substrate interaction involved in cell signaling pathways. To date, the PTM site information, as well as corresponding kinases, were implemented into several databases, such as PhosphoSitePlus (123), PHOSIDA (124), and Phospho.ELM (125). Based on these kinase substrate interactions, several bioinformatic tools, such as Inference of Kinase Activities from Phosphoproteomics (IKAP) (126) and Kinase Set Enrichment Analysis (KSEA) (127), can be directly applied to site-centric datasets. Such tools exploit the kinase-substrate relations to derive the activity of a kinase from the phosphorylation state of its substrates. Recently, an R package called InKA (Integrative Inferred Kinase Activity) was developed to infer kinases activities by integrating kinase-centric (i.e., kinome and activation loop) and substrate-centric (i.e., PhosphoSitePlus and NetworKIN) information (128). This approach achieves an optimized ranking of inferred kinase activities based on MS-derived phosphoproteomics data for single samples.

Site-centric signaling analysis

Recent integrative approaches have been developed to help get closer to extracting meaningful information rather than just providing long lists of PTM sites. The PHOTON tool (129) takes sets of differentially quantitated PTMs and protein-protein interaction network information as inputs, and through network-based statistical modeling, generates scores to identify significantly functional signaling proteins. Alternatively, Krug and colleagues (122) developed an approach that integrates modified versions of the single sample gene set enrichment analysis approach tailored to the PTM specific context (PTM-SEA making use of PTMsigDB database).

Proteins do not work independently but interact with each other in stable or transient multi-protein complexes of distinct composition. These complexes have essential roles in the cell cycle, transcription and translation, signaling cascades and cellular functions. In the fourth chapter, we are trying to develop a deep learning framework to predict protein-protein interaction probabilities, and protein complexes. Therefore, in the next section, we will introduce the basics of machine learning, which covers unsupervised machine learning, supervised machine learning, and basics in machine learning modeling.

2.3 Machine learning algorithms for proteomics data analysis

The exponential growth of the amount of proteomics data raises a daunting problem: the extraction of useful information buried under data. This problem is one of the main challenges in bioinformatics and computational biology. Fortunately, machine learning (ML) algorithms provide new promising approaches, designed to facilitate pattern recognition, classification, group clustering, and prediction, based on models derived from existing data. Machine learning algorithms can parse through voluminous data and pick up expression patterns or clusters that would otherwise go unrecognized to the human eyes. This approach has allowed biologists to uncover underlying biology of large-scale omics datasets.

Machine learning is increasingly applied to MS-derived proteomics data research problems. Recent increases in the amount and sensitivity of proteomics data collection have aggravated the requirement of machine learning models. Models have been applied to predict peptide properties (130) from only a primary sequence, including tandem mass spectra (131), ion mobility (132), and retention time (133). Furthermore, machine learning algorithms have been utilized in peptide identification and protein inference (134, 135). Apart from applications on pre-processing of the MS data, machine learning approaches have been widely used for downstream analysis of MS data. For example Deeb et al. (136) used protein expression profiles to perform classification for patients with diffuse large B-cell lymphoma. Dan et al. (137) used a support vector machine classifier to identify diagnostic markers for tuberculosis by proteomic fingerprinting of serum. Drew et al., (138) integrated over 9,000 mass spectrometry experiments and built a support vector machine classifier to predict protein-protein interaction probabilities.

Depending on how an algorithm is being trained and on the basis of availability of the output while training, machine learning paradigms can be classified into unsupervised learning and supervised learning (139).

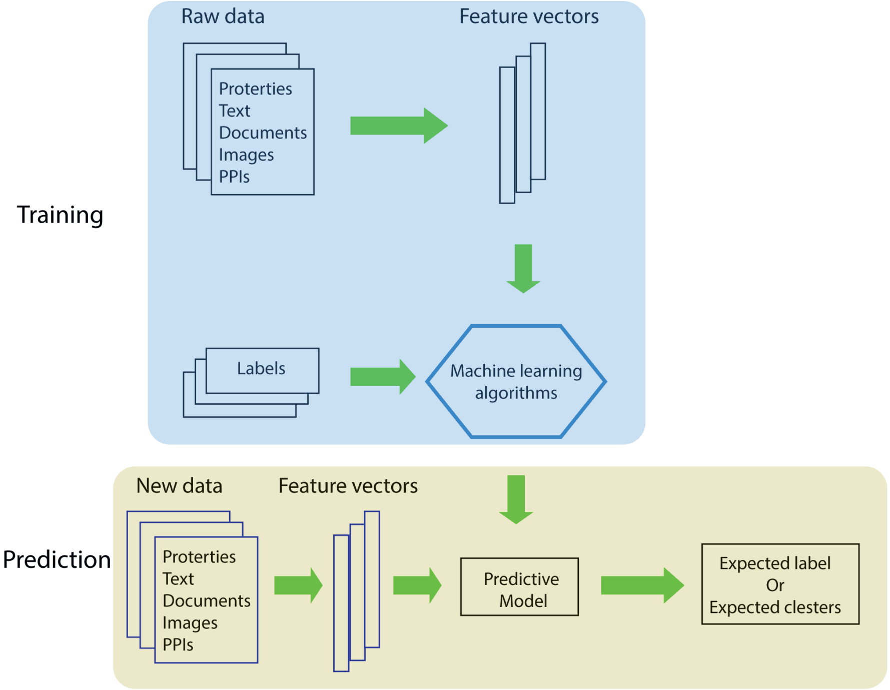


Figure 5. Simplified overview of a machine learning workflow. It consists of two main parts: training and prediction. The training process is executed on input raw data gathered from text, documents, images, and protein-protein interactions et.al. Once an optimal machine learning model is obtained, it can be applied to make predictions for new input data.

Unsupervised learning

Unsupervised learning is a form of machine learning that requires no labelled data for the training process (140). The training dataset is used in training the machine learning model, whereas the testing dataset helps in predicting the correct values and improve model accuracy. The machine predicts the outcome based on past experiences and learns from the previously introduced features to predict the real-valued outcome. The classic example of unsupervised learning is clustering (K-means clustering, principal component analysis) (141), which is designed to create groups or clusters automatically.

Supervised machine learning

In contrast to unsupervised classification or clustering, the supervised machine learning involves training a model based on data inputs that have specific class labels associated with them (140). Specifically, the supervised machine learning is the process of learning a set of rules from instances, creating a classifier that can be used to generalize from new instances (142). A variety of supervised machine learning algorithms have been frequently carried out, including Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), and Neural Networks (Perceptron) et.al. As we have employed the classification strategies Support Vector Machine (SVM) and Deep Learning (Neural Networks) in our thesis, we will introduce these two methods in the following sections.

Support vector machine is one of the most popular supervised machine learning algorithms proposed by Vapnik and co-workers (143-145). It differs from traditional methods which minimize the empirical training error, while the SVM algorithm aims at minimizing an upper bound of the generalization error through maximizing the margin between the separating hyperplane (line) and the data (Figure 6). This linear classifier tries to extend data into a high-dimensional (possibly infinite-dimensional) feature space, by using of the kernel functions (146). The general training process consists of two steps: in the first step a primary kernel is used to obtain support vectors; in the second step the modified kernel is used to obtain the final optimal classifier.

The SVM algorithm has been successfully applied to MS data analysis. For example, Wang et.al constructed an SVM scorer for peptide identification (147, 148). Fernández and co-workers developed an SVM classifier to detect ovarian cancer from metabolomics liquid chromatography/mass spectrometry data (149). A majority of SVM-based PPI-prediction methods are based on protein-primary sequences as input features (150, 151), since sequence information is more accessible for most proteins. Although the wide application of SVM, more care should be taken when using this promising algorithm because of its overfitting, in which the models showed high performance with internal testing datasets while behaving less in their prediction ability concerning new data.

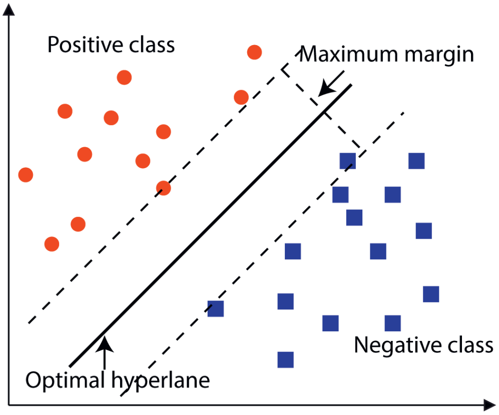


Figure 6. An overview of a two-class support vector machine classifier. The SVM classifier tries to find a separating optimal hyperplane (line) that best splits the data into classes. The larger the margin between the data and the hyperplane the better the separation.

Deep learning also known as deep structured learning, is part of the machine learning methods based on artificial neural networks (Figure 7A), which is inspired by the human nervous system and the structure of the brain (152). In other words, deep learning focuses on training computers to mimic how people learn things.

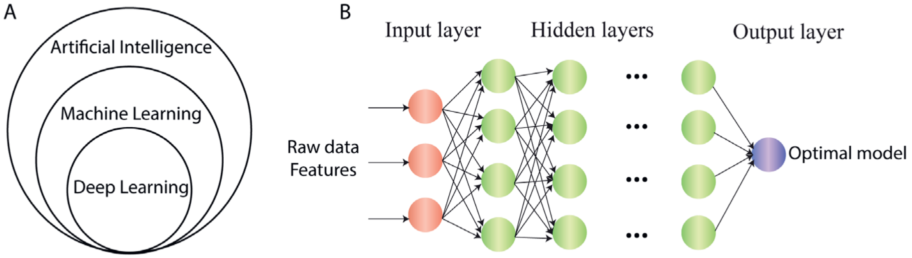


Figure 7. Simplified deep learning. (A) An overview of the relationship between artificial intelligence, machine learning, and deep learning. (B) The simplified structure of the deep learning strategy that consists of an input layer, one or several hidden layers, and an output layer.

Traditionally, the performance of machine-learning models highly relied on the goodness of the representation of the input data (features). A good data representation is a necessary requirement for obtaining models with higher performance. Therefore, for an extended period, feature engineering and selection have been crucial steps in machine learning, which focuses on building features from raw data. Comparatively, deep learning algorithms could automatically perform feature extraction (153, 154). Deep learning consists of an input layer, hidden layers, and an output layer (Figure 7B). The nodes or units in each layer are neurons that are interconnected in adjacent layers. In neural networks, inputs and outputs are connected to the neuron by weights, which are linear operators that multiply the previous value. Then, the units undergo a transformation with a specific activation function, which in most cases is a sigmoid function, tan hyperbolic, or rectified linear unit (ReLU).

Deep learning has already been applied to various fields of biological research, including the analyses of medical image data, gene expression data, DNA and protein sequence data, protein quantification data, etc. For instance, several deep learning-based tools have been developed for the prediction of retention time, including DeepRT (155), DeepMass (156), DeepDIA (157), and DeepLC (158); for the prediction of MS/MS spectra, such as pDeep (159), DeepMass:Prism (156), DeepDIA (157); for the prediction of Post-Translational Modifications (PTMs), such as the prediction of PTM sites for phosphorylation (160), ubiquitination, acetylation, and glycosylation; for the prediction of protein-protein interactions (161). Deep learning technology has great potential in many research tasks. With continuous improvements to the deep learning algorithms and the generation of high-quality proteomics data, we expect deep learning will have a profound impact on the application of proteomics data analysis.

Machine learning modeling

Training and testing. When establishing a new machine learning model, it is standard practice to divide data into two sets: the ‘training set’ used to train the model, and the ‘test set’, used to evaluate the performance of the final model. In practice, it is prone to overestimating if we pull all of our data to train up a model, and then use the same set of data for testing. Thus, a general rule of data splitting is based on a ratio of 70/30 or 80/20 for producing training and testing datasets. In general, the size of the training set has an important impact on the prediction ability of the ML models. The more data seen by the training process the better the chance of the algorithm to detect characteristics for making predictions. Ideally, an abundance of both training and test data is available for the machine learning model prediction to be optimal.

Cross validation. Cross-validation is similar to splitting training and test data but it contains more subsets. In practice, we split data into equal-sized subsets, termed folds. For each fold, we remove it from the training set, build a model on the other folds and then test on the withheld portion. If we have k folds, then this is called k-fold cross-validation. For instance, if we use 15 folds cross-validation, which means the dataset is divided into 15 equal-sized subsets, the first subset is taken as the test set, and the remaining 14 subsets are the training sets. Next, we put the first fold back into the training set but take out the second fold as a test set, and repeat these training and testing processes. Doing this ensures that each piece of the dataset is helpful for training and testing and therefore helps to shrink the bias.

Model performance evaluation. A crucial step in machine learning modeling is to evaluate the model performance. Most measures in use today focus on a classifier’s ability to identify classes correctly. These measures are built from a confusion matrix which records correctly and incorrectly recognized examples for each class (162). Table 3 presents a confusion matrix for binary classification, where TP is true positive, FP is false positive, FN is false negative, and TN is true negative counts. Specifically, a true positive (TP) is an outcome where the model correctly predicts the positive class. Similarly, a true negative (TN) is an outcome where the model correctly predicts the negative class. A false positive (FP) is an outcome where the model incorrectly predicts the positive class; and a false negative (FN) is an outcome where the model incorrectly predicts the negative class.

Table 3. A confusion matrix for binary classification

Classes	**Positive**	**Negative**
Positive (P)	True positive (TP)	False negative (FN)
Negative (N)	False positive (FP)	True negative (TN)

The most commonly used measures now are accuracy, precision, recall, F-measure (or F-score), and ROC analysis (163). The accuracy is the proportion of true results, i.e., both true positives and true negatives, among the total number of cases examined. Accuracy can be determined using the equation:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

The precision attempts to indicate the proportion of positive identifications was actually correct. It is calculated using the equation:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

The recall attempts to answer what proportion of actual positives was identified correctly. It can be calculated using the equation:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Both the precision and recall focus only on the positive examples and predictions, although these measures capture some information about the rates of errors made. However, neither of them captures information about how well the model handles negative cases. Therefore, the F-measure or F-score was proposed, which is the geometric and harmonic means of the precision and recall.

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

Another way to test the performance of classifier can be obtained by the ROC curve (receiver operating characteristic curve).

$$\text{True positive rate} = Recall = \frac{TP}{TP + FN} \quad (5)$$

$$\text{False positive rate} = \frac{FP}{FP + TN} \quad (6)$$

This curve plots two parameters: true positive rate, and false positive rate, where X axis represents the false positive rate and the Y axis represents true positive rate. The area under the ROC Curve is called for AUC, which provides an aggregate measure of performance across all possible classification thresholds.

These measures are commonly used to present results for binary decision problems in machine learning. However, when dealing with highly skewed datasets, Precision-Recall (PR) curves give a more informative picture of a model’s performance. Therefore, it’s important to choose a proper evaluation measure depending on the dataset.

3. The application of bioinformatics tools for (phospho)proteomics data analysis

Recent advances in mass spectrometry (MS)-based proteomics and phosphoproteomics have enabled tremendous progress in the uncovering cellular mechanisms, disease progression, and the relationship between genotype and phenotype. The general downstream workflow of bioinformatics analysis in mass spectrometry-based (phospho)proteomics experiments could be summarized into 6 steps: 1, Data cleaning, and normalization; 2, Differential analysis and heatmap for protein abundance with expression clustering, or volcano plot highlighting differentially expressed hits; 3, Clustering analysis to check the reproducibility for biological replicates; 4, Functional annotation to uncover the biological relevance of differentially expressed proteins; 5; Network inference to show the protein-protein interaction relationships and indicate the key regulators for the network/module; 6, Inferring of kinase activities or phosphosite-based signature enrichment for phosphoproteomics data interpretation. This general workflow could be adjusted according to one’s experimental design to answer specific research question.

3.1 Melanoma and treatment strategies

Melanoma is a type of skin cancer that develops from the pigment-producing cells known as melanocytes, which typically occur in the skin but may occur in eyes, inner ear, and leptomeninges. Although melanoma accounts for about 1% of all skin malignant cancer cases, the malignant melanoma occupies the most aggressive and the deadliest form of skin cancer (164). Studies have shown that melanoma is associated with a high variety of somatic mutations (165), most frequently involving BRAF (35–45%) and NRAS (15–25% of melanoma patients) genes, but also c-KIT and PTEN (166). Among these frequently mutated genes, BRAF is a serine-threonine kinase involved in the RAF-MEK-MAPK pathway, which plays a role in regulating MAP kinase/ERKs signaling pathway and affects cell division, differentiation, and secretion. Firstly, extracellular signals and stimulus bind to tyrosine kinase receptors, leading to the activation of RAS, and then activating BRAF. Then, the activated BRAF can phosphorylate and activate MEK1/2 kinases, which in turn phosphorylates and activates ERK1/2, leading to cellular proliferation and survival (167). However, oncogenic mutations (mostly V600E) of BRAF account for around 40 to 60% of melanomas, resulting in the constitutive activation of downstream MAPK signaling and unregulated cell growth (168).

Small-molecule targeted therapies that works by blocking the mutant BRAF V600E involved pathways (169) showed major tumor responses compared to chemotherapy. Inhibitors such as vemurafenib and dabrafenib are the most effective, approved treatments for BRAF positive melanoma (170). Although, BRAF inhibitors showed a lot of potential in melanoma treatment, with remarkable response rates and overall survival, the majority of patients develop resistance rapidly (171). To increase the molecular understanding of this drug dependency, we applied a mass spectrometry-based proteomic approach on BRAFi-resistant melanoma cells (**Chapter 2**), in which ERK1, ERK2 and JUNB were silenced separately using CRISPR–Cas9. By applying the above bioinformatics methods for the downstream proteomics and phosphoproteomics data analysis, we depict how ERK1, ERK2 and JUNB influence the proteome response of drug addicted melanoma cells upon drug withdrawal.

More recently, immunotherapy shows promising clinical efficacy in the treatment of melanoma. Immunotherapy is aimed at stimulating the immune system against the tumor, via an enhanced ability to recognize and kill cancer cells (172). The most successful immunotherapy on melanoma is the immune checkpoint inhibiting, including the CTLA-4 inhibitors (ipilimumab and tremelimumab) and PD-1/PD-L1 inhibitors (pembrolizumab, pidilizumab, and nivolumab) (173). In **Chapter 3**, we focused our research on CD8+ T cells to increase our understanding of the T cell activation process. Moreover, to investigate the role of PD-1 in regulating T cell activation, we measured the proteomics and phosphoproteomics profiling on resting and activated CD8+ T cells, in which PD-1 was silenced using CRISPR–Cas9. With the application of bioinformatics tools, we find that silencing of PD-1 induced more phosphorylation events in regulating mTOR signaling and activated the epidermal growth factor and corresponding downstream MAPK pathway.

3.2 Identification of protein complexes

Proteins do not work independently but interact with each other in stable or transient multi-protein complexes of distinct composition. Moreover, proteins can interact with other molecules, such as DNA (174), RNA (175) or metabolites (176), highlighting the importance of identifying protein interactions and protein complexes. A workflow for identification of protein complexes can be simplified into protein-protein interaction data acquisition, PPI scoring, and identification of complexes (Figure 8).

Probing protein-protein interaction

Different experimental techniques have been developed to measure protein-protein interactions; these methods vary considerably, not at the least in terms of the data they produce. The two most well-established methods are the yeast two-hybrid (Y2H) system (177) and affinity purification followed by mass spectrometry (AP-MS) (178). The Y2H system assays whether two proteins physically interact with each other by genetically modifying yeast strains to express a 'bait' and a 'prey' protein, which, if they interact, trigger the expression of a reporter gene. The high-throughput Y2H method can construct broad maps of binary PPIs, irrespective of protein abundances, including those that connect different complexes and those that are of a highly transient nature, which are difficult to target using alternative approaches. However, the quality of Y2H data sets has been controversial, since different Y2H systems have been shown to detect markedly different interactions in the same interactome, requiring tools to determine the confidence of the interactions (179).

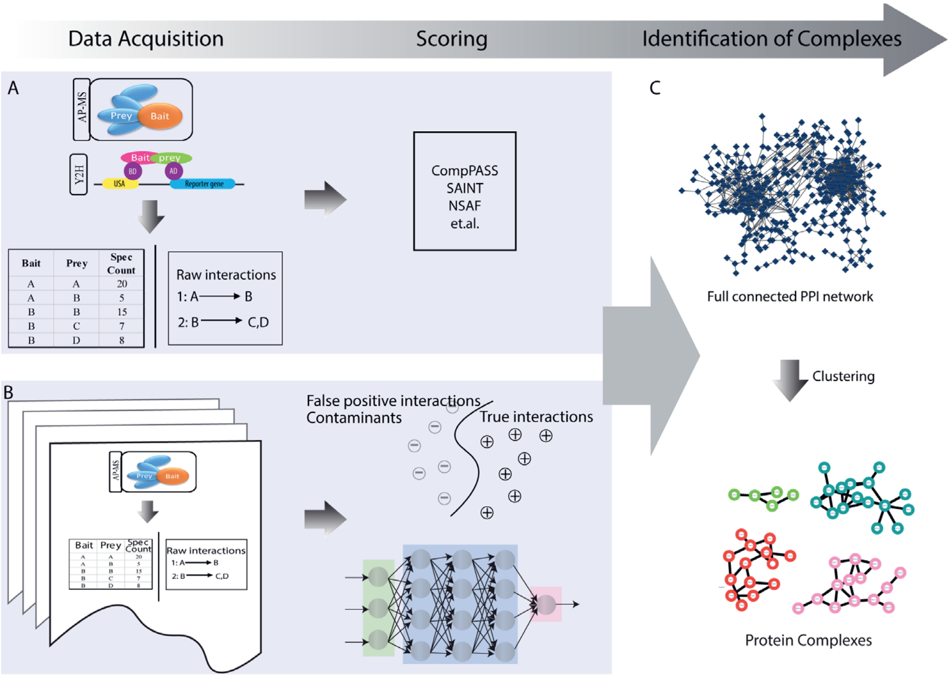


Figure 8. Workflow for identification of protein complexes. The workflow can be summarized into protein-protein interaction data acquisition, PPI scoring, and identification of complexes. (A) Characterization of PPI in a single dataset, the bait proteins are used to capture the sets of prey proteins, followed by PPI scoring. (B) Characterization of PPI by integrating multiple datasets. The PPI scores are predicted by a machine learning model generated based on existing protein-protein interaction data. (C) A full connected PPI network constructed by connecting all PPI obtained from (A) or (B), followed by a clustering strategy to identify protein complexes.

In an AP-MS experiment, a protein of interest (bait protein) is fused to a protein fragment (the 'tag'). This modified or tagged protein is expressed and purified from the cell extract, while proteins that bind to the tagged protein (prey proteins) are co-purified and subsequently identified by mass spectrometry. The AP-MS technique can delineate the dynamics of interactions at almost physiological conditions, thus explaining its success for the determination of protein complex compositions and has been performed for yeast and human proteins (94, 180, 181). Therefore, AP of tagged proteins of interest and identification of the co-purified protein components by MS (AP-MS) has become a preferred method for the analysis of PPIs because the data it produces reflect more closely the actual multidirectional complexity of a network in the cell. The following section focuses on introducing the identification of protein complexes by employing the AP-MS technology.

A concern in the protein complex analysis is the identification of true and specific PPIs as opposed to nonspecifically co-purified proteins. An effective strategy to address possible contaminants or nonspecific interactors is by using quantitative interaction proteomics to assign a bait-prey pair a interacting score to identify “true” interactors. To this end, several bioinformatics computational methods have been developed, including scoring strategies on the single AP-MS dataset and integrative large-scale AP-MS datasets.

Scoring on single AP-MS dataset. The Comparative Proteomics Analysis Software Suite (CompPASS) (182) was developed to identify high confidence interacting proteins in AP-MS experiments using spectral counts. A composite interacting score is calculated based on the bait-prey spectral count, reproducibility of interactions, and the frequency of appearance of the prey. Another scoring method, the Significance Analysis of INTeractome (SAINT) (183), uses quantitative data and generates separate Poisson distributions for true and false interactions to derive the interaction probability. The final PPI score reflects the probability of the observed spectral count belonging to the true interaction distribution. In another empirical method, Sardiu et al. (184) converts the normalized spectral abundance factor (NSAF) into the posterior probability of true interaction between a bait-prey pair using simple heuristics. To improve the accuracy of the true interacting protein pairs from those of nonspecific interacting protein pairs, the integrative modeling approach is being increasingly popular for AP-MS data analysis. For instance, the CRAPome (185) project built a contaminant repository collected from public AP-MS studies and combined these with existing scoring methods such as SAINT. Such an approach overcomes the drawback that most AP-MS studies do not capture a complete background protein set.

The performance of the above introduced algorithms is determined by the nature of the AP-MS data. For instance, scoring methods are affected by the topology of the protein-protein interaction network, the number and scale of baits library, the level of bait expression, replicates and control experiments, etc. Therefore, in practice, multiple scoring algorithms strategies are suggested to be applied to a given AP-MS dataset and then assessed on a case-by-case basis in the testing stage.

Scoring on integrative AP-MS datasets by machine learning modeling. Despite large efforts have been devoted to characterize the interactome, the existing methods for PPIs are limited by the fact that protein interaction datasets are usually incomplete. Besides, prior high-throughput protein interaction assays in yeast and humans have generally tended to show limited overlap (94, 186-188), suggesting that interactions from different studies tend to be bait-dependent and tissue-bias. Therefore, Drew et al., (138) constructed a comprehensive map of protein complexes by integrating large-scale mass spectrometry experimental datasets and employing a support vector machine (SVM) classifier to predict the interaction score. In that study, over 9000 MS-based protein interaction datasets from a variety of human and animal cells and tissues were integrated into the analysis. Interestingly, the combined map revealed thousands of PPIs that were not identified by any individual mass spectrometry experiment and increased the overlap of PPIs between different experiments.

The interactome is data-intensive, which turns researchers to looking for a higher efficient computational tool. Fortunately, deep learning has demonstrated breakthrough gains over existing best-in-class machine learning algorithms in the application of even big datasets (189, 190). A neural network consists of layers (i.e., input layer, hidden layers, and output layer) of interconnected compute units (neurons) (191). Easy-to-use software packages, such as Keras (192), have brought the neural network out of the specialist’s toolkit to a broad research area. Successful application of deep learning in research domains include regulatory genomics (193), drug discovery (194), and biomedicine (195). Inspired by the high performance and accuracy of this algorithm, in **Chapter 4,** we developed a deep learning framework that incorporates multiple sources of data (including interacting and non-interacting data) to predict protein-protein interaction probabilities, ultimately generate a comprehensive map of complexes. Our deep learning technique-based classifier significantly outperformed recently published SVM prediction models.

References

1. Crick, F., Central dogma of molecular biology. Nature 1970, 227, (5258), 561-563.
2. Venter, J. C., et al., The sequence of the human genome. science 2001, 291, (5507), 1304-1351.
3. Lander, E. S., et al., Initial sequencing and analysis of the human genome. 2001.
4. Barrett, J. C., et al., Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nature genetics 2009, 41, (6), 703-707.
5. Paternoster, L., et al., Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. Nature Genetics 2015, 47, (12), 1449-+.
6. Tachibana, C., Transcriptomics today: Microarrays, RNA-seq, and more. Science 2015, 349, (6247), 544-546.
7. Maier, T.; Guell, M.; Serrano, L., Correlation of mRNA and protein in complex biological samples. Febs Letters 2009, 583, (24), 3966-3973.
8. Aebersold, R.; Mann, M., Mass spectrometry-based proteomics. Nature 2003, 422, (6928), 198-207.
9. Han, X.; Aslanian, A.; Yates III, J. R., Mass spectrometry for proteomics. Current opinion in chemical biology 2008, 12, (5), 483-490.

10. Noor, Z.; Ahn, S. B.; Baker, M. S.; Ranganathan, S.; Mohamedali, A., Mass spectrometry–based protein identification in proteomics—a review. Briefings in bioinformatics 2021, 22, (2), 1620-1638.

11. Holmes, J.; Morrell, F., Oscillographic mass spectrometric monitoring of gas chromatography. Applied Spectroscopy 1957, 11, (2), 86-87.

12. Gohlke, R. S.; McLafferty, F. W., Early gas chromatography/mass spectrometry. Journal of the American Society for Mass Spectrometry 1993, 4, (5), 367-371.

13. Sandra, K., et al., Highly efficient peptide separations in proteomics: Part 1. Unidimensional high performance liquid chromatography. Journal of Chromatography B 2008, 866, (1-2), 48-63.

14. Altelaar, A. F. M.; Munoz, J.; Heck, A. J. R., Next-generation proteomics: towards an integrative view of proteome dynamics. Nature Reviews Genetics 2013, 14, (1), 35-48.

15. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M., Electrospray ionization for mass spectrometry of large biomolecules. Science 1989, 246, (4926), 64-71.

16. Karas, M.; Hillenkamp, F., Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Analytical chemistry 1988, 60, (20), 2299-2301.

17. Liu, Z.; Schey, K. L., Fragmentation of multiply-charged intact protein ions using MALDI TOF-TOF mass spectrometry. Journal of the American Society for Mass Spectrometry 2008, 19, (2), 231-238.

18. Wells, J. M.; McLuckey, S. A., Collision-induced dissociation (CID) of peptides and proteins. Methods in enzymology 2005, 402, 148-185.

19. Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F., Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proceedings of the National Academy of Sciences 2004, 101, (26), 9528-9533.

20. Fort, K. L., et al., Implementation of Ultraviolet Photodissociation on a Benchtop Q Exactive Mass Spectrometer and Its Application to Phosphoproteomics. Anal Chem 2016, 88, (4), 2303-10.

21. Greisch, J. F., et al., Expanding the mass range for UVPD-based native top-down mass spectrometry. Chemical Science 2019, 10, (30), 7163-7171.

22. Frese, C. K., et al., Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. Analytical chemistry 2012, 84, (22), 9668-9673.

23. Scigelova, M.; Makarov, A., Orbitrap mass analyzer–overview and applications in proteomics. Proteomics 2006, 6, (S2), 16-21.

24. Schwartz, J. C.; Senko, M. W.; Syka, J. E., A two-dimensional quadrupole ion trap mass spectrometer. Journal of the American Society for Mass Spectrometry 2002, 13, (6), 659-669.

25. Ens, W.; Standing, K. G., Hybrid quadrupole/time-of-flight mass spectrometers for analysis of biomolecules. Methods in enzymology 2005, 402, 49-78.

26. Ong, S.-E.; Mann, M., Mass spectrometry–based proteomics turns quantitative. Nature chemical biology 2005, 1, (5), 252-262.

27. Gerber, S. A.; Rush, J.; Stemman, O.; Kirschner, M. W.; Gygi, S. P., Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proceedings of the National Academy of Sciences of the United States of America 2003, 100, (12), 6940-6945.

28. Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B., Quantitative mass

spectrometry in proteomics: a critical review. Analytical and Bioanalytical Chemistry 2007, 389, (4), 1017-1031.

29. Bantscheff, M.; Lemeer, S.; Savitski, M. M.; Kuster, B., Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. Analytical and Bioanalytical Chemistry 2012, 404, (4), 939-965.

30. Gygi, S. P., et al., Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nature Biotechnology 1999, 17, (10), 994-999.

31. Oda, Y.; Huang, K.; Cross, F. R.; Cowburn, D.; Chait, B. T., Accurate quantitation of protein expression and site-specific phosphorylation. Proceedings of the National Academy of Sciences of the United States of America 1999, 96, (12), 6591-6596.

32. Gouw, J. W.; Krijgsveld, J.; Heck, A. J. R., Quantitative Proteomics by Metabolic Labeling of Model Organisms. Molecular & Cellular Proteomics 2010, 9, (1), 11-24.

33. Smolka, M. B.; Zhou, H. L.; Purkayastha, S.; Aebersold, R., Optimization of the isotope-coded affinity tag-labeling procedure for quantitative proteome analysis. Analytical Biochemistry 2001, 297, (1), 25-31.

34. Reynolds, K. J.; Yao, X. D.; Fenselau, C., Proteolytic O-18 labeling for comparative proteomics: Evaluation of endoprotease Glu-C as the catalytic agent. Journal of Proteome Research 2002, 1, (1), 27-33.

35. Yao, X. D.; Freas, A.; Ramirez, J.; Demirev, P. A.; Fenselau, C., Proteolytic O-18 labeling for comparative proteomics: Model studies with two serotypes of adenovirus (vol 73, pg 2836, 2001). Analytical Chemistry 2004, 76, (9), 2675-2675.

36. Thompson, A., et al., Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Analytical chemistry 2003, 75, (8), 1895-1904.

37. Ross, P. L., et al., Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Molecular & Cellular Proteomics 2004, 3, (12), 1154-1169.

38. Ong, S. E., et al., Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Molecular & Cellular Proteomics 2002, 1, (5), 376-386.

39. Chelius, D.; Bondarenko, P. V., Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. Journal of Proteome Research 2002, 1, (4), 317-323.

40. Sinitcyn, P.; Rudolph, J. D.; Cox, J., Computational Methods for Understanding Mass Spectrometry-Based Shotgun Proteomics Data. Annual Review of Biomedical Data Science, Vol 1 2018, 1, 207-234.

41. Prianichnikov, N., et al., MaxQuant Software for Ion Mobility Enhanced Shotgun Proteomics. Molecular & Cellular Proteomics 2020, 19, (6), 1058-1069.

42. Washburn, M. P.; Wolters, D.; Yates, J. R., Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nature biotechnology 2001, 19, (3), 242-247.

43. Liu, H. B.; Sadygov, R. G.; Yates, J. R., A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Analytical Chemistry 2004, 76, (14), 4193-4201.

44. Rogers, J. C.; Bomgarden, R. D., Sample Preparation for Mass Spectrometry-Based Proteomics; from Proteomes to Peptides. Modern Proteomics - Sample Preparation, Analysis and Practical Applications 2016, 919, 43-62.

45. Olsen, J. V.; Ong, S. E.; Mann, M., Trypsin cleaves exclusively C-terminal to arginine and lysine residues. Molecular & Cellular Proteomics 2004, 3, (6), 608-614.

46. Rodriguez, J.; Gupta, N.; Smith, R. D.; Pevzner, P. A., Does trypsin cut before proline? Journal of proteome research 2008, 7, (01), 300-305.

47. Swaney, D. L.; Wenger, C. D.; Coon, J. J., Value of Using Multiple Proteases for Large-Scale Mass Spectrometry-Based Proteomics. Journal of Proteome Research 2010, 9, (3), 1323-1329.

48. Bian, Y. Y., et al., Improve the Coverage for the Analysis of Phosphoproteome of HeLa Cells by a Tandem Digestion Approach. Journal of Proteome Research 2012, 11, (5), 2828-2837.

49. Choudhary, G.; Wu, S. L.; Shieh, P.; Hancock, W. S., Multiple enzymatic digestion for enhanced sequence coverage of proteins in complex proteomic mixtures using capillary LC with ion trap MS/MS. Journal of Proteome Research 2003, 2, (1), 59-67.

50. Dephoure, N., et al., A quantitative atlas of mitotic phosphorylation. Proceedings of the National Academy of Sciences of the United States of America 2008, 105, (31), 10762-10767.

51. Taouatas, N., et al., Strong Cation Exchange-based Fractionation of Lys-N-generated Peptides Facilitates the Targeted Analysis of Post-translational Modifications. Molecular & Cellular Proteomics 2009, 8, (1), 190-200.

52. Wisniewski, J. R.; Nagaraj, N.; Zougman, A.; Gnad, F.; Mann, M., Brain Phosphoproteome Obtained by a FASP-Based Method Reveals Plasma Membrane Protein Topology. Journal of Proteome Research 2010, 9, (6), 3280-3289.

53. Delmotte, N.; Lasaosa, M.; Tholey, A.; Heinzle, E.; Huber, C. G., Two-dimensional reversed-phase x ion-pair reversed-phase HPLC: An alternative approach to high-resolution peptide separation for shotgun proteome analysis. Journal of Proteome Research 2007, 6, (11), 4363-4373.

54. Pinkse, M. W., et al., Highly robust, automated, and sensitive online TiO2-based phosphoproteomics applied to study endogenous phosphorylation in Drosophila melanogaster. Journal of proteome research 2008, 7, (2), 687-697.

55. Beausoleil, S. A., et al., Large-scale characterization of HeLa cell nuclear phosphoproteins. Proceedings of the National Academy of Sciences of the United States of America 2004, 101, (33), 12130-12135.

56. Meissner, F.; Mann, M., Quantitative shotgun proteomics: considerations for a high-quality workflow in immunology. Nature Immunology 2014, 15, (2), 112-117.

57. Karpievitch, Y. V.; Dabney, A. R.; Smith, R. D., Normalization and missing value imputation for label-free LC-MS analysis. Bmc Bioinformatics 2012, 13.

58. Tyanova, S., et al., The Perseus computational platform for comprehensive analysis of (prote)omics data. Nature Methods 2016, 13, (9), 731-740.

59. Valikangas, T.; Suomi, T.; Elo, L. L., A systematic evaluation of normalization methods in quantitative label-free proteomics. Briefings in Bioinformatics 2018, 19, (1), 1-11.

60. Bolstad, B. M.; Irizarry, R. A.; Astrand, M.; Speed, T. P., A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003, 19, (2), 185-193.

61. Wu, W.; Xing, E. P.; Myers, C.; Mian, I. S.; Bissell, M. J., Evaluation of normalization methods for cDNA microarray data by k-NN classification. Bmc Bioinformatics 2005, 6.

62. Lazar, C.; Gatto, L.; Ferro, M.; Bruley, C.; Burger, T., Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare

Imputation Strategies. Journal of Proteome Research 2016, 15, (4), 1116-1125.

63. Kim, H. J., et al., PhosR enables processing and functional analysis of phosphoproteomic data. Cell Reports 2021, 34, (8).

64. Hendrickson, E. L.; Xia, Q. W.; Wang, T. S.; Leigh, J. A.; Hackett, M., Comparison of spectral counting and metabolic stable isotope labeling for use with quantitative microbial proteomics. Analyst 2006, 131, (12), 1335-1341.

65. Kim, T. K., T test as a parametric statistic. Korean Journal of Anesthesiology 2015, 68, (6), 540-546.

66. Smyth, G. K., Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology 2004, 3, (1).

67. Bender, R.; Lange, S., Adjusting for multiple testing—when and how? Journal of clinical epidemiology 2001, 54, (4), 343-349.

68. Noble, W. S., How does multiple testing correction work? Nature Biotechnology 2009, 27, (12), 1135-1137.

69. Bland, J. M.; Altman, D. G., Multiple Significance Tests - the Bonferroni Method .10. British Medical Journal 1995, 310, (6973), 170-170.

70. Hochberg, Y.; Tamhane, A. C., Multiple comparison procedures. John Wiley & Sons, Inc.: 1987.

71. Kumar, C.; Mann, M., Bioinformatics analysis of mass spectrometry-based proteomics data sets. Febs Letters 2009, 583, (11), 1703-1712.

72. Ashburner, M., et al., Gene Ontology: tool for the unification of biology. Nature Genetics 2000, 25, (1), 25-29.

73. Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005, 102, (43), 15545-50.

74. Hollander, M.; Wolfe, D. A.; Chicken, E., Nonparametric statistical methods. John Wiley & Sons: 2013.

75. Carnielli, C. M.; Winck, F. V.; Leme, A. F. P., Functional annotation and biological interpretation of proteomics data. Biochimica Et Biophysica Acta-Proteins and Proteomics 2015, 1854, (1), 46-54.

76. Zhou, Y. Y., et al., Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nature Communications 2019, 10.

77. Maere, S.; Heymans, K.; Kuiper, M., BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 2005, 21, (16), 3448-3449.

78. Huang, D. W.; Sherman, B. T.; Lempicki, R. A., Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols 2009, 4, (1), 44-57.

79. Warde-Farley, D., et al., The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Research 2010, 38, W214-W220.

80. Kanehisa, M.; Goto, S., KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 2000, 28, (1), 27-30.

81. Croft, D., et al., The Reactome pathway knowledgebase. Nucleic Acids Research 2014, 42, (D1), D472-D477.

82. Saito, R., et al., A travel guide to Cytoscape plugins. Nature Methods 2012, 9, (11), 1069-1076.

83. Carazzolle, M. F., et al., IIS - Integrated Interactome System: A Web-Based Platform for the Annotation, Analysis and Visualization of Protein-Metabolite-Gene-Drug Interactions by Integrating a Variety of Data Sources and Tools. Plos One 2014, 9, (6).

84. Snel, B.; Lehmann, G.; Bork, P.; Huynen, M. A., STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Research 2000, 28, (18), 3442-3444.

85. Barabasi, A. L.; Oltvai, Z. N., Network biology: Understanding the cell's functional organization. Nature Reviews Genetics 2004, 5, (2), 101-U15.

86. Ideker, T.; Nussinov, R., Network approaches and applications in biology. Plos Computational Biology 2017, 13, (10).

87. Vidal, M.; Cusick, M. E.; Barabási, A.-L., Interactome networks and human disease. Cell 2011, 144, (6), 986-998.

88. Liu, C., et al., Computational network biology: data, models, and applications. Physics Reports 2020, 846, 1-66.

89. Gao, F.; Foat, B. C.; Bussemaker, H. J., Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. BMC bioinformatics 2004, 5, (1), 1-10.

90. Yıldırım, M. A.; Goh, K.-I.; Cusick, M. E.; Barabási, A.-L.; Vidal, M., Drug—target network. Nature biotechnology 2007, 25, (10), 1119-1126.

91. Bensimon, A.; Heck, A. J.; Aebersold, R., Mass spectrometry–based proteomics and network biology. Annual review of biochemistry 2012, 81, 379-405.

92. Bin Goh, W. W.; Guo, T.; Aebersold, R.; Wong, L., Quantitative proteomics signature profiling based on network contextualization. Biology Direct 2015, 10.

93. Collins, B. C., et al., Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. Nature Methods 2013, 10, (12), 1246-+.

94. Huttlin, E. L., et al., The BioPlex Network: A Systematic Exploration of the Human Interactome. Cell 2015, 162, (2), 425-440.

95. Oughtred, R., et al., The BioGRID interaction database: 2019 update. Nucleic Acids Research 2019, 47, (D1), D529-D541.

96. Licata, L., et al., MINT, the molecular interaction database: 2012 update. Nucleic Acids Research 2012, 40, (D1), D857-D861.

97. Salwinski, L., et al., The Database of Interacting Proteins: 2004 update. Nucleic Acids Research 2004, 32, D449-D451.

98. Kerrien, S., et al., The IntAct molecular interaction database in 2012. Nucleic Acids Research 2012, 40, (D1), D841-D846.

99. Prasad, T. S. K., et al., Human Protein Reference Database-2009 update. Nucleic Acids Research 2009, 37, D767-D772.

100. Szklarczyk, D., et al., The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Research 2021, 49, (D1), D605-D612.

101. Snel, B.; Lehmann, G.; Bork, P.; Huynen, M. A., STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res 2000, 28, (18), 3442-4.

102. Szklarczyk, D., et al., STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 2015, 43, (Database issue), D447-52.

103. Szklarczyk, D., et al., STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019, 47, (D1), D607-D613.

104. Dunn, R.; Dudbridge, F.; Sanderson, C. M., The use of edge-betweenness clustering to investigate biological function in protein interaction networks. Bmc Bioinformatics 2005, 6.

105. Adamcsek, B.; Palla, G.; Farkas, I. J.; Derenyi, I.; Vicsek, T., CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics 2006, 22, (8), 1021-1023.

106. Zaki, N.; Singh, H.; Mohamed, E. A., Identifying Protein Complexes in Protein-Protein Interaction Data Using Graph Convolutional Network. Ieee Access 2021, 9, 123717-123726.

107. Enright, A. J.; Van Dongen, S.; Ouzounis, C. A., An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 2002, 30, (7), 1575-1584.

108. Macropol, K.; Can, T.; Singh, A. K., RRW: repeated random walks on genome-scale protein networks for local cluster discovery. Bmc Bioinformatics 2009, 10.

109. Nepusz, T.; Yu, H. Y.; Paccanaro, A., Detecting overlapping protein complexes in protein-protein interaction networks. Nature Methods 2012, 9, (5), 471-U81.

110. Wang, R. Q.; Wang, C. X.; Sun, L. Y.; Liu, G. X., A seed-extended algorithm for detecting protein complexes based on density and modularity with topological structure and GO annotations. Bmc Genomics 2019, 20, (1).

111. Collins, F. S.; Lander, E. S.; Rogers, J.; Waterston, R. H.; Conso, I. H. G. S., Finishing the euchromatic sequence of the human genome. Nature 2004, 431, (7011), 931-945.

112. Jensen, O. N., Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. Current Opinion in Chemical Biology 2004, 8, (1), 33-41.

113. Apweiler, R., et al., The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Research 2010, 38, D142-D148.

114. Choudhary, C., et al., Lysine Acetylation Targets Protein Complexes and Co-Regulates Major Cellular Functions. Science 2009, 325, (5942), 834-840.

115. Zielinska, D. F.; Gnad, F.; Wisniewski, J. R.; Mann, M., Precision Mapping of an In Vivo N-Glycoproteome Reveals Rigid Topological and Sequence Constraints. Cell 2010, 141, (5), 897-907.

116. Xu, G. Q.; Paige, J. S.; Jaffrey, S. R., Global analysis of lysine ubiquitination by ubiquitin remnant immunoaffinity profiling. Nature Biotechnology 2010, 28, (8), 868-U154.

117. Rontogianni, S., et al., Proteomic profiling of extracellular vesicles allows for human breast cancer subtyping. Communications Biology 2019, 2.

118. Zagorac, I., et al., In vivo phosphoproteomics reveals kinase activity profiles that predict treatment outcome in triple-negative breast cancer. Nature Communications 2018, 9.

119. Olsen, J. V., et al., Quantitative Phosphoproteomics Reveals Widespread Full Phosphorylation Site Occupancy During Mitosis. Science Signaling 2010, 3, (104).

120. Humphrey, S. J., et al., Dynamic Adipocyte Phosphoproteome Reveals that Akt Directly Regulates mTORC2. Cell Metabolism 2013, 17, (6), 1009-1020.

121. Humphrey, S. J.; Azimifar, S. B.; Mann, M., High-throughput phosphoproteomics reveals in vivo insulin signaling dynamics. Nature Biotechnology 2015, 33, (9), 990-U142.

122. Krug, K., et al., A Curated Resource for Phosphosite- specific Signature Analysis. Molecular & Cellular Proteomics 2019, 18, (3), 576-593.

123. Hornbeck, P. V., et al., PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Research 2015, 43, (D1), D512-D520.

124. Gnad, F., et al., PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biology 2007, 8, (11).

125. Dinkel, H., et al., Phospho.ELM: a database of phosphorylation sites-update 2011. Nucleic Acids Research 2011, 39, D261-D267.

126. Mischnik, M., et al., IKAP: A heuristic framework for inference of kinase activities from Phosphoproteomics data. Bioinformatics 2016, 32, (3), 424-431.

127. Casado, P., et al., Kinase-Substrate Enrichment Analysis Provides Insights into the Heterogeneity of Signaling Pathway Activation in Leukemia Cells. Science Signaling 2013, 6, (268).

128. Beekhof, R., et al., INKA, an integrative data analysis pipeline for phosphoproteomic inference of active kinases. Molecular Systems Biology 2019, 15, (4).

129. Rudolph, J. D.; de Graauw, M.; van de Water, B.; Geiger, T.; Sharan, R., Elucidation of Signaling Pathways from Large-Scale Phosphoproteomic Data Using Protein Interaction Networks. Cell Systems 2016, 3, (6), 585-+.

130. Guan, S. H.; Moran, M. F.; Ma, B., Prediction of LC-MS/MS Properties of Peptides from Sequence by Deep Learning. Molecular & Cellular Proteomics 2019, 18, (10), 2099-2107.

131. Zeng, W. F., et al., MS/MS Spectrum Prediction for Modified Peptides Using pDeep2 Trained by Transfer Learning. Analytical Chemistry 2019, 91, (15), 9724-9731.

132. Meier, F., et al., Deep learning the collisional cross sections of the peptide universe from a million experimental values. Nature Communications 2021, 12, (1).

133. Moruz, L.; Käll, L., Peptide retention time prediction. Mass spectrometry reviews 2017, 36, (5), 615-623.

134. Demichev, V.; Messner, C. B.; Vernardis, S. I.; Lilley, K. S.; Ralser, M., DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. Nature Methods 2020, 17, (1), 41-+.

135. Kim, M.; Eetemadi, A.; Tagkopoulos, I., DeepPep: Deep proteome inference from peptide profiles. Plos Computational Biology 2017, 13, (9).

136. Deeb, S. J., et al., Machine Learning-based Classification of Diffuse Large B-cell Lymphoma Patients by Their Protein Expression Profiles. Molecular & Cellular Proteomics 2015, 14, (11), 2947-2960.

137. Agranoff, D., et al., Identification of diagnostic markers for tuberculosis by proteomic fingerprinting of serum. Lancet 2006, 368, (9540), 1012-1021.

138. Drew, K., et al., Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. Molecular Systems Biology 2017, 13, (6).

139. Alzubi, J.; Nayyar, A.; Kumar, A. In Machine learning from theory to algorithms: an overview, Journal of physics: conference series, 2018; IOP Publishing: 2018; p 012012.

140. Swan, A. L.; Mobasheri, A.; Allaway, D.; Liddell, S.; Bacardit, J., Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology. Omics-a Journal of Integrative Biology 2013, 17, (12), 595-610.

141. Ghahramani, Z. In Unsupervised learning, Summer school on machine learning, 2003; Springer: 2003; pp 72-112.

142. Kotsiantis, S. B.; Zaharakis, I.; Pintelas, P., Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering

2007, 160, (1), 3-24.

143. Boser, B. E.; Guyon, I. M.; Vapnik, V. N. In A training algorithm for optimal margin classifiers, Proceedings of the fifth annual workshop on Computational learning theory, 1992; 1992; pp 144-152.

144. Cortes, C.; Vapnik, V., Support-vector networks. Machine learning 1995, 20, (3), 273-297.

145. Vapnik, V., The nature of statistical learning theory. Springer science & business media: 1999.

146. Amari, S.; Wu, S., Improving support vector machine classifiers by modifying kernel functions. Neural Networks 1999, 12, (6), 783-789.

147. Wang, H., et al., An SVM scorer for more sensitive and reliable peptide identification via tandem mass spectrometry. In Biocomputing 2006, World Scientific: 2006; pp 303-314.

148. Webb-Robertson, B. J., Support vector machines for improved peptide identification from tandem mass spectrometry database search. Methods Mol Biol 2009, 492, 453-60.

149. Guan, W., et al., Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. BMC Bioinformatics 2009, 10, 259.

150. Ben-Hur, A.; Noble, W. S., Kernel methods for predicting protein-protein interactions. Bioinformatics 2005, 21 Suppl 1, i38-46.

151. You, Z. H., et al., Detecting protein-protein interactions with a novel matrix-based protein sequence representation and support vector machines. Biomed Res Int 2015, 2015, 867516.

152. Shrestha, A.; Mahmood, A., Review of deep learning algorithms and architectures. IEEE access 2019, 7, 53040-53065.

153. Pouyanfar, S., et al., A survey on deep learning: Algorithms, techniques, and applications. ACM Computing Surveys (CSUR) 2018, 51, (5), 1-36.

154. Guo, Y. M., et al., Deep learning for visual understanding: A review. Neurocomputing 2016, 187, 27-48.

155. Ma, C. W., et al., Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning. Analytical Chemistry 2018, 90, (18), 10881-10888.

156. Tiwary, S., et al., High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. Nature Methods 2019, 16, (6), 519-+.

157. Yang, Y., et al., In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. Nature Communications 2020, 11, (1).

158. Bouwmeester, R.; Gabriels, R.; Hulstaert, N.; Martens, L.; Degroeve, S., DeepLC can predict retention times for peptides that carry as-yet unseen modifications. Nature Methods 2021, 18, (11), 1363-+.

159. Zhou, X. X., et al., pDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. Analytical Chemistry 2017, 89, (23), 12690-12697.

160. Fenoy, E.; Izarzugaza, J. M. G.; Jurtz, V.; Brunak, S.; Nielsen, M., A generic deep convolutional neural network framework for prediction of receptor-ligand interactionsNetPhosPan: application to kinase phosphorylation prediction. Bioinformatics 2019, 35, (7), 1098-1107.

161. Zhou, G. Y., et al., Mutation effect estimation on protein-protein interactions using deep contextualized representation learning. Nar Genomics and Bioinformatics 2020, 2, (2).

162. Stehman, S. V., Selecting and interpreting measures of thematic classification

accuracy. Remote Sensing of Environment 1997, 62, (1), 77-89.

163. Sokolova, M.; Japkowicz, N.; Szpakowicz, S. In Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation, Australasian joint conference on artificial intelligence, 2006; Springer: 2006; pp 1015-1021.

164. American cancer society. Cancer facts & figures 2021. In Atlanta: American Cancer Society; 2021: 2021.

165. Hodis, E., et al., A Landscape of Driver Mutations in Melanoma. Cell 2012, 150, (2), 251-263.

166. Rossi, A., et al., Drug resistance of BRAF-mutant melanoma: Review of up-to-date mechanisms of action and promising targeted agents. European Journal of Pharmacology 2019, 862.

167. Sumimoto, H.; Imabayashi, F.; Iwata, T.; Kawakami, Y., The BRAF-MAPK signaling pathway is essential for cancer-immune evasion in human melanoma cells. Journal of Experimental Medicine 2006, 203, (7), 1651-1656.

168. Lito, P., et al., Relief of Profound Feedback Inhibition of Mitogenic Signaling by RAF Inhibitors Attenuates Their Activity in BRAFV600E Melanomas. Cancer Cell 2012, 22, (5), 668-682.

169. Crosby, T.; Fish, R.; Coles, B.; Mason, M., Systemic treatments for metastatic cutaneous melanoma (vol 6, CD011123, 2018). Cochrane Database of Systematic Reviews 2018, (2).

170. Maverakis, E., et al., Metastatic Melanoma - A Review of Current and Future Treatment Options. Acta Dermato-Venereologica 2015, 95, (5), 516-524.

171. Robert, C., et al., Five-Year Outcomes with Dabrafenib plus Trametinib in Metastatic Melanoma. New England Journal of Medicine 2019, 381, (7), 626-636.

172. Sanlorenzo, M., et al., Melanoma immunotherapy. Cancer biology & therapy 2014, 15, (6), 665-674.

173. West, H. J., Immune checkpoint inhibitors. JAMA oncology 2015, 1, (1), 115-115.

174. Walhout, A. J. M., Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. Genome Research 2006, 16, (12), 1445-1454.

175. Castello, A., et al., Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. Cell 2012, 149, (6), 1393-1406.

176. Li, X. Y.; Gianoulis, T. A.; Yip, K. Y.; Gerstein, M.; Snyder, M., Extensive In Vivo Metabolite-Protein Interactions Revealed by Large-Scale Systematic Analyses. Cell 2010, 143, (4), 639-650.

177. Fields, S.; Song, O. K., A Novel Genetic System to Detect Protein Protein Interactions. Nature 1989, 340, (6230), 245-246.

178. Rigaut, G., et al., A generic protein purification method for protein complex characterization and proteome exploration. Nature Biotechnology 1999, 17, (10), 1030-1032.

179. von Mering, C., et al., Comparative assessment of large-scale data sets of protein-protein interactions. Nature 2002, 417, (6887), 399-403.

180. Gavin, A. C., et al., Proteome survey reveals modularity of the yeast cell machinery. Nature 2006, 440, (7084), 631-636.

181. Huttlin, E. L., et al., Architecture of the human interactome defines protein communities and disease networks. Nature 2017, 545, (7655), 505-+.

182. Sowa, M. E.; Bennett, E. J.; Gygi, S. P.; Harper, J. W., Defining the human

deubiquitinating enzyme interaction landscape. Cell 2009, 138, (2), 389-403.

183. Choi, H., et al., SAINT: probabilistic scoring of affinity purification-mass spectrometry data. Nature Methods 2011, 8, (1), 70-U100.

184. Sardiu, M. E., et al., Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. Proceedings of the National Academy of Sciences of the United States of America 2008, 105, (5), 1454-1459.

185. Mellacheruvu, D., et al., The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. Nature Methods 2013, 10, (8), 730-+.

186. Gandhi, T. K. B., et al., Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nature Genetics 2006, 38, (3), 285-293.

187. Hein, M. Y., et al., A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. Cell 2015, 163, (3), 712-723.

188. Wan, C. H., et al., Panorama of ancient metazoan macromolecular complexes. Nature 2015, 525, (7569), 339-+.

189. Angermueller, C.; Parnamaa, T.; Parts, L.; Stegle, O., Deep learning for computational biology. Molecular Systems Biology 2016, 12, (7).

190. Ching, T., et al., Opportunities and obstacles for deep learning in biology and medicine. Journal of the Royal Society Interface 2018, 15, (141).

191. Schmidhuber, J., Deep learning in neural networks: An overview. Neural networks 2015, 61, 85-117.

192. Keras. https://keras.io/.

193. Park, Y.; Kellis, M., Deep learning for regulatory genomics. Nature biotechnology 2015, 33, (8), 825-826.

194. Gawehn, E.; Hiss, J. A.; Schneider, G., Deep learning in drug discovery. Molecular informatics 2016, 35, (1), 3-14.

195. Mamoshina, P.; Vieira, A.; Putin, E.; Zhavoronkov, A., Applications of Deep Learning in Biomedicine. Molecular Pharmaceutics 2016, 13, (5), 1445-1454.

Chapter 2

Proteomics and Phosphoproteomics Profiling of Drug-Addicted BRAFi-Resistant Melanoma Cells

Proteomics and Phosphoproteomics Profiling of Drug-Addicted BRAFi-Resistant Melanoma Cells

Bohui Li[1,2], Xiangjun Kong[3], Harm Post[1,2], Linsey Raaijmakers[1,2], Daniel S. Peeper[3] and Maarten Altelaar*[,1,2,4]

1 Biomolecular Mass Spectrometry and Proteomics Group, Utrecht Institute for Pharmaceutical Science, Utrecht University, Utrecht, The Netherlands.

2 Netherlands Proteomics Center, Padualaan 8, 3584 CH Utrecht, The Netherlands

3 Division of Molecular Oncology and Immunology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands

4 Mass Spectrometry and Proteomics Facility, The Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands

J Proteome Res. 2021, 20(9): 4381-4392.

Abstract

Acquired resistance to MAPK inhibitors limits the clinical efficacy in melanoma treatment. We and others have recently shown that BRAF inhibitors (BRAFi)-resistant melanoma cells can develop a dependency on the therapeutic drugs to which they have acquired resistance, creating a vulnerability for these cells that can potentially be exploited in cancer treatment. In drug addicted melanoma cells, it was shown that this induction of cell death was preceded by a specific ERK2-dependent phenotype switch, however, the underlying molecular mechanisms are largely lacking. To increase the molecular understanding of this drug dependency, we applied a mass spectrometry-based proteomic approach on BRAFi-resistant BRAFMUT 451Lu cells, in which ERK1, ERK2 and JUNB were silenced separately using CRISPR–Cas9. Inactivation of ERK2 and, to a lesser extent, JUNB prevents drug addiction in these melanoma cell while, conversely, knock out of ERK1 fails to reverse this phenotype, showing a response similar to control cells. Our analysis reveals that ERK2 and JUNB share comparable proteome responses dominated by reactivation of cell division. Importantly, we find that EMT activation in drug addicted melanoma cells upon drug withdrawal is affected by silencing ERK2 but not ERK1. Moreover, transcription factor (regulator) enrichment shows that PIR acts as an effector of ERK2 and phosphoproteome analysis reveals that silencing of ERK2 but not ERK1 leads to amplification of GSK3 kinase activity. Our results depict possible mechanisms of drug addiction in melanoma, which may provide a guide for strategies in drug-resistant melanoma.

Introduction

Oncogenic mutations that cause activation of BRAF occur regularly in melanoma, with approximately 40 to 60% of cutaneous melanomas carrying mutations in BRAF (e.g. BRAF-V600E). Such mutations lead to constitutive activation of downstream signaling through the RAF/MEK/ERK mitogen-activated protein kinase (MAPK) pathway (1), making BRAF an attractive target for anti-melanoma therapy. Thus, small molecules (inhibitors) were designed to target the MAPK pathway, such as vemurafenib and dabrafenib, which are selective BRAF mutant inhibitors (2). Although these BRAF inhibitors (BRAFi) showed a lot of potential in melanoma treatment, with remarkable response rates and overall survival (3), the clinical benefit is hindered by the rapid development of acquired resistance.

Many routes to the acquisition of BRAFi resistance are described, such as BRAF allele amplification or splice variants (4), re-activation of the MAPK pathway and substitutive pathways (5, 6). The main mechanisms leading to MAPK reactivation and sustained ERK signaling involve alterations in BRAF, NRAS, MEK, and neurofibromin1 (NF1) (7, 8). The compensatory PI3K-mTOR cascade is the most commonly activated in drug resistant melanoma, via gene mutation or deletion of PTEN, or the activation of receptor tyrosine kinases (RTKs) (9-11). Interestingly, several studies have shown that discontinued drug treatment in the resistant melanoma cells causes massive cell mortality, in other words, these resistant cells become addicted to the very drugs that initially served to eliminate them (12-14).

Typically, the BRAFi-addicted melanoma cells experience a transient cell-cycle slowdown followed by cell-death upon drug withdrawal. This specific phenotype induces pERK reactivation that up-regulates p38–FRA1–JUNB–CDKN1A expression and slows down proliferation, and a robust pERK reactivation can result in DNA damage and parthanatos-related cell death (14). Moreover, ERK2, but not ERK1, was shown to be a “switch” in cancer drug addiction, since drug withdrawal induced cell death in melanoma could be reversed by genetic inactivation of ERK2 (12). Transcription factors JUNB, FRA1 and MITF, were found to play key roles in such Erk2-dependent drug addiction switch, by reprogramming the ERK2-JUNB-FRA1-MITF pathway (12).

To systematically depict the alteration of the proteome and phosphoproteome involved in drug addicted melanoma, we present a proteomic and phosphoproteomic study of BRAFi addicted melanoma cells (i.e. 451Lu cell line) in response to BRAFi withdrawal. To shed light on the role of ERK1, ERK2, and JUNB in response to drug withdrawal, we genetically silenced these genes separately by CRISPR–Cas9, in these BRAFi addicted melanoma cells, followed by systematic proteomic and phosphoproteomic profiling.

Materials and methods

Cell culture and colony formation

The BRAF inhibitor dabrafenib, the MEK inhibitor trametinib, and the ERK inhibitor SCH772984 were purchased from Selleck Chemicals. 451Lu cells were obtained from J. Villanueva (The Wistar Institute). The A375, Mel888, and A101D cells were from the Peeper laboratory cell line stock. Cells were routinely tested for mycoplasma contamination and authenticated by STR profiling (Promega). Next cells were cultured in DMEM supplemented with 9% fetal bovine serum (Sigma), plus 100 units per ml penicillin and 0.1 mg ml^{-1} streptomycin (Gibco). To generate BRAFi or BRAFi + MEKi-resistant cells, parental drug-sensitive cells were exposed to increasing concentrations of BRAFi dabrafenib (from 0.01 μM to 5 μM) or BRAFi dabrafenib + MEKi trametinib (from 0.01 μM + 0.001 μM to 0.5 μM + 0.05 μM) for 3–5 months. The drug resistance phenotype of cells was verified by colony formation assay. The generation ERK2, ERK1 or JUNB knockout pools, colony formation assay to determine cell viability in the presence or absence of BRAFi were performed as previously described (12). Resistant cells were stained with crystal violet (1% in 50% methanol) and photographed at day0 with drug and day3 without drug. The relative colony area was calculated using the plugin “ColonyArea” in ImageJ (15). Samples were collected at day0 with drug, day1 and day3 without drug for further mass spectrometry analysis.

Immunoblotting

Immunoblotting was performed as previously described (16). The BRAFi + MEKi-resistant (Mel888 BMR, A101 BMR and A375 BMR) or BRAFi-resistant (451Lu BR) cells were treated with corresponding inhibitors (BRAFi + MEKi for BMR cells, BRAFi for BR cells), with an ERK inhibitor, or drug was withdrawn, all for 24 hours. Cells were harvested and total cell lysates were prepared and submitted for immunoblotting. The antibody for Pirin was obtained from BD Bioscience.

Protein digestion

Cells were lysed, reduced and alkylated in lysis buffer (1% sodium deoxycholate (SDC), 10 mM tris(2-carboxyethyl)-phosphinehydroxchloride (TCEP), 40 mM chloroacetamide, 100 mM TRIS, pH 8.0 supplemented with protease inhibitor (cOmplete mini EDTA-free, Roche) and phosphatase inhibitor (PhosSTOP, Merck) and heated for 5 min at 95$^{\circ \mathrm{C}}$ followed by sonication with a Bioruptor Plus (Diagenode) for 15 cycles of 30 s. Samples were diluted 1:10 with 50mM ammoniumbicarbonate (AMBIC), pH 8.0 and digested with Lys-C (1:200 ratio w/w, Wako) and trypsin (1:50 ratio w/w, Merck) at 37°C, overnight. Digestion was quenched by acidification to 2% formic acid and followed by desalting using 1cc Sep-Pak C18 cartridges (Waters Corporation).

Phosphopeptide enrichment

Phosphorylated peptides were enriched using Fe(III)-NTA cartridges (Agilent technologies) in an automated fashion with the AssayMAP Bravo Platform (Agilent Technologies) as described (16). Briefly, Fe(III)-NTA cartridges were primed using 0.1% TFA in ACN and equilibrated with loading buffer (80% ACN/0.1% TFA). Samples were dissolved in loading buffer and loaded onto the cartridge, washed with loading buffer, and the phosphorylated peptides were eluted with 1% ammonia directly into 10% formic acid and dried down.

Mass spectrometry: RP-nanoLC-MS/MS

Re-suspended peptides were subjected to LC-LC MS/MS using an Agilent 1290 Infinity coupled to an Orbitrap Q Exactive HF mass spectrometer (Thermo Scientific, Bremen, Germany) using a 160 min gradient for the full proteome samples and a 100 min gradient for the phospho enriched samples. Peptides were first trapped (Dr Maisch Reprosil C18, 3 µm, 2 cm × 100 µm) before being separated on an analytical column (Agilent Poroshell EC-C18, 2.7 µm, 50 cm × 75 µm). Trapping was performed for 5 min in solvent A (0.1 M acetic acid in water) at 5 µl min-1. The LC flow during the gradient was passively split to 300 nL min−1. The 160 min gradient was as follows: 13−44% solvent B (0.1 M acetic acid in 80% ACN) in 152 min, 44−100% in 3 min and 100% for 4 min. The 100 min gradient was as follows: 9−36% solvent B (0.1 M acetic acid in 80% ACN) in 93 min, 36−100% in 3 min and 100% for 4 min. The mass spectrometer was operated in data-dependent mode. Full-scan MS spectra from m/z 375−1600 were acquired at a resolution of 60 000 at m/z 200 after accumulation to a target value of 3×10^6. Up to 15 most intense precursor ions were selected for fragmentation for the full proteome samples and up to 12 most intense precursor ions were selected for the phosphopeptide enriched samples. HCD fragmentation was performed at a normalized collision energy of 27.

Data processing

All raw MS files were searched using MaxQuant software (version 1.5.8.3) (17). MS/MS spectra were searched by Andromeda against a reviewed Homo sapiens database (20,197 entries, August, 2016) using the following parameters: trypsin digestion; maximum of two missed cleavages; cysteine carbamidomethylation as fixed modification; oxidized methionine, protein N-terminal acetylation, and serine/threonine/tyrosine phosphorylation (for the phosphoproteome data analysis only) as variable modifications. Mass tolerance was set to 4.5 and 20 ppm for the MS1 and MS2, respectively. The protein and PSM False Discovery Rate (FDR) were set to 1%. Peptide identifications by MS/MS were transferred between runs to replace missing values for quantification, with a 0.7-min window after retention time alignment.

Data analysis and statistics

All data were analyzed under the R-3.5.1 environment (18) and Microsoft Excel. Raw intensities extracted from MaxQuant were first log2 transformed and then normalized by the median of replicates for label-free quantification. We assumed that missing values were below detectability, thus sample minimums were used to replace missing values. Furtherly, t-Distributed Stochastic Neighbor Embedding (t-SNE) (19) was employed to assess the reproducibility of the experiments within biological replicates ($n = 4$). To obtain the differentially expressed (DE) proteins between ERK1-silenced cells and non-silenced cells, we fitted two linear regression models, where the first linear regression model indicates the observed differences between conditions that associate with gene silencing, which was fitted for each protein using the “lm” function in R. The second linear regression model, indicates no difference between the conditions. Next, the likelihood ratio test (LRT) was used to compare these two models to get the differential p-value for each protein using the “anova” function with the parameter “test = ‘LRT’” in R. The log2 fold change of the intensity for each protein was calculated by log2(As/An), where “As” is the mean protein abundance in the Erk1-silenced cells, and the “An” is the mean abundance in the control cells. The same analyses were performed to obtain the DE proteins between ERK2-silenced cells and non-silenced cells, and JUNB-silenced cells and non-silenced cells. Next, the DE proteins with FDR-corrected $p\text{-value} < 0.05$ and $\log2 \text{ fold change} > 1$ were further used for unsupervised clustering using Euclidean as distance measurement and then cut into four clusters based on expression patterns over time. Subsequently, DE proteins in each cluster was used for the hallmark enrichment analysis in Metascape (a gene annotation & analysis resource) online platform (20). Following, the top 10 enriched (min overlap: 3; p-value cutoff: 0.01; min enrichment: 1.5) hallmarks for each cluster were used to show the enrichment difference.

Transcription factors (regulators) enrichment

The RegEnrich R package that integrates proteome expression profiling, transcription factors (regulators) and WGCNA co-expression networks, was performed to define the key transcription factors (21, 22). Briefly, it involves three steps: 1) construction of data-driven co-expression networks using proteome profiles in control and each sgRNA cells; 2) deducing genes of interest (i.e., using differential changed protein lists to obtain a TF connected sub-network); and 3) referring importance to TFs by Fisher's exact test. An enrichment score for TFs was given by incorporating the exhibited significance of differential expression ($P_D < 0.05$) and significance of enrichment ($P_E < 0.05$). The overall scores of TFs were calculated by the following fomular.

$$score = norm(-log(p_E)) + norm(-log(p_D)), norm(x) = \frac{x - min(x)}{max(x) - min(x)}$$

Kinase activity

In our phosphoproteome profiling, we quantified 22,117 phosphosites, of which 15,124 had a localization probability over 0.75, mapping to 4,070 proteins. A single sample-based public R package InKA (Integrative Inferred Kinase Activity) that integrates kinase-centric (e.g. kinome and activation loop) and substrate-centric (e.g. PhosphoSitePlus and NetworKIN) information was applied to infer active kinases (23). Following, the top 20 activated kinases in each sample was used to show the activity difference.

Data availability

All mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (24) partner repository with the dataset identifier PXD026557.

Results and Discussion

Drug addiction phenotype switching

In this study, we sought to increase our knowledge on the dynamic response of the proteome and associated signaling networks of drug addicted melanoma cells in the presence and absence of drug, using a (phospho)proteomics approach. ERK1, ERK2, and JUNB genes were separately silenced by CRISPR–Cas9 in dabrafenib (BRAF inhibitor)-addicted 451Lu cells (Fig. 1A). These cells carry a BRAF V600E mutation as well as a MEK1 K57N-activating mutation (25). Interestingly, ceasing drug administration in drug-addicted melanoma cells triggered massive cell death in the control condition (Fig. 1B). As observed before (12), KO of ERK2 can drastically reverse this phenotype, i.e., ERK2 knockout prevents drug addiction and cell death upon drug withdrawal. Conversely, inactivation of ERK1 failed to prevent drug addiction and resulted in severe cell death upon drug withdrawal, comparable to the control situation (Fig. 1B). Loss of JUNB showed comparable results to ERK2 KO, but to a lesser extent. These results suggest that loss of ERK2 or JUNB in BRAFi addicted melanoma cells could prevent the drug addiction phenotype.

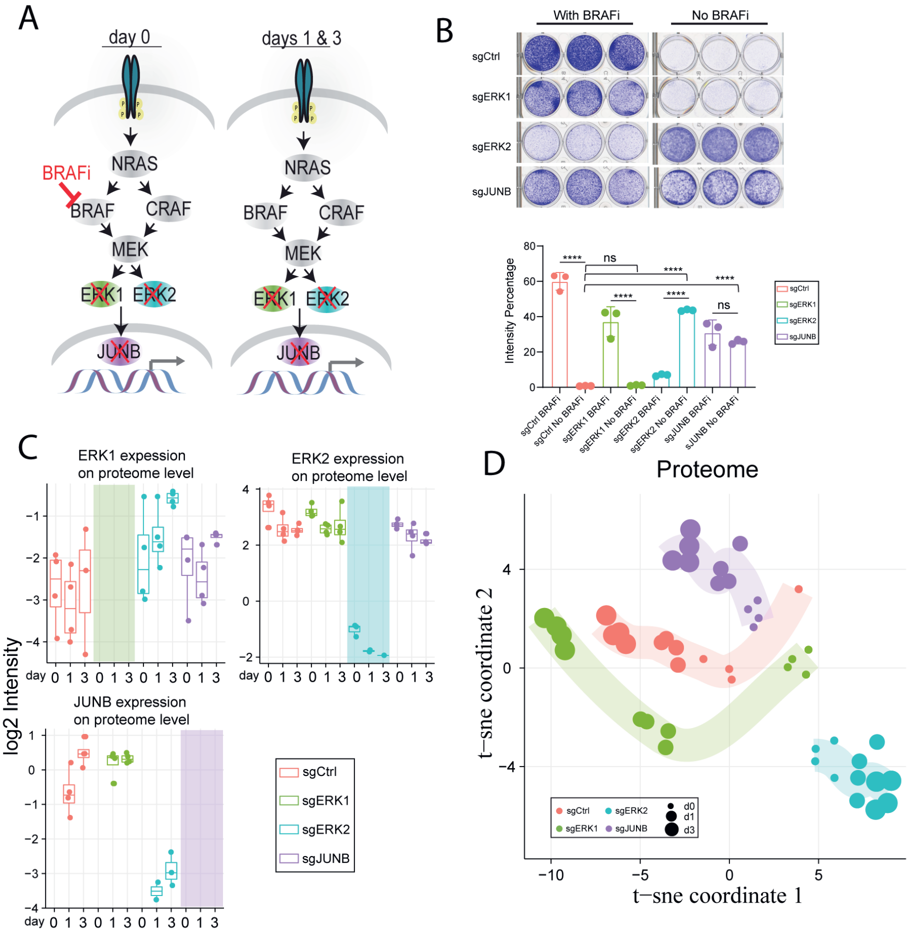


Figure 1. Phenotype and proteomic trajectory. A, Separate knockout of ERK1, ERK2, and JUNB in BRAF inhibitor (dabrafenib) addicted 451Lu cells, with drug on (day 0) and drug off (days 1 & 3). B top, control and 3 CRISPR–Cas9 knockout cells were cultured with or without BRAF inhibitor dabrafenib, followed by staining for the live cells. B bottom, the relative colony area was calculated using ImageJ. Statistical differences were analyzed by one-way ANOVA with Tukey post-hoc testing (****, $P < 0.0001$, error bars denote ±SD). C, Box plot illustrating protein abundance of each knockout gene. The points are not shown when the abundance below detectability. D t-SNE plot, based on proteome profiling, shows the trajectory of how different cells proceeded upon drug withdrawal.

To gain insight into resistance mechanisms to BRAFi addicted melanoma, we employed a label-free quantitative (phospho)proteomics approach. Briefly, cells were collected at day0 (with drug), day1 and day3 (without drug), lysed and subsequently the protein extracts were in-solution digested by LysC/trypsin and analyzed by liquid chromatography-tandem mass spectrometry (nanoLC-MS/MS) on a high-resolution mass spectrometer (Q-Exactive HF). The quantified protein abundances show depletion of ERK1, JUNB and ERK2, compared to control (Fig. 1C). Interestingly, sgCtrl, sgERK1 and sgERK2 cells show JUNB activation upon drug withdrawal, which is consistent with previous results (12).

In total, 5,720 proteins with at least two unique peptides were quantified (Table S1A), on which we performed t-Distributed Stochastic Neighbor Embedding (t-SNE) analysis. Here, we found that cells within biological replicates cluster tightly together, while cells that carry different gene KOs followed distinct trajectories over time (Fig. 1D). Of note, the cells in which ERK2 was depleted followed an opposite trajectory to the other cells over time. These results together suggest that ERK2-knockout considerably influences the proteome profiles of BRAFi-addicted melanoma cells.

Proteome response in drug addicted cells upon drug withdrawal

To compare proteome changes between sgERK2 and sgCtrl upon drug exposure and withdrawal, 2,502 significantly changed proteins between sgERK2 and sgCtrl were identified using ANOVA with 5% FDR and > 2-fold change. Next, an unsupervised hierarchical clustering method was applied to identify protein clusters with similar expression trends (Fig. 2A). A total of four expression patterns were identified, with C2 (cluster 2) and C3 (cluster 3) displaying upregulated and C1 (cluster 1) and C4 (cluster 4) downregulated patterns (Fig. 2A and 2B) upon drug removal. Pearson correlation using 2,502 differential expressed proteins revealed samples upon drug withdrawal showed higher correlation (Fig. S1A).

Clusters C2 and C3 contain upregulated proteins, where C2 contains 726 proteins that increase in expression immediately after drug removal and are stable afterward, C3 contains 512 proteins, showing proteins that continuously increase in expression upon drug withdrawal. Proteins in these clusters show strongest enrichment related to mTORC1 signaling, cytokine (IL-2 and TNF-alpha) signaling, cell division and DNA repair processes (Fig. 2D). Together, these results highlight that drug withdrawal in melanoma cells with inactivated ERK2 leads to reactivation of cell division, dominating the molecular processes observed. Among the downregulated clusters (Fig. 2B), C1 showed continuous down regulation upon drug withdrawal and C4 includes proteins downregulated at day 1 and remaining stable afterwards. Hallmark analysis showed oxidative phosphorylation as the most predominant pathways in both clusters (Fig. 2D).

To understand the different molecular mechanisms involved between sgERK2 and sgERK1 cells, we next analyzed significantly changing proteins of ERK1 depleted cells upon drug withdrawal compared to sgCtrl. This analysis identified 4 clusters with similar expression patterns, with more than half of the proteins showing down-regulation after drug withdrawal (Fig. 2A right panel and 2C). Pearson correlation with 1465 differential proteins showed higher correlation within bio-replicates, where the correlation between biological replicates generally exceeded 0.9 (Fig. S1B). Upregulated proteins upon drug removal in C1 and C4 enriched in EMT, cytokine and mTORC1 signaling (Fig. 2E). Cluster C2 and C3, showing down-regulation of proteins, significantly enriched in cell cycle (E2F targets and G2M checkpoint), followed by energy metabolism and apoptosis (Fig. 2E).

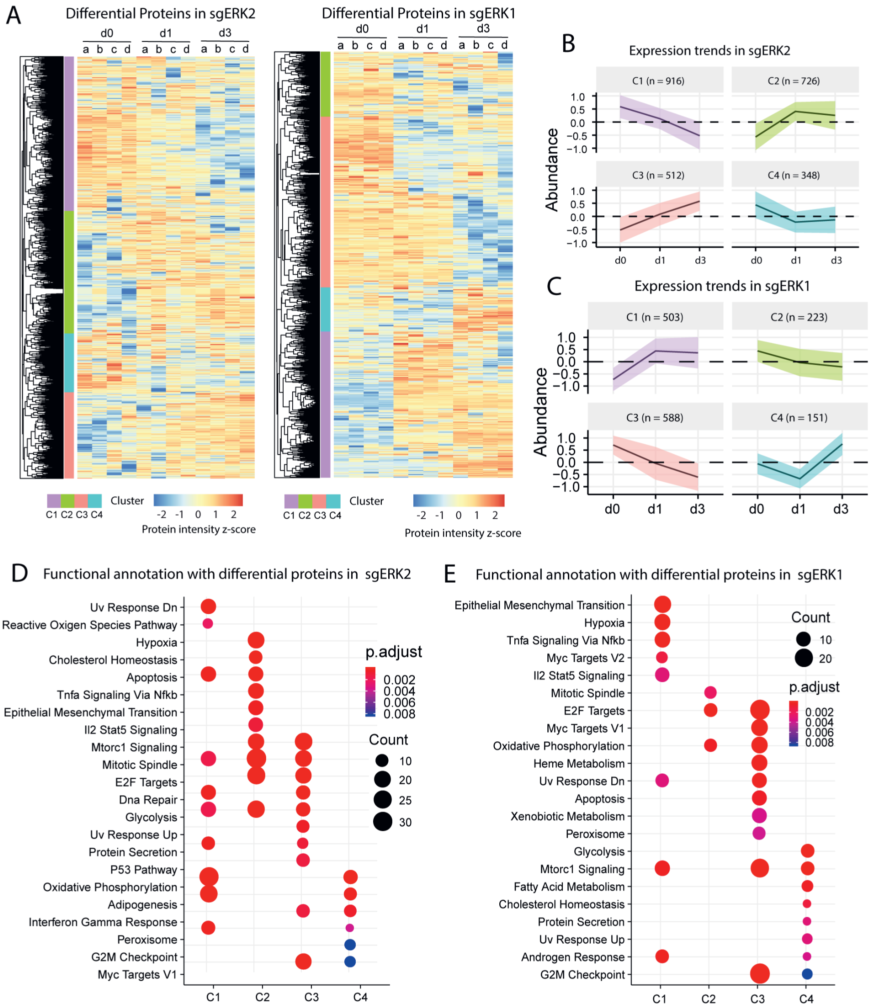


Figure 2. Proteome response in drug addicted cells upon drug withdrawal. A, Heatmap of differential expression proteins in ERK2 knockout (left) and ERK1 knockout (right), constructed using unsupervised hierarchical clustering upon drug on (day0 a, b, c, and d) and drug off (day 1 & day3 a, b, c, and d), which shows four distinct expression patterns. B & C, Average abundance trends for clusters in sgERK2 and sgERK1, shading denotes ±1 SD. D & E, Hallmark enrichment according to proteins in the corresponding cluster.

EMT is characterized by loss of typical epithelial histologic features and gain mesenchymal characteristics (26). These changes enhance cell migratory capacity and increase invasiveness, which enable the transition from melanoma in situ to aggressive, invasive melanoma (27). Our previous work showed that cells lacking ERK2, but not ERK1, failed to undergo the EMT-like changes following drug withdrawal. Consistently, we found here that silencing of ERK2 maintains EMT-related proteins at a low expression level upon drug withdrawal (Fig. 3A).

However, an opposite phenomenon was observed in sgCtrl, sgERK1 and sgJUNB cells, i.e., the EMT-related proteins were upregulated upon drug removal (Fig. 3A). These include important EMT proteins, such as fibronectin (FN1), integrin beta-1 (ITGB1), and integrin alpha-5 (ITGA5) (Fig. 3B) (28). Fibronectin is an established marker for (EMT), and has been linked to promote cancer growth, including in melanoma (29). ITGB1 was reported to enhance EMT via FAK-AKT signaling pathway (30), and an increased expression of ITGB1 has been associated with breast tumor progression (31). Further, ITGA5 has been shown to induce EMT transition and invasion in human cancer cells, after being cooperatively upregulated by twist1 and AP-1 (32). Taking these data together shows that silencing of ERK2 results in no change or slight downregulation of proteins involved in EMT, yet ERK1 KO shows a strong increase in EMT related proteins upon drug withdrawal, similar to the control and JUNB KO condition but slightly enhanced. These results suggest that EMT activation in drug addicted melanoma cells upon drug withdrawal is affected by silencing ERK2 but not ERK1 or JUNB.

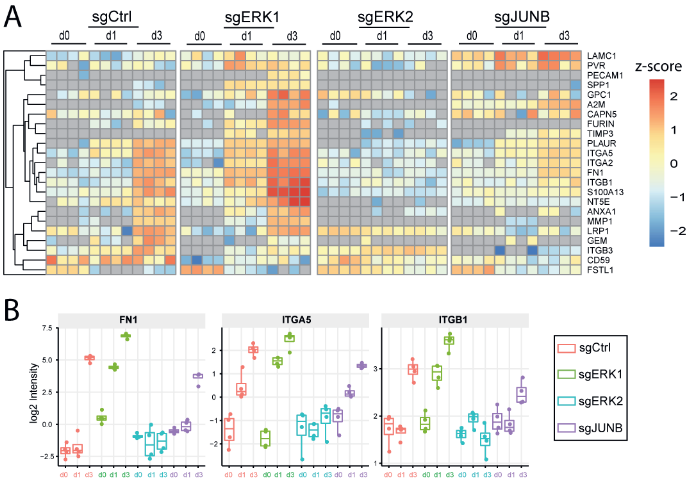


Figure 3. A, Proteins enriched in epithelial mesenchymal transition (EMT), missing values are represented in grey. B, Box plots presenting three important EMT-related proteins FN1, ITGA5, and ITGB1 in control (red), sgERK1 (green), sgERK2 (blue) and sgJUNB (purple) cells.

Depletion of JUNB shows similar behavior as sgERK2

Consistent with the t-SNE plot in figure 1D, unsupervised hierarchical clustering of the Pearson correlations (1673 differentially expressed proteins) showed sgJUNB samples clustered together over time (i.e. d0, d1 and d3), where the correlation between biological replicates generally exceeded 0.9 (Fig. S2A). More than half (937/1673) of these proteins were up-regulated upon drug withdrawal (Fig. S2B and S2C). Among the up-regulated clusters, C3 shows continuously up-regulation after drug withdrawal, while C4 shows first down-regulation of expression at day 1 and then amplification at day 3 (Fig. S2B). Biological enrichment reveals proteins in C3 to be involved in the regulation of cytokine (interferon-alpha and TNF-alpha) signaling, cell cycle (E2F targets and mitotic spindle) pathways and mTORC1 signaling. Proteins in C4 are involved in cell division and interferon-gamma signaling (Fig. S2D). Consistent with ERK2 knockout cells, the continuously downregulated proteins in cluster C2 regulate oxidative phosphorylation, apoptosis and adipogenesis (Fig. S2C and S2D).

Loss of JUNB in drug addicted melanoma cells results in largely the same proteome response upon drug withdrawal as in the ERK2 depleted cells, showing amplification of the cell cycle and activation of the mTORC1 and cytokine pathway upon drug withdrawal, meanwhile, suppressing energy metabolism. As shown in Fig. S2E, silencing of ERK2 shows more unique differential expressed proteins, and this number increases upon drug withdrawal. Moreover, more common proteins between sgERK2 and sgJUNB cells were detected, which is more obvious over time upon drug withdrawal. In total, 528 differential proteins were in common between sgERK2 and sgJUNB upon drug withdrawal day 3, showing enrichment in, amongst others, cell cycle signaling (E2F targets and mitotic spindle), oxidative phosphorylation (Fig. S2F), suggesting a common phenotype where cells survive after drug withdrawal in combination with depletion of JUNB and ERK2 in melanoma cell.

WGCNA module-based transcription factor (regulators) enrichment

The RAS-RAF-MEK-ERK signaling pathway, directly and indirectly, interacts with transcription factors and their regulators, thereby controlling cell survival and proliferation (33). To systematically depict the underlying inter-connectivity (e.g., interactions between proteins and transcription factors) in BRAFi-resistant melanoma cells, a module based weighted protein (gene) co-expression network analysis (WGCNA) was carried out to explore correlations between differentially expressed proteins and regulators (21). Next, a one-tailed hypergeometric test was used to determine the importance of highlighted transcription factors and regulators (34-36). An enrichment score for TFs was given by incorporating the exhibited significant differential expression ($p < 0.05$) and significant enrichment ($p < 0.05$) (see Methods).

Figure 4A shows the top-scored transcription factors and regulators in ERK2 KO at d0 (with drug present). NR2F2, also known as COUP-TFII (chicken ovalbumin upstream promoter-transcription factor II), which acts as a major angiogenesis regulator in the tumor microenvironment by regulating the transcription of Angiopoietin-1 (37), and promotes metastasis by loss of miR-101 and miR-27a, thereby inducing FOXM1 and CENPF in prostate cancer (38). Furthermore, the top enriched regulators include tumor suppressor genes PML, APC and GATA4 (39-41), which are upregulated when drug was present. For instance, the tumor suppressor PML is involved in regulating the p53 response to oncogenic signals and overexpression of PML induces senescence in a p53-dependent manner (42). Knocking out PML impaired the p53-regulatory pathway for apoptosis and the induction of p53 target genes such as Bax and p21 upon γ-irradiation (43).

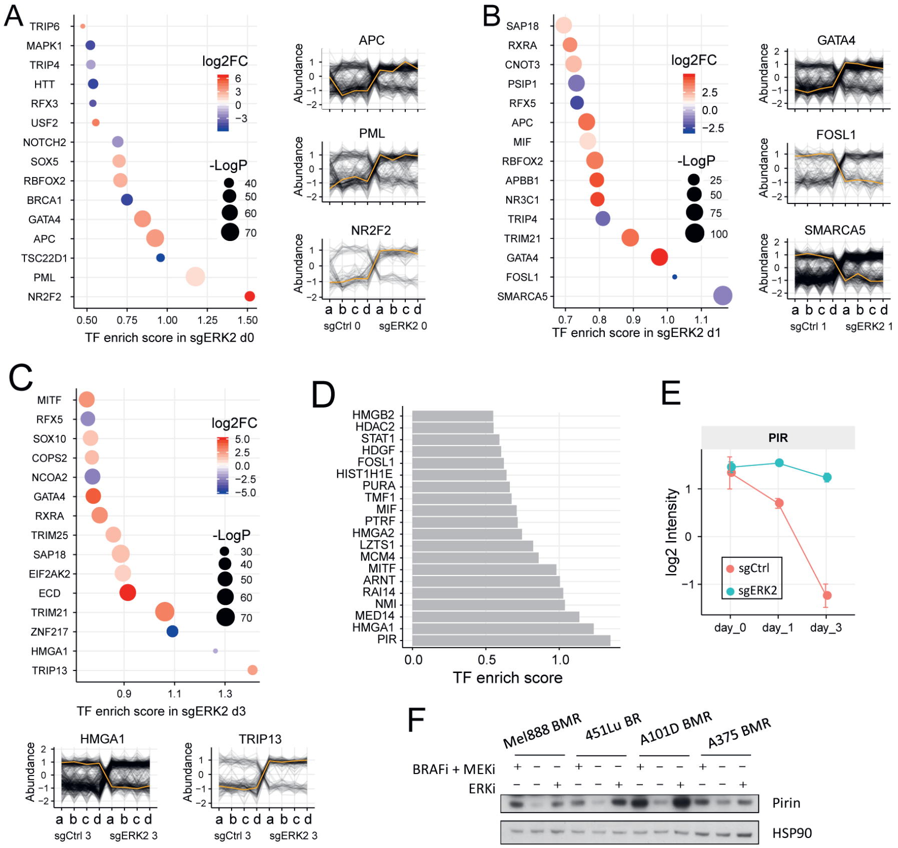


Figure 4. Transcription factors (regulators) enrichment. A-C, (left) Top 15 significant transcription regulators enriched in sgERK2 on day0, day1, and day3 separately, shades by log2 fold change of expression intensity, and TF enrich –log P-value donates dots size; (right) Abundance of 3 TFs (highlighted in yellow line) with co-expression proteins (black lines). D, Top significant TFs with pooled of differential changed proteins in sgERK2. E, PIR (encode of Pirin) expression intensity in sgCtrl (red) and sgERK2 (blue), error bar presents ±1 SE. F, Four drug-addicted melanoma cell line panels (BR, for cells resistant to BRAFi, and BMR for cells resistant to the combined BRAFi and MEKi) were cultured 1) with or without MAPK inhibitors, 2) with or without ERK inhibitor, and then immunoblotted.

Next, we looked at the top fifteen transcription factors showing high correlation with the differentially expressed proteins at d1 and d3 after drug removal (Fig. 4B and 4C). Consistent with the results from Kong et al. (12), the FOSL1 (encodes FRA1) induction after drug withdrawal was diminished in sgERK2 cells. Previous research showed that FOSL1 can act oncogenic to transform melanocytes, enabling subcutaneous tumor growth, through downregulation of MITF in a HMGA1-dependent manner (42). Interestingly, we observed a strong decrease of MITF abundance in sgCtrl and sgERK1 cells upon drug withdrawal, whereas a high expression level was maintained in sgERK2 and sgJUNB cells upon drug withdrawal (Fig. S3). Conversely, HMGA1 (one of the co-expression proteins for MITF in our prediction result) showed an opposite trend compared to MITF, i.e., upregulation in sgCtrl and sgERK1 cells, while maintaining a low abundance level in sgERK2 and sgJUNB cells.

Furthermore, the transcription factor MITF was reported to be co-regulated by SOX10, both of these two TFs show enrichment at d3. It has been shown that downregulation of SOX10 results in a simultaneous reduction of MITF and increased SOX9 expression (44), and a low level of MITF was detected in ‘invasive’ type of melanoma (45). The most significantly enriched transcription factors at d1 and d3 were SMARCA5 (also known as SNF2H) (Fig. 4B) and TRIP13 (a key mitosis regulator) (Fig. 4C), respectively. SMARCA5, involved in preventing genomic instability (46) and interacts with the miR-99 family to regulate the DNA damage response (47), shows lower expression in sgERK2 upon drug withdrawal at d1. TRIP13 is reported to promote colorectal cancer progression by modulating EMT related protein YWHAZ (14-3-3 protein zeta/delta) (48).

Compared to sgERK2, sgERK1 alters a completely different set of transcription factors and regulators (Fig. S4A and 4B), mainly showing down regulation at d1-d3. The observed down regulation of many of these factors can be explained by the decreased cell viability at drug withdrawal. This can be illustrated by the down regulation of MCM proteins (MCM2, MCM3 and MCM5), which are involved in governing DNA replication and the cell cycle process (49). At d2, the strongest upregulation is observed for NAB2, a suppressor of the inducible zinc finger transcription factors EGR1 and EGR2, which regulate the expression of genes involved in differentiation, growth, and response to extracellular signals (50).

Interestingly, silencing of JUNB and ERK2 display a more common profile of significant transcription factors (regulators) (Fig. S4C and Fig.4). For example, the transcriptional repressor TSC22D1 (TSC22 domain family protein 1) (51), which acts as a negative feedback regulator of Ras/Raf signaling (51), shows reduced expression in in the presence of drug in both sgERK2 and sgJUNB. In addition, changed expression of the chromatin remodeling factor-HMGA1 was observed upon drug withdrawal. HMGA1 has been reported to function in melanocyte progression to melanoma and involved in EMT. Maurus et.al showed that siRNA-mediated reduction of HMGA1 partially prevented the FOSL1-mediated reduction of MITF on RNA and protein level (52). Pegoraro et.al demonstrated that knock-down of HMGA1 induces the mesenchymal to epithelial transition and dramatically decreases stemness and self-renewal in basal-like breast cancer (53).

Next, we pooled our data over all three days, which results in PIR, which encodes pirin, a transcriptional co-regulator of nuclear factor (NF)κB, to be most significantly enriched in sgERK2 (Fig. 4D). Pirin has been described to inhibit melanocytic cell senescence (54) and regulate migration of melanoma cells (55). Previous research revealed that pirin can bind to Bcl3 that interacts with NFκB, thereby enhancing cell survival, proliferation and tumor malignancy (56-58). Pirin expresses stably upon drug withdrawal in sgERK2 cells, while it is strongly downregulated in sgCTRL cells (Fig. 4E). Moreover, Pirin expression is high in four different melanoma cells (e.g. Mel888BMR, 451LuBR, A101DBMR and A375BMR) treated with either a combination of BRAFi and MEKi or ERKi (Fig. 4F). However, pharmacological inhibition of PIR failed to prevent lethality caused by drug removal in these BRAFi-resistant cells. Next, we wondered whether restoring expression of pirin in control cells could reverse the observed cell death upon drug withdrawal. For this we overexpressed PIR in BRAFi-resistant 451Lu as well as BRAFi + MEKi-resistant Mel888 cells (Fig. S5). However, we did not observe such a rescue effect, suggesting PIR acts downstream of ERK2 (being a biomarker rather than a functional mediator) and ERK2 KO prevents its degradation.

Phosphoproteome profiling inferred kinases activity

To further investigate the cellular signaling dynamics in these different melanoma cells we performed phosphoproteome profiling, quantifying 22,117 phosphosites, 15,124 with a localization probability over 0.75, which mapped to 4,070 proteins (Table S2A). Similar to our proteome profile, the tSNE plot shows that the different genetic clones followed distinct trajectories over time, while biological replicates group tightly together (Fig. 5A). In our data we observed the activating close-proximity phosphosites T202/Y204 on ERK1 and T185/Y187 on ERK2 (59), showing significant upregulation upon drug withdrawal in all conditions except in the conditions where either of these genes were silenced (Fig. S7).

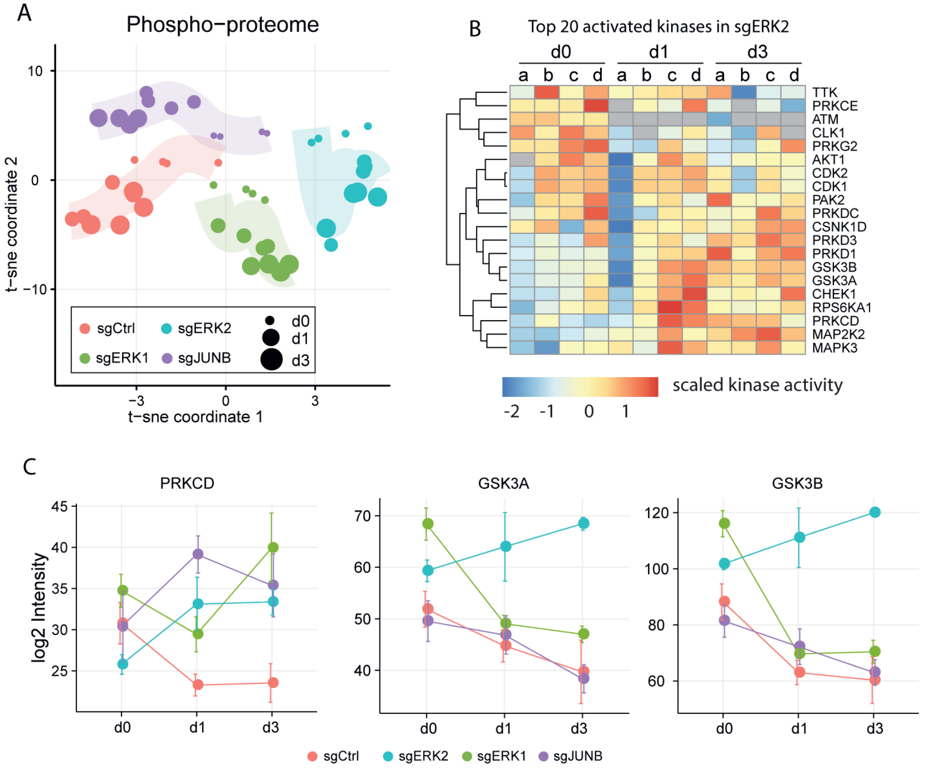


Figure 5. Phospho-proteome inferred kinases activity. A, phosphoproteomic profiling based t-SNE plot, shows the trajectory of how different cells proceeded upon drug removal. B, Top 20 activated kinases in sgERK2. C, Kinase activities in control (red), sgERK2 (blue), sgERK1 (green), and sgJUNB (purple), error bar presents ±1 SE.

To further unveil the role of the observed phosphorylation dynamics in these cells, an R pipeline ‘InKA’, which integrates kinome, activation loop, phosphoSitePlus and NetworKIN evidences, was applied to infer kinase activity (23). Based on this substrate-centric kinase activity prediction model, a total of 131 kinases were predicted with a relative activity score (Table S2B). We observed that ERK1, ERK2 and CDK1 are the most activated kinases, which is more obvious over time upon drug withdrawal in control, sgERK1 and sgJUNB, indicating their dominant roles in drug resistant melanoma. To obtain a high confident kinases activity, only those top 20 scoring kinases with an average score around 50 in each condition are presented (Fig. S7A). Of the top 20 kinases, silencing of ERK2 results in the highest number of activated kinases upon drug removal, followed by silencing of JUNB. Moreover, several kinases were predicted to be activated upon drug withdrawal both in sgERK2 and sgJUNB cells (Fig. 5B and Fig. S7C), including MAPK pathway members such as MAP2K2 (MEK2), AKT1 and MAPK3 (Fig. S7B), and cell cycle kinases CDK1/2. Notably, no changes were detected in protein levels of MAP2K2 (MEK2), while in our phosphoproteome data, we observed a significant amplification in activity upon drug withdrawal. Conversely, most of those kinases show decreasing activity in sgERK1 and control cells when ceasing drug administration. These results suggest that silencing of ERK2 and JUNB share more similarities in phosphorylation profiles, activating key members in MAPK and cell cycle signaling.

An interesting observation is the kinase PRKCD which has been reported to have contradicting roles in cell survival and death (60). The kinase is enriched in all conditions, however, shows maximum enrichment on day1 of drug withdrawal in Erk2 and JunB KO cells, while its maximum is reached on day3 in Erk1 KO cells (Fig. 5C). However, the most striking difference observed in kinase activation is that of glycogen synthase kinase-3 (GSK3). We found GSK3A and GSK3B hyper-activated in ERK2 KO melanoma cells, while their activity decreased in the other three conditions upon drug withdrawal (Fig. 5C and Fig. S7C). Previous research reveals that inhibition of ERK1/2 restores GSK3B activity and protein synthesis levels in a tuberous sclerosis model (61). Our data indicates that control of GSK3B in drug resistant melanoma cells is indeed controlled by ERK2 but not ERK1. It has been shown that the level of glycogen synthase kinase-3 (GSK3) in its active form is higher in tumor cells compared to normal tissue (62). Furthermore, GSK3 is active downstream both PI3K and Wnt pathways, and converges MAPK signaling to control MITF nuclear export (63) and promote cell survival and growth in human melanoma cells (64).

Conclusion

Here, we present a proteomics and phosphoproteomics profiling of BRAFi addicted melanoma cells (i.e. 451Lu cell line carries BRAFMUT) in response to BRAFi withdrawal. Silencing of ERK2 and JUNB could prevent drug addiction and reverse drug withdrawal induced cell death, in contrast, inactivation of ERK1 failed to do so. Depletion of ERK2 and JUNB share more similar proteome profiles upon drug withdrawal, while the proteome response in ERK1 depleted cells resembles that in control cells. Notably, we find a strong increase in

EMT related proteins upon drug withdrawal in both control and ERK1 depleted cells, which is abrogated by silencing of ERK2. These results suggest that EMT activation in drug addicted melanoma cells upon drug withdrawal is affected by silencing ERK2 but not ERK1. Moreover, we identify PIR as an effector of ERK2 and show amplification of GSK3 kinase activity upon silencing of ERK2 but not ERK1. Our results depict how ERK1, ERK2 and JUNB influence the proteome response of drug addicted melanoma cells upon drug withdrawal, which may help future strategies fighting drug-resistance.

Acknowledgement

This research was part of the Netherlands X-omics Initiative and partially funded by NWO (Project 184.034.019) and the China Scholarship Council (CSC) no. 201606300049.

References

1. Millington, G. W., Mutations of the BRAF gene in human cancer, by Davies et al. (Nature 2002; 417: 949-54). Clin Exp Dermatol 2013, 38, (2), 222-3.

2. Domingues, B.; Lopes, J. M.; Soares, P.; Populo, H., Melanoma treatment in review. Immunotargets Ther 2018, 7, 35-49.

3. Robert, C., et al., Five-Year Outcomes with Dabrafenib plus Trametinib in Metastatic Melanoma. N Engl J Med 2019, 381, (7), 626-636.

4. Vido, M. J.; Le, K.; Hartsough, E. J.; Aplin, A. E., BRAF Splice Variant Resistance to RAF Inhibitor Requires Enhanced MEK Association. Cell Rep 2018, 25, (6), 1501-1510 e3.

5. Kozar, I.; Margue, C.; Rothengatter, S.; Haan, C.; Kreis, S., Many ways to resistance: How melanoma cells evade targeted therapies. Biochimica Et Biophysica Acta-Reviews on Cancer 2019, 1871, (2), 313-322.

6. Shi, H. B., et al., Acquired Resistance and Clonal Evolution in Melanoma during BRAF Inhibitor Therapy. Cancer Discovery 2014, 4, (1), 80-93.

7. Dietrich, P.; Kuphal, S.; Spruss, T.; Hellerbrand, C.; Bosserhoff, A. K., Wild-type KRAS is a novel therapeutic target for melanoma contributing to primary and acquired resistance to BRAF inhibition. Oncogene 2018, 37, (7), 897-911.

8. Stark, M. S., et al., miR-514a regulates the tumour suppressor NF1 and modulates BRAFi sensitivity in melanoma. Oncotarget 2015, 6, (19), 17753-17763.

9. Zuo, Q., et al., AXL/AKT axis mediated-resistance to BRAF inhibitor depends on PTEN status in melanoma. Oncogene 2018, 37, (24), 3275-3289.

10. Irvine, M.; Stewart, A.; Pedersen, B.; Boyd, S.; Kefford, R.; Rizos, H., Oncogenic PI3K/AKT promotes the step-wise evolution of combination BRAF/MEK inhibitor resistance in melanoma. Oncogenesis 2018, 7, (9), 72.

11. Miller, M. A., et al., Reduced Proteolytic Shedding of Receptor Tyrosine Kinases Is a Post-Translational Mechanism of Kinase Inhibitor Resistance. Cancer Discovery 2016, 6, (4), 382-399.

12. Kong, X. J., et al., Cancer drug addiction is relayed by an ERK2-dependent phenotype switch. Nature 2017, 550, (7675), 270-+.

13. Moriceau, G., et al., Tunable-combinatorial mechanisms of acquired resistance limit

the efficacy of BRAF/MEK cotargeting but result in melanoma drug addiction. Cancer Cell 2015, 27, (2), 240-56.

14. Hong, A., et al., Exploiting Drug Addiction Mechanisms to Select against MAPKi-Resistant Melanoma. Cancer Discov 2018, 8, (1), 74-93.

15. Guzman, C.; Bagga, M.; Kaur, A.; Westermarck, J.; Abankwa, D., ColonyArea: an ImageJ plugin to automatically quantify colony formation in clonogenic assays. PLoS One 2014, 9, (3), e92444.

16. Post, H., et al., Robust, Sensitive, and Automated Phosphopeptide Enrichment Optimized for Low Sample Amounts Applied to Primary Hippocampal Neurons. J Proteome Res 2017, 16, (2), 728-737.

17. Cox, J.; Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 2008, 26, (12), 1367-72.

18. R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 2017. URL https://www.R-project.org/.

19. van der Maaten, L.; Hinton, G., Visualizing Data using t-SNE. Journal of Machine Learning Research 2008, 9, 2579-2605.

20. Zhou, Y., et al., Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nature communications 2019, 10, (1), 1-10.

21. Tao, W.; Radstake, T. R.; Pandit, A., RegEnrich: An R package for gene regulator enrichment analysis reveals key role of ETS transcription factor family in interferon signaling. bioRxiv 2021.

22. Langfelder, P.; Horvath, S., WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008, 9, 559.

23. Beekhof, R., et al., INKA, an integrative data analysis pipeline for phosphoproteomic inference of active kinases. Mol Syst Biol 2019, 15, (5), e8981.

24. Perez-Riverol Y,; Csordas A,; Bai J,; Bernal-Llinares M,; Hewapathirana S,; Kundu DJ,; Inuganti A,; Griss J,; Mayer G,; Eisenacher M,; Pérez E,; Uszkoreit J,; Pfeuffer J,; Sachsenberg T,; Yilmaz S,; Tiwary S,; Cox J,; Audain E,; Walzer M,; Jarnuczak AF,; Ternent T,; Brazma A, Vizcaíno JA, The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res 2019, 47(D1):D442-D450

25. Kemper, K., et al., BRAF(V600E) Kinase Domain Duplication Identified in Therapy-Refractory Melanoma Patient-Derived Xenografts. Cell Rep 2016, 16, (1), 263-277.

26. Kalluri, R.; Weinberg, R. A., The basics of epithelial-mesenchymal transition. J Clin Invest 2009, 119, (6), 1420-8.

27. Lin, K.; Baritaki, S.; Militello, L.; Malaponte, G.; Bevelacqua, Y.; Bonavida, B., The Role of B-RAF Mutations in Melanoma and the Induction of EMT via Dysregulation of the NF-kappaB/Snail/RKIP/PTEN Circuit. Genes Cancer 2010, 1, (5), 409-420.

28. Zhao, M.; Kong, L.; Liu, Y.; Qu, H., dbEMT: an epithelial-mesenchymal transition associated gene resource. Sci Rep 2015, 5, 11459.

29. Park, J.; Schwarzbauer, J. E., Mammary epithelial cell interactions with fibronectin stimulate epithelial-mesenchymal transition. Oncogene 2014, 33, (13), 1649-57.

30. Ju, L.; Zhou, C., Integrin beta 1 enhances the epithelial-mesenchymal transition in association with gefitinib resistance of non-small cell lung cancer. Cancer Biomark 2013, 13, (5), 329-36.

31. Yang, J., et al., Twist induces epithelial-mesenchymal transition and cell motility in

breast cancer via ITGB1-FAK/ILK signaling axis and its associated downstream network. Int J Biochem Cell Biol 2016, 71, 62-71.

32. Nam, E. H.; Lee, Y.; Moon, B.; Lee, J. W.; Kim, S., Twist1 and AP-1 cooperatively upregulate integrin alpha5 expression to induce invasion and the epithelial-mesenchymal transition. Carcinogenesis 2015, 36, (3), 327-37.

33. Sanchez, I., et al., Role of SAPK/ERK kinase-1 in the stress-activated pathway regulating transcription factor c-Jun. Nature 1994, 372, (6508), 794-8.

34. Ramirez, R. N.; El-Ali, N. C.; Mager, M. A.; Wyman, D.; Conesa, A.; Mortazavi, A., Dynamic Gene Regulatory Networks of Human Myeloid Differentiation. Cell Syst 2017, 4, (4), 416-429 e3.

35. Goode, D. K., et al., Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation. Dev Cell 2016, 36, (5), 572-87.

36. Langfelder, P.; Luo, R.; Oldham, M. C.; Horvath, S., Is My Network Module Preserved and Reproducible? Plos Computational Biology 2011, 7, (1): e1001057.

37. Qin, J.; Chen, X. P.; Xie, X.; Tsai, M. J.; Tsai, S. Y., COUP-TFII regulates tumor growth and metastasis by modulating tumor angiogenesis. Proceedings of the National Academy of Sciences of the United States of America 2010, 107, (8), 3687-3692.

38. Lin, S.-C., et al., Dysregulation of miRNAs-COUP-TFII-FOXM1-CENPF axis contributes to the metastasis of prostate cancer. Nature communications 2016, 7, (1), 1-14.

39. Salomoni, P.; Pandolfi, P. P., The role of PML in tumor suppression. Cell 2002, 108, (2), 165-170.

40. Muthusamy, V., et al., Epigenetic silencing of novel tumor suppressors in malignant melanoma. Cancer Res 2006, 66, (23), 11187-93.

41. Kang, C., et al., The DNA damage response induces inflammation and senescence by inhibiting autophagy of GATA4. Science 2015, 349, (6255), aaa5612.

42. Guo, A., et al., The function of PML in p53-dependent apoptosis. Nat Cell Biol 2000, 2, (10), 730-6.

43. Pearson, M., et al., PML regulates p53 acetylation and premature senescence induced by oncogenic Ras. Nature 2000, 406, (6792), 207-10.

44. Shakhova, O., et al., Sox10 promotes the formation and maintenance of giant congenital naevi and melanoma. Nat Cell Biol 2012, 14, (8), 882-90.

45. Verfaillie, A., et al., Decoding the regulatory landscape of melanoma reveals TEADS as regulators of the invasive cell state. Nat Commun 2015, 6, 6683.

46. Toiber, D., et al., SIRT6 recruits SNF2H to DNA break sites, preventing genomic instability through chromatin remodeling. Mol Cell 2013, 51, (4), 454-68.

47. Mueller, A. C.; Sun, D.; Dutta, A., The miR-99 family regulates the DNA damage response through its target SNF2H. Oncogene 2013, 32, (9), 1164-72.

48. Sheng, N., et al., TRIP13 promotes tumor growth and is associated with poor prognosis in colorectal cancer. Cell Death Dis 2018, 9, (3), 402.

49. Evrin, C., et al., A double-hexameric MCM2-7 complex is loaded onto origin DNA during licensing of eukaryotic DNA replication. Proc Natl Acad Sci U S A 2009, 106, (48), 20240-5.

50. Kumbrink, J.; Kirsch, K. H.; Johnson, J. P., EGR1, EGR2, and EGR3 activate the expression of their coregulator NAB2 establishing a negative feedback loop in cells of neuroectodermal and epithelial origin. J Cell Biochem 2010, 111, (1), 207-17.

51. Kester, H. A.; Blanchetot, C.; den Hertog, J.; van der Saag, P. T.; van der Burg, B., Transforming growth factor-β-stimulated clone-22 is a member of a family of leucine zipper proteins that can homo-and heterodimerize and has transcriptional repressor activity. Journal of Biological Chemistry 1999, 274, (39), 27439-27447.

52. Maurus, K., et al., The AP-1 transcription factor FOSL1 causes melanocyte reprogramming and transformation. Oncogene 2017, 36, (36), 5110-5121.

53. Pegoraro, S., et al., HMGA1 promotes metastatic processes in basal-like breast cancer regulating EMT and stemness. Oncotarget 2013, 4, (8), 1293-308.

54. Licciulli, S., et al., Pirin inhibits cellular senescence in melanocytic cells. Am J Pathol 2011, 178, (5), 2397-406.

55. Miyazaki, I.; Simizu, S.; Okumura, H.; Takagi, S.; Osada, H., A small-molecule inhibitor shows that pirin regulates migration of melanoma cells. Nature chemical biology 2010, 6, (9), 667-673.

56. Dechend, R., et al., The Bcl-3 oncoprotein acts as a bridging factor between NF-κB/ Rel and nuclear co-regulators. Oncogene 1999, 18, (22), 3316-3323.

57. Park, S. G.; Chung, C.; Kang, H.; Kim, J.-Y.; Jung, G., Up-regulation of cyclin D1 by HBx is mediated by NF-κB2/BCL3 complex through κB site of cyclin D1 promoter. Journal of Biological Chemistry 2006, 281, (42), 31770-31777.

58. Courtois, G.; Gilmore, T. D., Mutations in the NF-kappaB signaling pathway: implications for human disease. Oncogene 2006, 25, (51), 6831-43.

59. Goetz, E. M.; Ghandi, M.; Treacy, D. J.; Wagle, N.; Garraway, L. A., ERK mutations confer resistance to mitogen-activated protein kinase pathway inhibitors. Cancer Res 2014, 74, (23), 7079-89.

60. Kronfeld, I.; Kazimirsky, G.; Lorenzo, P. S.; Garfield, S. H.; Blumberg, P. M.; Brodie, C., Phosphorylation of protein kinase Cdelta on distinct tyrosine residues regulates specific cellular functions. J Biol Chem 2000, 275, (45), 35491-8.

61. Pal, R.; Bondar, V. V.; Adamski, C. J.; Rodney, G. G.; Sardiello, M., Inhibition of ERK1/2 Restores GSK3beta Activity and Protein Synthesis Levels in a Model of Tuberous Sclerosis. Sci Rep 2017, 7, (1), 4174.

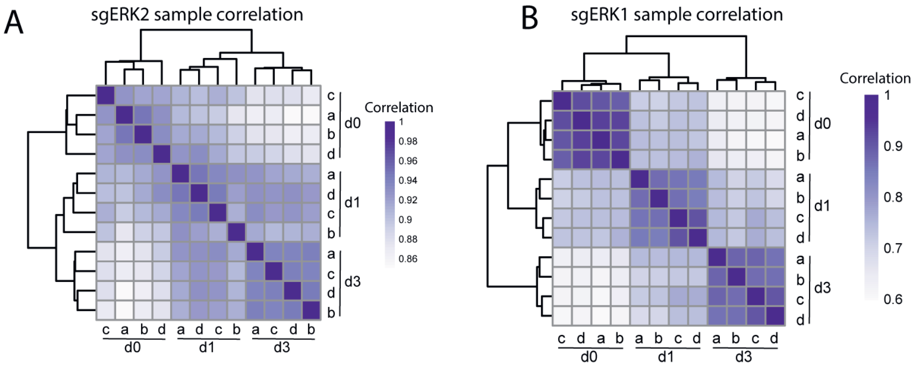
62. Shakoori, A., et al., Deregulated GSK3beta activity in colorectal cancer: its association with tumor cell survival and proliferation. Biochem Biophys Res Commun 2005, 334, (4), 1365-73.

63. Ngeow, K. C., et al., BRAF/MAPK and GSK3 signaling converges to control MITF nuclear export. Proc Natl Acad Sci U S A 2018, 115, (37), E8668-E8677.

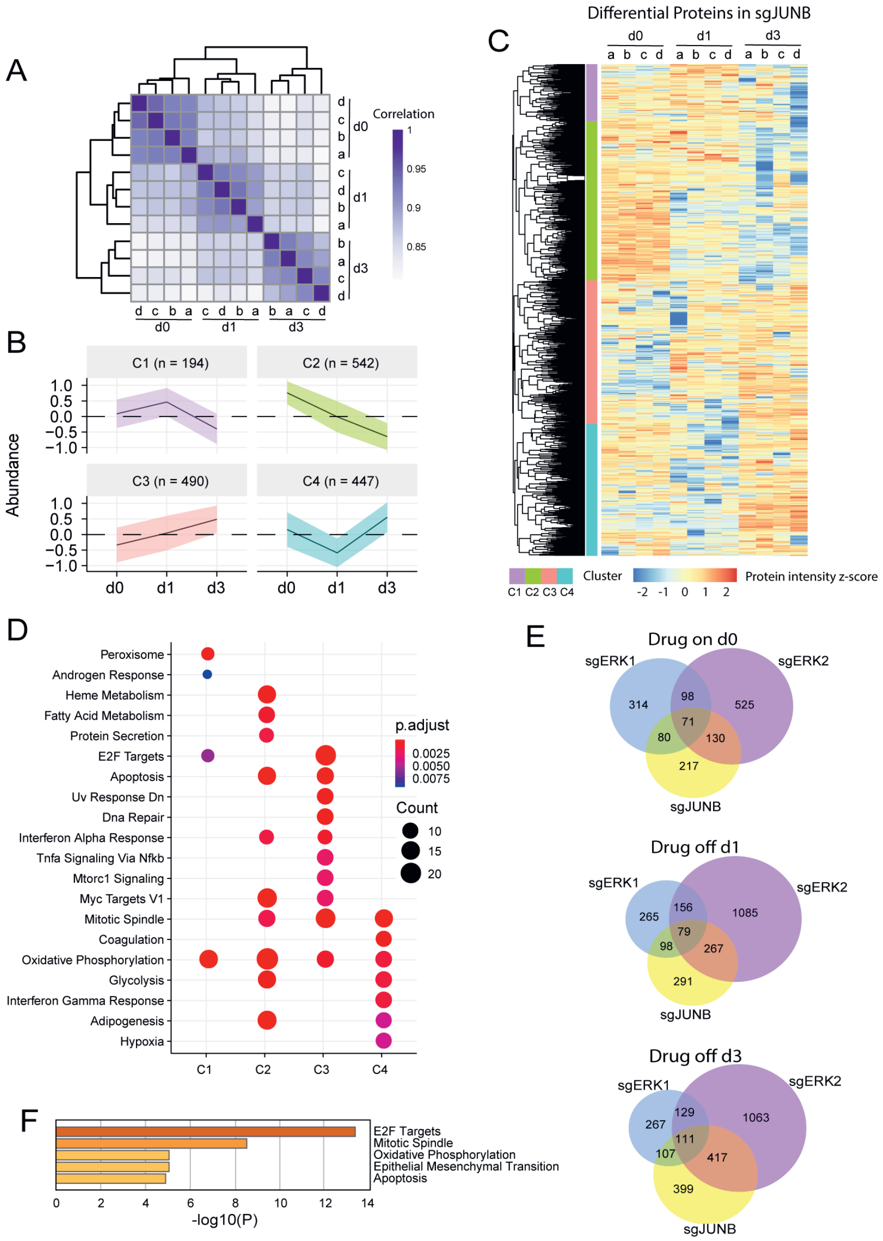
64. Kubic, J. D.; Mascarenhas, J. B.; Iizuka, T.; Wolfgeher, D.; Lang, D., GSK-3 promotes cell survival, growth, and PAX3 levels in human melanoma cells. Mol Cancer Res 2012, 10, (8), 1065-76.

Supplementary Material

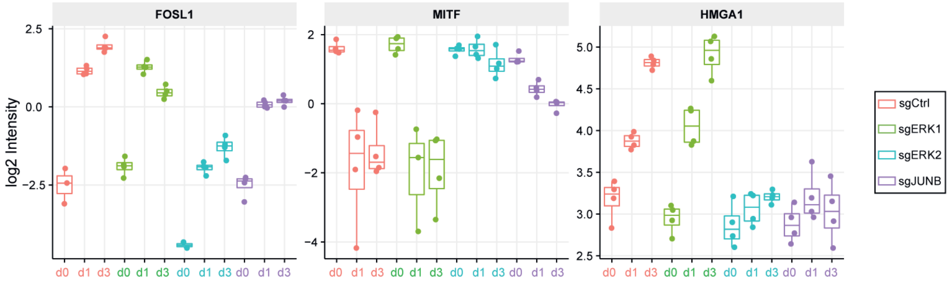
Supplementary Figures



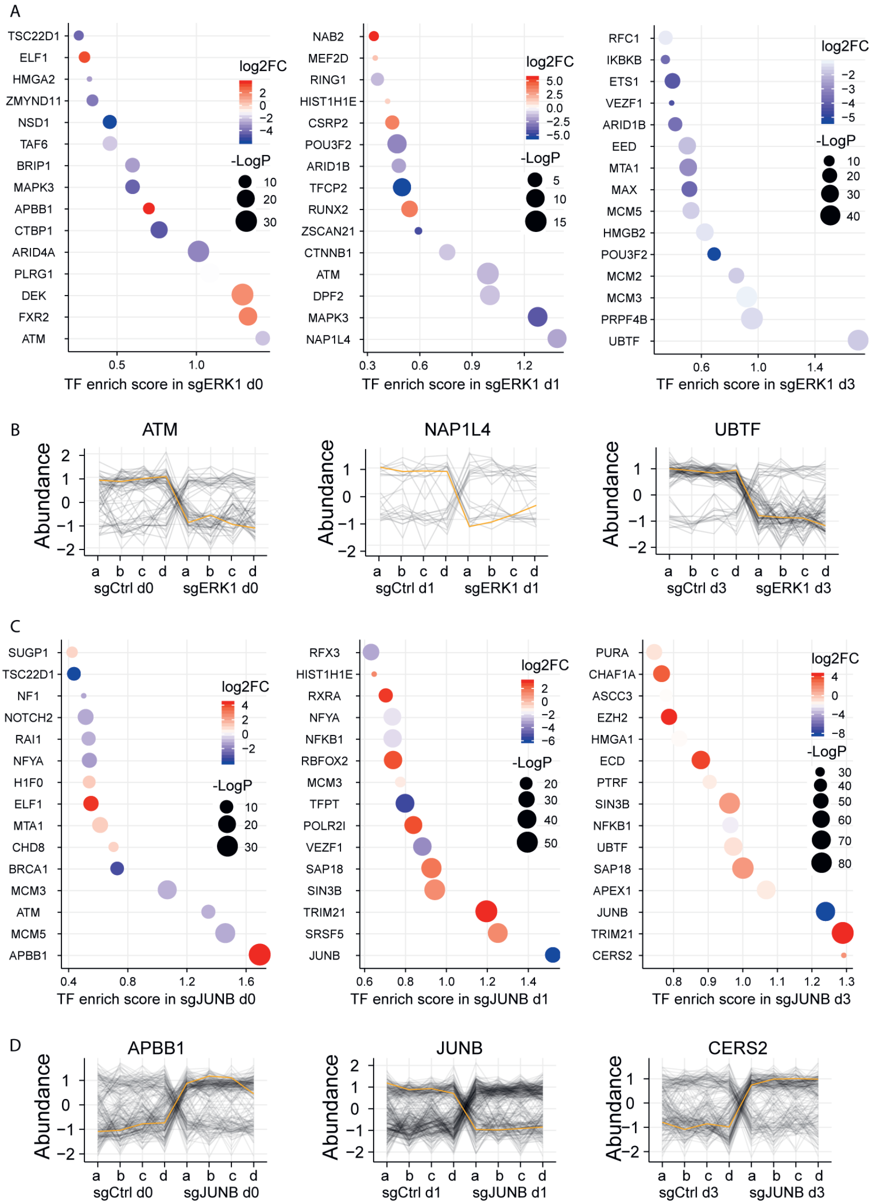
Supplementary Figure 1. Heatmap of sample Pearson correlation in sgERK2 and sgERK1 cells.



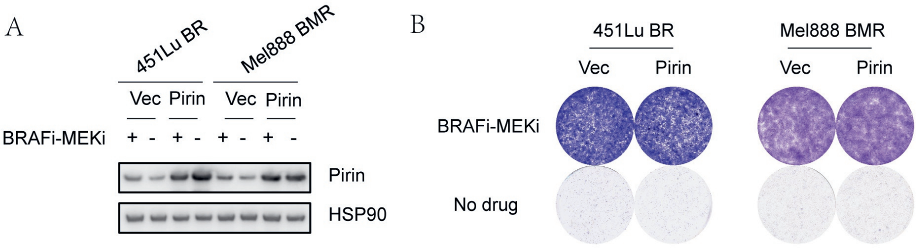
Supplementary Figure 2. Differential proteins in sgJUNB. A, Hierarchical clustering of sample Pearson correlation. B, Mean abundance of proteins in each cluster over time; shading denotes ±1 SD. C, Heatmap with differential changed proteins in sgJUNB compared to sgCtrl upon drug on (day0 a, b, c and d) and drug withdrawal (day 1 & 3 a, b, c and d). D, Hallmark enrichment using proteins in each cluster, dots size shows gene count in specific hallmark term, color presents enrichment p-value. E, Venn diagram with significantly changed proteins in three sgRNA cells, drug on day 0; drug withdrawal day 1 and day 3. F, Hallmark enrichment using differential proteins in sgERK2 and sgJUNB upon drug removal day 3.



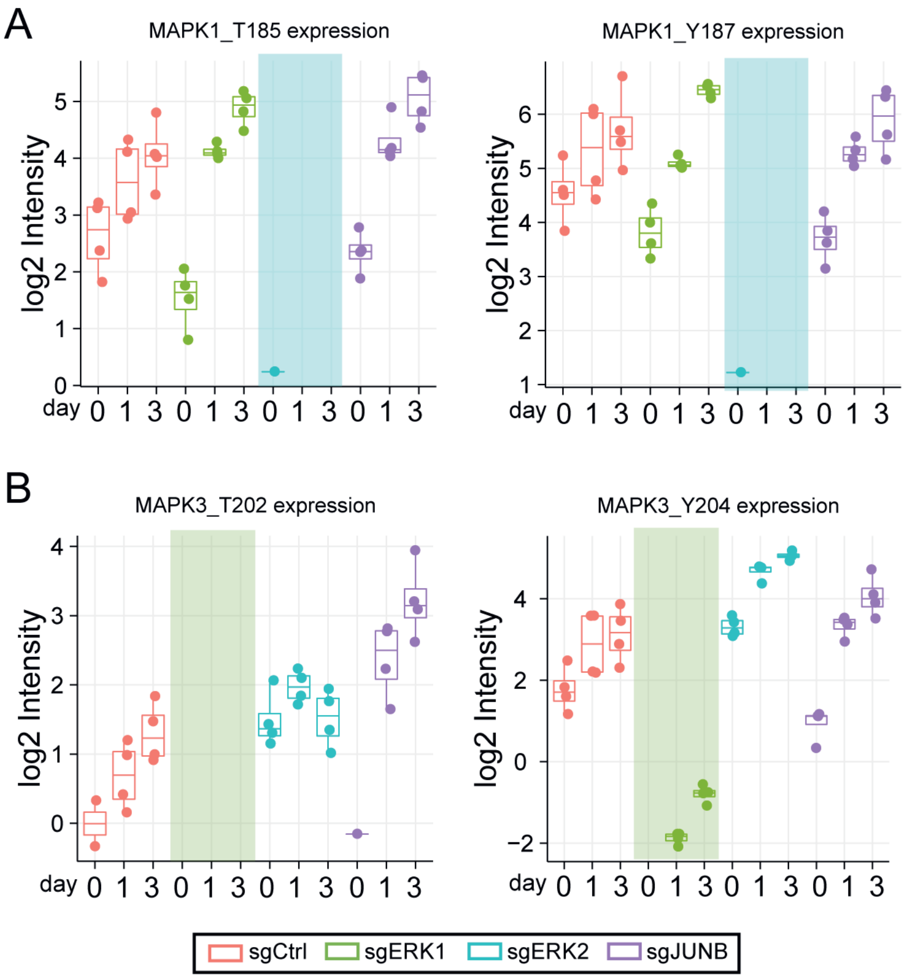
Supplementary Figure 3. Box plot representing the intensity of three proteins FOSL1, MITF, and HMGA1 in control (red), sgERK1 (green), sgERK2 (blue) and sgJUNB (purple) cells.



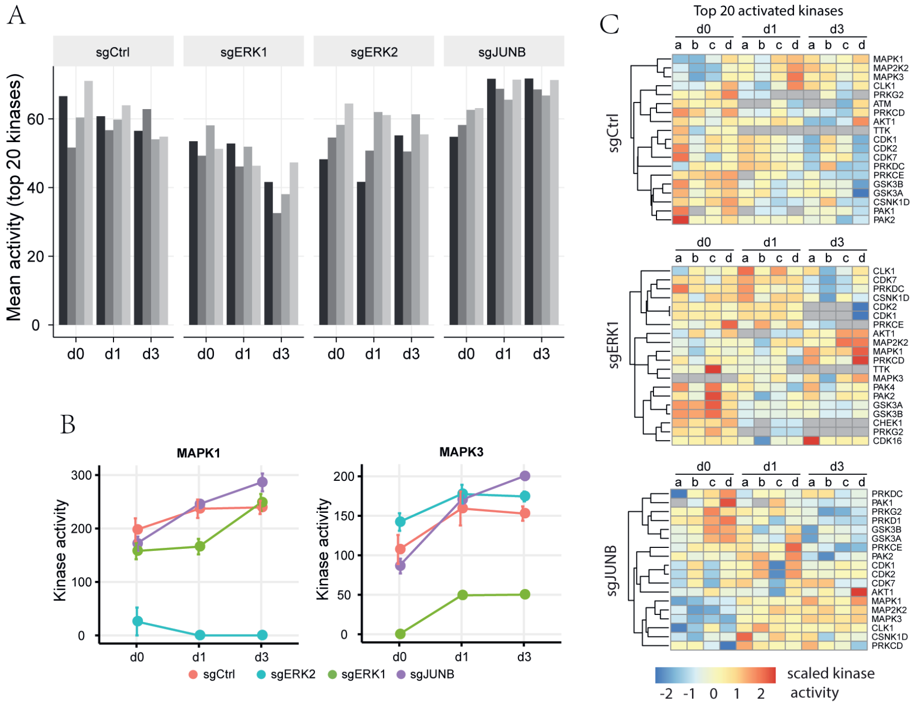
Supplementary Figure 4. Transcription factors (regulators) enrichment. A and C, Top 15 significant transcription factors (regulators) enriched in sgERK1 and sgJUNB in three time point, shades by log2 fold change of protein abundance, and size by enriched –log P-value. B and D, Express abundance of transcription factors (yellow highlighted line) with co-expression proteins (black lines).



Supplementary Figure 5. Pirin overexpression in drug resistant melanoma cells. A) PIR was overexpressed in 451LuBR, Mel888BMR cells, respectively. (BR, BRAFi-resistant, BMR, BRAFi + MEKi-resistant). B) PIR overexpression does not lead to reversal of the drug addiction phenotype.



Supplementary Figure 6. Phosphorylation of MAPK1 (ERK2) and MAPK3 (ERK1) in control (red), sgERK2 (blue), sgERK1 (green), and sgJUNB (purple) cells.



Supplementary Figure 7. Kinase activity inferred from the phosphoproteome. A, Barplot shows the average activities for the top 20 activated kinases in each sample. B, MAPK1 and MAPK3 activity over time, error bar presents ±1 SE, control (red), sgERK2 (blue), sgERK1 (green), and sgJUNB (purple) C, Top activated kinases in control, sgERK1, and sgJUNB cells.

Chapter 3

Exploring the role of PD-1 in CD8$^+$ T cell activation by Proteomic and phosphoproteomics profiling

Exploring the role of PD-1 in CD8$^+$ T cell activation by Proteomic and phosphoproteomics profiling

Bohui Li[1,2], David W Vredevoogd[3], Daniel S. Peeper[3] and Maarten Altelaar*[,1,2,4]

1 Biomolecular Mass Spectrometry and Proteomics Group, Utrecht Institute for Pharmaceutical Science, Utrecht University, Utrecht, The Netherlands.

2 Netherlands Proteomics Center, Padualaan 8, 3584 CH Utrecht, The Netherlands

3 Division of Molecular Oncology and Immunology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands

4 Mass Spectrometry and Proteomics Facility, The Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands

Manuscript under revision

Abstract

CD8$^+$ T cells play crucial roles in the adaptive immune response to clear pathogens. PD-1 (programmed cell death-1) is one of the central inhibitory receptors regulating CD8$^+$ T cell activation. Interestingly, it has recently shown that PD-1 was expressed by activated CD8$^+$ T cells during T cell receptor (TCR) stimulation and acute viral infection, but the role of PD-1 in regulating T cell activation is not well defined. To increase the knowledge of the role of PD-1 in T cell activation, we applied a mass spectrometry-based proteomic approach on resting and anti-CD3 stimulated CD8$^+$ T cells, in which PD-1 was silenced using CRISPR–Cas9. PD-1 was upregulated upon activation in wild type CD8$^+$ T cells while, conversely, low to almost none PD-1 was detected in PD-1 knockout CD8$^+$ T cells. Our quantitative mass spectrometry profile reveals that T cell receptors activated T cells reconstructed their proteome and phosphoproteome marked by activating of mTORC1 pathway. Importantly, we find that silencing of PD-1 altered the E3 ubiquitin-protein ligases, and increased glucose and lactate transporters. Moreover, the phosphorylation peptides-centric signaling analysis shows that knocking out of PD-1 evokes more phosphorylation events on mTORC1 pathway, and activates epidermal growth factor and its downstream MAPK pathway in regulating CD8$^+$ T cell activation. Our results depict possible mechanisms of PD-1 in response to TCR stimulation in CD8$^+$ T cells, which may provide a guide in immune homeostasis and immune checkpoint therapy.

Introduction

CD8$^+$ T cells play a critical role in controlling viral, intracellular bacterial and parasitic infections. Precise regulation of antigen specific CD8$^+$ T cell activation and proliferation is crucial for acquiring effector function, enabling them to specifically lyse target cells. Upon stimulation of antigens, the naive CD8$^+$ T cells trigger a signaling cascade to initiate T-cell surface receptors (TCRs), and modulate costimulatory and inhibitory receptors and signals, to ultimately become activated T cells (1). Receptors such as CD3 and CD28 transduce signals necessary to activate T cells. In contrast, receptors like programmed cell death protein 1 (PD-1) transduce signals that are inhibitory to lymphocyte activation (2). The intricate balance between positive and negative costimulatory signals is thought to enable effective immune response while preventing unnecessary T cell activation and maintaining immunological homeostasis (3, 4).

PD-1 is the most investigated inhibitory receptor in cell types such as T- and B-lymphocytes (5), showing its importance in immune regulation. The C57BL/6 mice that lack PD-1 were reported to develop lupus-like arthritis and glomerulonephritis (6) and BALB/c PD-1-deficient mice develop fatal dilated cardiomyopathy with IgG deposition (7). Recent studies have highlighted the importance of PD-1 in cancer and immunotherapy (2). Therapeutic blockade of the PD-1 or PD-L1 inhibitory receptors by antibodies has been approved in clinical cancer treatment, including patients with melanoma and non–small cell lung cancer (NSCLC) (8-10).

Considering the importance of PD-1 in regulating immune tolerance and cancer treatment, studies deciphering the characteristics of PD-1 in T cell activation are urgently needed.

PD-1 conveys its negative signals through two tyrosine-based structural motifs; one is the immunoreceptor tyrosine-based inhibitory motif (ITIM), the other one is the immunoreceptor tyrosine-based switch motif (ITSM) (11). When engaged with a ligand, PD-1 becomes phosphorylated at these motifs, leading to the recruitment of protein tyrosine phosphatases, such as SHP-2 (12). These tyrosine phosphatases affect downstream signaling pathways through the dephosphorylation of kinases and other signaling molecules , including phosphoinositide 3-kinase (PI3K)–AKT and the mitogen activated protein kinase pathway (MAPK) RAS-MEK-ERK (13-15), resulting in a decrease in T cell activation, proliferation, and survival as well as altered metabolism. Naïve T cells manifest a metabolically quiescent phenotype and acquire energy via oxidative phosphorylation (OXPHOS) by breaking down glucose, fatty acids, and amino acids (16). During activation via T cell receptor stimulation, T cells undergo a metabolic reprogramming to aerobic glycolysis, enabling T cells to meet their energy requirements for differentiation and proliferation (17). Interestingly, there is increasing evidence for a connection between the PD-1 pathway and metabolic reprogramming in T cell activation. PD-1 was reported to modulate metabolic reprogramming during naive T cell activation by inhibiting the upregulation of glucose and glutamine metabolism (14, 18). Besides, the PD-1 pathway can also promote lipolysis and fatty acid oxidation in activated CD4$^+$ T cells (18).

Therefore, we reasoned that depicting how proteome and phosphorylation events modulating T cell activation are affected by PD-1 is of major interest to better understand immune responses and immune checkpoint therapy. In the present study, high-resolution mass spectrometry was used to analyze the proteome and phosphoproteome of naïve and TCR-stimulated CD8$^+$ T cells. To shed light on the role of PD-1, a CRISPR–Cas9 approach was employed to genetically knockout PD-1. We quantified >5,000 proteins and >16,800 phosphorylation sites, providing a valuable resource that shows how immune activation and PD-1 reshape the proteomics and phosphoproteomics landscape of naïve and activated CD8$^+$ T cells.

Materials and methods

Cell culture and PD-1 knockout

Murine T cells were handled as described elsewhere (Vredevoogd et al., 2022; man. in rev.). Briefly, murine CD8$^+$ T cells were isolated from the spleens of OT-I/Cas9 mice and maintained in RPMI with 9% fetal bovine serum, penicillin (100U/mL), streptomycin (100µg/mL), 2-Mercaptoethanol (50µM, Merck), murine IL-2 (10ng/mL, ImmunoTools), murine IL-7 (0.5ng/mL, ImmunoTools) and murine IL-15 (1ng/mL, ImmunoTools). CD8$^+$ T cells were isolated from spleens by using the Dynabeads Untouched Mouse CD8 Cells kit (Thermo Fisher Scientific) following manufacturer's instructions and activated on non-tissue culture treated 24-well plates (Corning) coated with anti-CD3 (0.25µg per well, Thermo Fisher Scientific) and anti-CD28 antibodies (2.5µg per well, Thermo Fisher Scientific). Two days after isolation, cells were transduced with a retrovirus encoding sgRNAs targeting Pdcd1 (5’-GCTCAAACCATTACAGAAGG-3’) or a non-targeting control (5’-GTATTACTGATATTGGTGGG-3’) and, after a further two days, selected with puromycin (4µg/mL, Sigma). Ten days after transduction, cells were reactivated, or not, with plate-bound anti-CD3 (2.5µg per well in a 12-well plate) and analyzed by mass spectrometry or flow cytometry. For flow cytometric analysis, cells were stained in 50µL of a 0.1% BSA in PBS solution, with a 1:50 dilution of PD-1-PE antibody, a 1:100 dilution of CD137-APC antibody and a 1:100 dilution of CD8-FITC antibody (all Miltenyi). Dead cells were marked with DAPI (BD). Samples were analyzed on an LSRFortessa Flow Cytometer (BD).

Protein digestion

T-cell pellets were lysed, reduced and alkylated in heated Guanidine-HCl buffer as described previously (19). After dilution to <2mM Guanidine-HCl, aliquots containing equal protein amounts (as determined with Pierce Coomassie (Bradford) Protein Assay Kit) were digested twice (o/n and 4h) with trypsin (Sigma-Aldrich; enzyme/substrate ratio 1:75) at 37°C. Digestion was quenched with formic acid (5% final concentration), after which peptides were desalted on Sep-Pak C18 cartridges (Waters, MA, USA). After desalting, eluate aliquots (5%) were taken from each sample for proteome analysis, the remaining digest (900 µg) intended for phosphopeptide enrichment. All eluate fractions were vacuum dried in a SpeedVac and stored at -80°C until LC-MS/MS analysis or phosphopeptide enrichment. Phosphopeptides were enriched using the High-Select Fe-NTA Phosphopeptide Enrichment Kit (Thermo Scientific) according to the manufacturer’s instructions. Phosphopeptide eluates were vacuum dried and stored at -80°C until LC-MS/MS analysis. Prior to mass spectrometry, digests were reconstituted in 2% formic acid.

Mass spectrometry

For single-shot proteome analysis, peptide mixtures (1 µg aliquots) were analyzed by nanoLC-MS/MS on a Q Exactive HF-X Hybrid Quadrupole-Orbitrap Mass Spectrometer equipped with an EASY-NLC 1200 system (Thermo Scientific). Samples were directly loaded onto the analytical column (ReproSil-Pur 120 C18-AQ, 2.4µm, 75 µm × 500 mm, packed in-house). Solvent A was 0.1% formic acid/water and solvent B was 0.1% formic acid/80% acetonitrile. Peptides were eluted from the analytical column at a constant flow of 250 nl/min in a non-linear 210-min gradient containing a 5-min increase from 2% to 10% solvent B, a 125-minute increase to 24% B, followed by a 40-min increase to 35% B, a 20-min increase to 60% B and a 5-min ramp to 100% solvent B, ending with a 15-min wash. Mass spec settings were as follows: full MS scans (375-1,500 m/z) were acquired at 60,000 resolution with an AGC target of 3×10^6 charges and max injection time of 45 ms. Loop count was set to 20 and only precursors with charge state 2-7 were sampled for MS2 using 15,000 resolution, an MS2 isolation window of 1.4 m/z, 1 × 10^5 AGC target, a maximum injection time of 22 ms and a normalized collision energy of 26.

For phosphoproteome analysis, peptides were analyzed on an Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific) equipped with the same LC setup as described for proteomes. Phosphopeptides were eluted from the analytical column at a constant flow of 250 nl/min in a linear 210-min gradient containing a 180-min increase from 7% to 28% solvent B. Mass spec settings were as follows: full MS scans (375-1,500 m/z) were acquired at 60,000 resolution with an AGC target of 1 × 10^6 charges and max injection time of 120 ms. The machine was operated in top-speed mode with 3s cycle time. Only precursors with charge state 2-6 were sampled for MS2 using 60,000 resolution, an MS2 isolation window of 1.6 m/z, 1 × 10^5 AGC target, a maximum injection time of 180 ms, and normalized collision energy of 35.

Data processing

The proteome RAW files were analyzed by Proteome Discoverer (version 2.3.0.523) using standard settings for feature detection and precursor ion quantification. Spectra were searched against the Swissprot M. musculus database (2019_02, 17,005 entries) using SequestHT with 50 ppm and 0.06 Da as precursor and fragment mass tolerances, respectively. Results were filtered using Percolator to achieve 1%FDR at protein and peptide level and SequestHT PSM score Xcorr>1 was applied as additional filter. Trypsin/P was specified as cleavage specificity; carbamidomethylation (C) was used as fixed modification and oxidation (M), protein N-terminal acetylation and deamidation (N, Q) were specified as variable modifications.

The phosphoproteome RAW files were analyzed with label-free quantitation (LFQ) in MaxQuant (version 1.6.5.0) (20) using standard settings and ‘match between runs’. FDR 0.01 was used as cutoff at protein and peptide level. Trypsin/P was selected as enzyme; spectra were searched against the same database as described for proteomes and the same fixed and variable modifcations were specified in the searches, with the exception that phospho(S, T, Y) was included among the variable modifications.

Data analysis and statistics

All data analysis were conducted under R-3.5.1 environment (21) and Microsoft Excel. Raw intensities extracted from proteome discover (PD, proteome) and MaxQuant (phosphoproteome) were first log2 transformed and then normalized by the median of replicates. Furtherly, t-Distributed Stochastic Neighbor Embedding (t-SNE) (22) was carried out to assess the reproducibility of the experiments. Next, the Empirical Bayes Statistics for Differential Expression (eBayes) function in the limma R package (23) was carried out to analyze the differential expression (DE) proteins and phosphosites, followed by a Benjamini-Hochberg multiple testing correction with 5% FDR. The DE proteins and phosphosites with adjusted p-value < 0.05 and odds ratio > 1.5 were used for further analysis and functional annotation by Gene Ontology (GO), KEGG, and Hallmark databases (FDR < 0.05) (24).

Transcription factors (regulators) enrichment

Next, to depict the critical transcription factors/regulators in regulating T cell activation, we employed the RegEnrich R package on the proteome profile (25, 26). The RegEnrich integrity of proteome expression profiling, transcription factors (regulators) to construct a WGCNA co-expression network and define the importance of the regulators. Specifically, it can be simplified into the following steps: a) construction a data-driven co-expression network using proteome profiles b) deducing subnetwork of interest (i.e., using differential changed protein list to obtain a TFs connected sub-network); and c)referring importance of TFs by Fisher's exact test based on subnetwork. An enrichment score for TFs was given by incorporating the exhibited significance of differential expression ($P_D < 0.05$) and significance of enrichment ($P_E < 0.05$). The overall scores of TFs were calculated by:

$$score = norm(-log(p_E)) + norm(-log(p_D)), norm(x) = \frac{x - min(x)}{max(x) - min(x)}$$

Phosphorylation-driven signature analysis

To identify phosphorylation-driven pathways across T cell activation, we performed the phosphosite-specific signature enrichment analysis (PTM-SEA) (27). The implementation of PTM-SEA available on GitHub (https://github.com/broadinstitute/ssGSEA2.0) was used on our phosphoproteome profiling. The following parameters were used to run PTM-SEA: weight: 0.75, statistic: "area.under.RES", output.score.type: "NES", nperm: 1,000, min.overlap: 5, correl.type: "z.score", ptmsigdb: "ptm.sig.db.all.flanking.mouse.v1.9.0.gmt". Signatures with FDR-corrected p-values < 0.1 in three out of six samples were considered to be differential between resting and activated T cells.

Results and Discussion

Quantitative proteomics and phosphoproteomics upon TCR Stimulation

To gain insight into the activation process of CD8$^+$ T lymphocytes (CTLs) and examine the role of programmed cell death-1 (PD-1, encoded by the Pdcd1 gene) receptor, in response to TCR stimulation, we employed a label-free quantitative (phospho)proteomics approach. Specifically, the naive CD8$^+$ T cells were purified from the OT-1 murine spleen and divided into two subpopulations. The WT subpopulation was subjected to vehicle sgRNA (i.e., WT, Fig. 1A), and the other subpopulation was subjected to Pdcd1 sgRNA (i.e., PD-1 KO, Fig. 1A). To improve reproducibility and knockout efficiency, we applied two different sgRNAs on both WT and PD-1 KO groups. Subsequently, stimulation of the T cell receptor (TCR) with anti-CD3 (a-CD3) was carried out for 24 hours. Samples were collected before and after TCR stimulation, lysed and in-solution digested by LysC/trypsin, followed by liquid chromatography-tandem mass spectrometry (nanoLC-MS/MS) on a high-resolution mass spectrometer (Q-Exactive

HFx) (Fig. 1A). 5% of the pool was used for whole proteome analysis, and the remaining 95% for phosphopeptide enrichment and analysis.

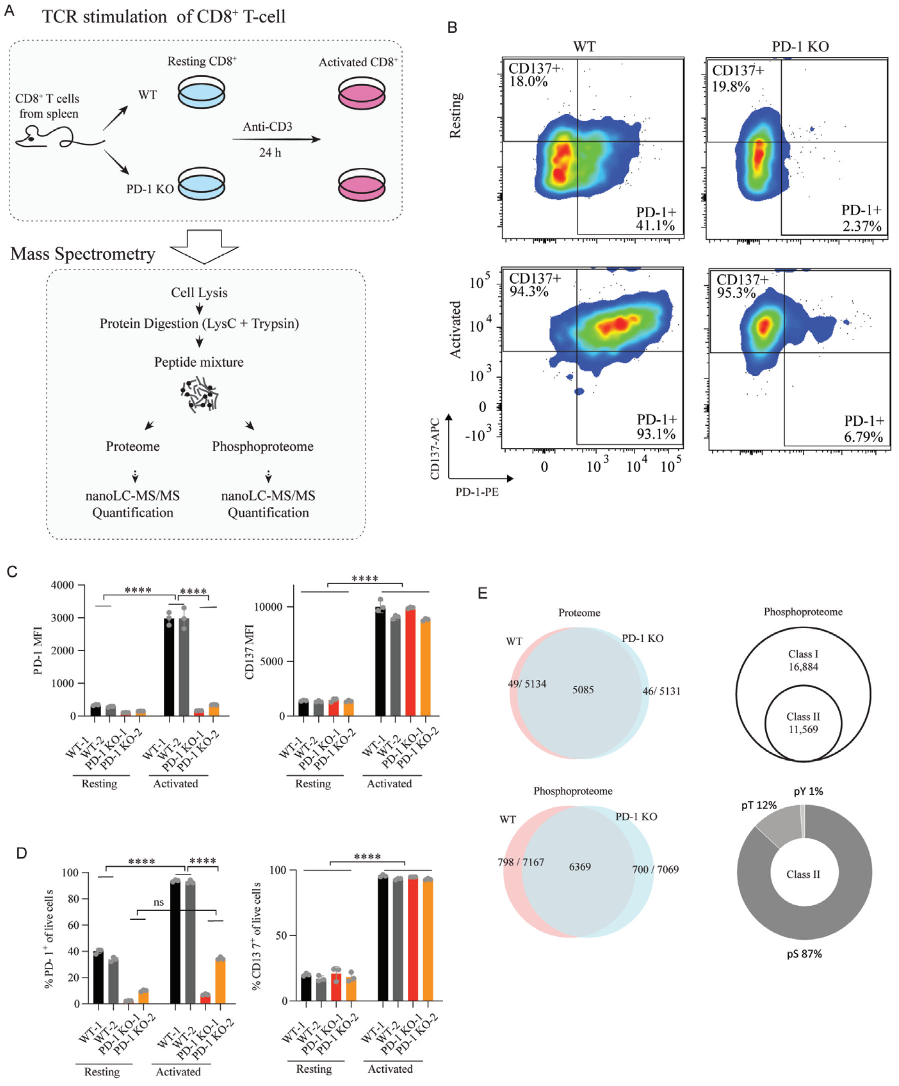


Figure 1. Experimental design and expression of PD-1 and CD137. A, CD8 T cells were isolated from the OT-1 murine spleen and divided into two subpopulations. Next, the WT cells were subjected to two vehicle single guide RNAs (sgRNA), each with 3 bio-replicates, respectively; the PD-1 KO cells were subjected to two Pdcd1 specific sgRNAs, each with 3 bio-replicates, respectively. Both WT and PD-1 KO group were stimulated with anti-CD3 for 24-h (24 hours) and followed by mass spectrometry analysis for (phospho) proteome. B, Cells were analyzed by flow cytometry. The representative FACS plots for gating strategy to define frequencies of CD137$^+$ CD8$^+$ T cells and PD-1$^+$ CD8$^+$ T cells were shown, left were WT cells and right were Pdcd1 KO cells. C, Statistical differences of median fluorescence intensity (MFI) in frequencies of PD1-expressing CD8$^+$ T cells between resting and activated by one-way ANOVA with Tukey post-hoc testing (****, $P < 0.0001$, error bars denote ±SD). D, Statistical bar plots showing the percentage of PD-1$^+$ CD8$^+$ T cells and CD137$^+$ CD8$^+$ T cells (****, $P < 0.0001$, one-way ANOVA, error bars denote ±SD). E, Venn plot compares the total number of quantified proteins and phosphorylation sites in WT and PD-1 KO cells. Total number of class I and class II phosphosites quantified. Distribution of Ser/Thr/Tyr phosphosites identified.

PD-1 plays a crucial role in CD8$^+$ T cell function and has been shown to function as an inhibitory receptor during the early stage of T cell activation (28). To further investigate PD-1 function in CD8$^+$ T cell activation upon TCR stimulation, we first assessed the expression of PD-1 using flow cytometry. We observed a significant upregulation of PD-1 expression upon TCR stimulation in WT CD8$^+$ T cells (from 41.1% to 93.1%), which was absent in PD-1 knockout CD8$^+$ T cells (Fig. 1B). Quantification of the FACS results in three biological replicates (Fig. 1C) shows significant upregulation (FDR < 0.0001) of PD-1 expression between resting and activated T cells in wild type, yet no difference is observed in PD-1 KO cells (Fig. 1C, left panel). The observation of PD-1$^+$ CD8$^+$ T cells further verifies the upregulation of PD-1 during T cell activation (Fig. 1D, left panel).

CD137 (also known as 4-1BB, encoded by the Tnfrsf9 gene) is a member of the tumor necrosis factor receptor (TNFR) family that is expressed on activated T cells (29), and can co-stimulate T cell activation and proliferation (30). Here, we evaluated T cell activation efficiency by monitoring the expression of CD137. Notably, we observed a significant increase of CD137 in both WT and PD-1 KO CD8$^+$ T cells after 24 hours of TCR stimulation (Fig. 1B and 1C, right panel). Moreover, also the percentage of live CD137$^+$ CD8$^+$ cells showed a significant increase in both conditions (Fig. 1D, right panel). Consistently, the expression of CD137 observed in our proteomics experiment showed an increase upon TCR stimulation in both WT and PD-1 KO T cells (Fig. S1). In conjunction with CD137, several other activation markers including CD30, Lag3, Irf8, Atp2a2, Pdcd4, and Txnip identified in our proteomics experiment, showed expression behavior in agreement with previously published results (Fig. S1) (1).

Next, having confirmed activation of our CD8$^+$ T cells and efficient KO of PD-1, we started mining our proteomics and phosphoproteomics datasets. In total, we quantified 5,134 and 5,131 proteins with at least two unique peptides in WT and PD-1 KO CD8$^+$ T cells, respectively (Fig. 1E and Table S1A). Around 99% of proteins were identified in both control and PD-1 KO cells. For the phosphoproteome analysis, we identified 16,884 phosphosites, of which 11,569 with a localization probability over 0.75. Stringently filtering of phosphosites (with quantitative values in at least four out of six replicates) resulted in 7,167 and 7,069 phosphosites in WT and PD-1 KO CD8$^+$ T cells, respectively (Fig. 1E and Table S2A). The t-Distributed Stochastic Neighbor Embedding (t-SNE) analysis using the whole proteome and phosphoproteome datasets shows a clear clustering of biological replicates withing the same treatment group, where cells within WT and PD-1 KO clustered tightly together, and a strong segregation between resting and activated populations (Fig. S2A). In agreement with these t-SNE results, unsupervised hierarchical clustering of the Pearson correlations within the proteome and phosphoproteome showed high correlation of the replicates within the same treatment group (Fig. S2B). These results together demonstrate the high reproducibility of the proteome and phosphoproteome datasets.

Proteome response upon T cell activation

To compare proteome changes between naive and activated CD8$^+$ CTLs upon antigen activation, the Empirical Bayes Statistics for Differential Expression (eBayes) function in the limma R package (23) was separately carried out for WT and PD-1 KO cells. 639 significantly changed proteins (Table S1B) during WT CD8$^+$ T cell activation were quantified with 5% FDR and > 1.5-fold change, where 333 proteins showed significant upregulation upon TCR stimulation (Fig. 2A, left). In PD-1 KO cells, 706 differential changed proteins (Table S1C) were quantified (5% FDR and > 1.5-fold change), where 385 proteins significantly increased in expression during activation (Fig. 2A, right).

Next, we compared the significantly changed proteins in WT and PD-1 KO T cells, which we functionally annotated using downregulated (DN) and upregulated (UP) protein subsets. We observed that around 1/3 of the proteins showing a change in expression were in common between WT and PD-1 KO CD8$^+$ T cells upon activation (Fig. 2B). Interesting enrichments observed for the downregulated proteins were, in the case of the overlapping proteins (DN overlap); IFNγ and IFNα responses. (Fig. 2C, left panel), while proteins with decreased expression specific to WT CD8$^+$ T cells were enriched in TCR signaling, cancer and cell cycle pathways and PD-1 KO CD8$^+$ T cells specific proteins enriched for MAPK and IL2 STAT5 signaling (Fig. 2C, left).

Hallmark analysis of the upregulated proteins in common between the two conditions (UP overlap) shows MYC targets (V1 and V2) and mTORC1 signaling as two of the most dominant pathways, in line with the reported metabolic reprogramming in the process of T cell activation (17). Moreover, cytokine signaling (IL-2 and TNFA signaling), and the unfolded protein response are strongly enriched (Fig. 2C, right). Interestingly, WT and PD-1 KO cells also show slight differences in the MYC targets (V1 and V2) and mTORC1 signaling enrichment terms, together with other relevant hallmarks such as ‘glycolysis’ and ‘complement’, indicating these processes are affected upon PD-1 KO. Thus, TCR stimulation for 24 hours resulted in profound reprogramming of the metabolic machineries, associated with activation of mTORC1 and MYC pathways, showing slight difference between cell lines.

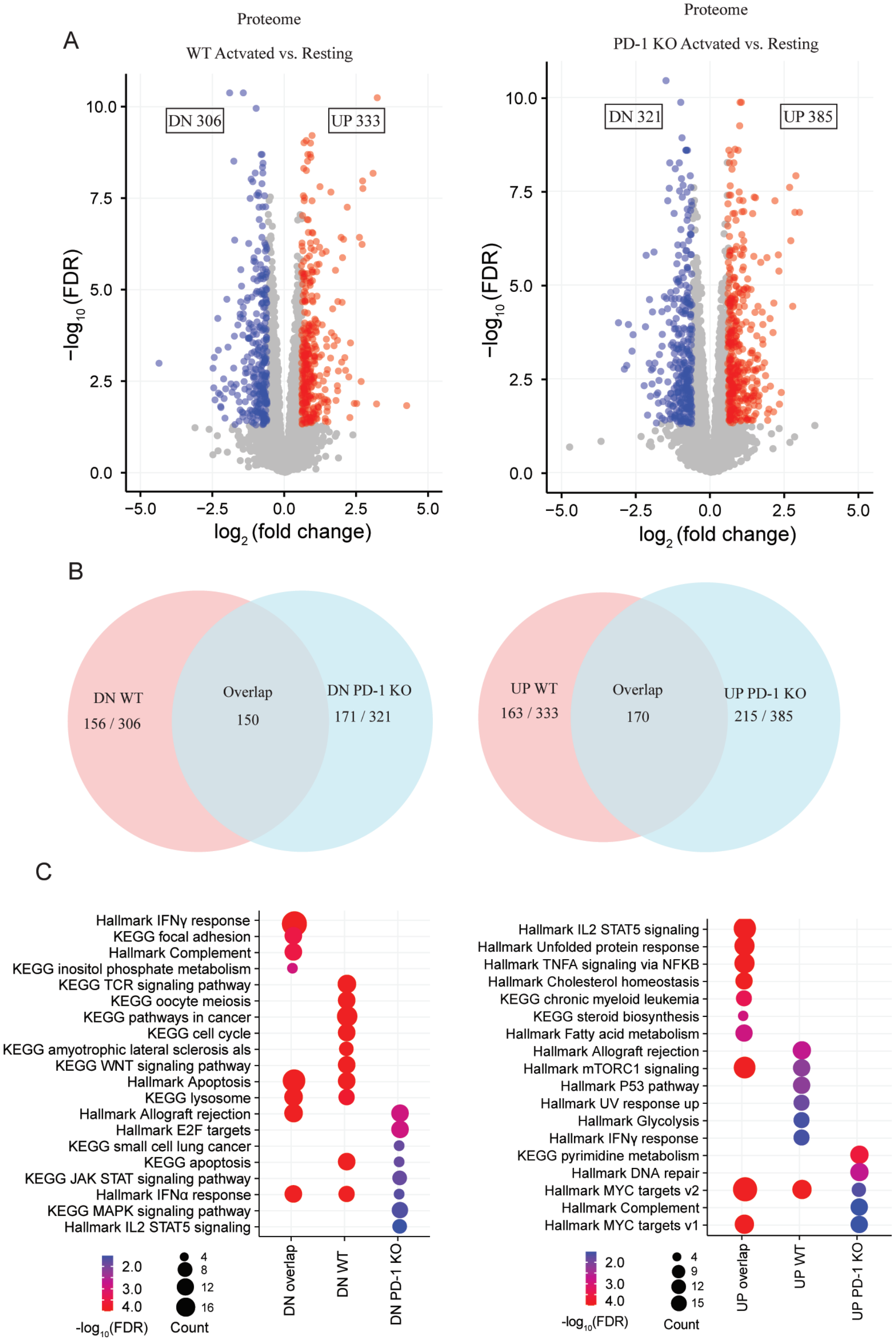


Figure 2. Differential expression proteins upon TCR stimulation and functional annotation. A, Volcano plots showing relative protein expression values (Log2-transformed and median normalized protein intensities) in WT (left) and PD-1 KO (right) cells. Red dots highlighted upregulation proteins (UP, FDR < 0.05, and >1.5 fold change) in activated T cells, blue dots highlighted downregulation proteins (DN, FDR

< 0.05, and <-1.5 fold change) in activated T cells, and grey dots are not significant proteins. B, Venn plots compares upregulated and downregulated proteins. DN indicates downregulation proteins and UP indicates upregulation proteins. C, KEGG pathway and Hallmark signaling enrichment with significantly downregulated proteins (left) and upregulated proteins (right).

PD-1 KO affects expression of E3 ubiquitin-protein ligases and metabolic transporters

To further investigate the role of PD-1 in TCR stimulated activation of CD8⁺ T cells, we compared the dynamic changes of the proteome under WT and PD-1 KO conditions using the "interaction models" comparison in the limma R package (23). A total of 595 proteins showed a significantly different protein expression change upon TCR activation (Table S1D) with a p-value smaller than 0.05 in WT versus PD-1 knockout cells. Unsupervised hierarchical clustering of these significant proteins identified four protein clusters with similar trends (Fig. 3A), showing opposite expression patterns between WT and PD-1 KO cells. Function annotation analysis using proteins in cluster 1 and 3 shows, amongst others, strong enrichment of ubiquitin mediated proteolysis, the cell cycle (E2F targets and G2M checkpoint), and metabolism (mTORC1 signaling, amino and nucleotide sugar metabolism, and pyrimidine metabolism) (Fig. 3B and 3C).

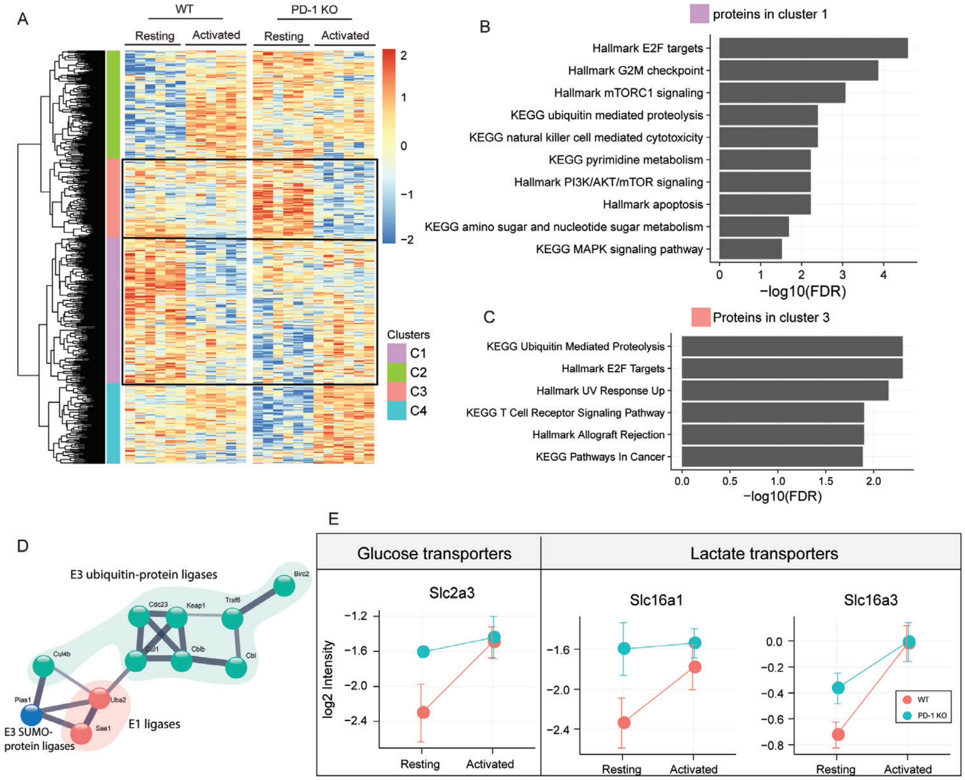


Figure 3. Proteome response to TCR stimulation and PD-1 KO. A, Heatmap showing significant proteins that evoked by TCR stimulation and PD-1 knockout, which colored by the relative protein expression values (z-scored proteins intensities) and clustered by unsupervised hierarchical clustering. B & C, Functional annotation with proteins in cluster 1 and cluster 3. D, Protein-protein interaction network of proteins enriched in protein ligases. E, Line plot for glucose and lactate transporters with relative protein intensities (log2 transformed and median normalized intensities) in WT (red) and PD-1 KO (blue), error bar presents ±1 SE.

Ubiquitin mediated proteolysis is the most significant pathway enriched in cluster 3 and also dominant in cluster 1, suggesting ubiquitin protein ligases may be affected by PD-1 KO in CD8$^+$ T cells. Therefore, we further investigated the quantitative expression of proteins involved in this pathway. We quantified 11 protein ligases of which eight are E3 ubiquitin-protein ligases, one E3 SUMO ligase, and two are SUMO-activating E1 ligases (Fig. 3D and S3). We found that PD-1 KO increases the expression of ubiquitination ligases including Cblb, Birc2, Uba2, and Cdc23 in resting CD8$^+$ T cells and returns to a level comparable to that of the WT T cells upon activation (Fig. S3). Conversely, ubiquitin ligases, such as Cul1, Cul4b, Cbl, and Traf6, were up-regulated in activated PD-1 KO cells (Fig. S3). Notably, several E3 ubiquitin ligases, such as Cbl and Cbl-b, have been demonstrated to be involved in regulating immune response during infection by targeting specific inhibitory molecules for proteolytic destruction (31). The Cbl-b (encoded by Cblb gene) and Traf6 were reported as negative regulators for maintaining immune homeostasis (32, 33).

Nutrient transporters are essential components of T cell environment-sensing machinery as they act as ‘gatekeepers’ to control the activity of nutrient-sensing kinases (34). PD-1 signaling was reported to modulate metabolic reprogramming during T cell activation by inhibiting the upregulation of glucose and glutamine metabolism (14, 18). The present data show that silencing of PD-1 could reverse the inhibitory effect by increasing the expression of one glucose transporter (Slc2a3) and two lactate transporters (Slc16a1 and Slc16a3) in resting CD8$^+$ T cells (Fig. 3E).

Collectively, these findings show that silencing of the PD-1 receptor in CD8$^+$ T cells altered the expression of several key players in the E3 ubiquitin-protein degradation system and affected metabolic pathways by regulating glucose and lactate transporters.

Transcription factor (regulators) analysis

Transcription factors and their regulators play a crucial role in controlling cell growth, cell differentiation, and proliferation. T cell differentiation has been reported to remodel the pattern of expression of transcription factors and regulators (34). Therefore, to depict transcription factors/regulators involved in T cell activation, a module based weighted protein (gene) co-expression network analysis (WGCNA) was carried out to explore correlations between differentially expressed proteins and transcription factors (Fig. 4A) (25, 35). Next, a one-tailed hypergeometric test was used to determine the importance of highlighted transcription factors and their regulators (36, 37). An enrichment score for TFs was given by incorporating the exhibited significant differential expression ($p < 0.05$) and significant enrichment ($p < 0.05$)

(detailed in Methods).

It was notable that there were more similarities than differences between WT and PD-1 KO CD8⁺ T cells, as most of the key transcription factors, including Irf4, JunB, Nfkb2, and Rel, are significant in both cell types (Fig. S4B and S4C). The transcriptional activator Irf4, which was significantly up-regulated in both WT and PD-1 KO T cells, has been demonstrated to regulate TCR affinity-mediated metabolic programming and clonal expansion of CD8⁺ T cells (38). The transcription factor Jun-B (one of the AP-1 family members), can dimerize with Batf in CD8⁺ T cells and enable the BATF–JUN heterodimer to interact with Irf4 and Irf8 (39). Furthermore, the Nfkb transcription factors can dimerize with proto-oncogenes, including RelA, RelB, or c-Rel (encoded by Rel gene), in regulation of T-lymphocyte differentiation and effector functions (40).

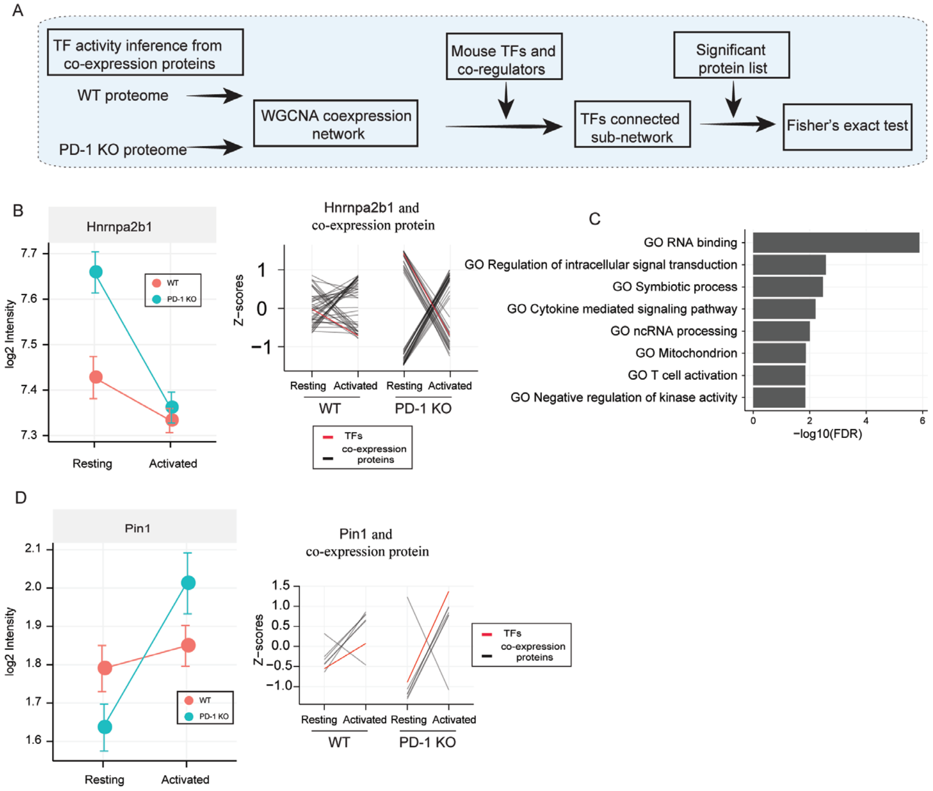


Figure 4. Transcription factors (regulators) enrichment. A, Overview of the WGCNA co-expression network-based transcription factors and regulators enrichment. B, (left) Expression of Hnrnpa2b1 in WT (red) and PD-1 KO (blue), error bar presents ±1 SE. (right) Hnrnpa2b1 (highlighted in red) and corresponding co-expression proteins (grey lines). C, Gene ontology enrichment of Hnrnpa2b1 co-expressing proteins. D, (left) Expression of Pin1 in WT (red) and PD-1 KO (blue) cells, error bar presents ±1 SE. (right) Pin1 (highlighted in red) and corresponding co-expression proteins (grey lines).

Next, we generated a co-expression network using the proteome from both WT and PD-1 KO cells and looked for enrichment of proteins that significantly differ in expression upon antigen stimulation in PD-1 KO cells. This step resulted in heterogeneous nuclear ribonucleoproteins A2/B1 (Hnrnpa2b1) as the most significant regulator (Fig. 4B). Notably, we observed that Hnrnpa2b1 and proteins showing co-expression portray differences in expression in resting PD-1 KO T cells versus WT T cells, which is reversed upon TCR stimulation (Fig. 4B). Functional annotation analysis revealed the Hnrnpa2b1 co-expressing proteins are involved in signal transduction, cytokine mediated signaling and T cell activation (Fig. 4C). Hnrnpa2b1 was reported to play a role as an anti-inflammatory regulator in patients with autoimmune endocrine disorders (41). Moreover, it was reported that heterogeneous nuclear ribonucleoprotein A2/B1 (Hnrnpa2b1) could interact with vol hippel lindau (Vhlα), and consequently modulate pyruvate kinase (Pkm) transcript splicing and reprogram cellular glucose metabolism (42, 43). Besides, we observed a significant increase of expression of Pin1 (Peptidyl-prolyl cis-trans isomerase 1) and co-expressing proteins upon TCR stimulation in PD-1 KO cells (Fig. 4D). Importantly, Pin1 has been described to regulate cells that participate in immune system (44, 45). Pin1 also involved in the activation of T cells by modulating the activity of the transcription factor NFAT and regulating activation-induced cytokine production (46, 47). A recently study has revealed that Pin1 induces lysosomal degradation of PD-L1 in pancreatic ductal adenocarcinoma (48).

Phosphoproteome profiling across T cell activation

T lymphocyte activation is a complex process involving a variety of associated receptors and kinases, which initiates multiple signal transduction pathways. To investigate the cellular signaling cascade and to understand the initial activation steps in CD8$^+$ T cells, we performed phosphoproteome profiling. The Empirical Bayes Statistics for Differential Expression (eBayes) function in the limma R package (23) was separately carried out for the WT and PD-1 KO cells phosphoproteome to analyze differential regulated phosphosites during CD8$^+$

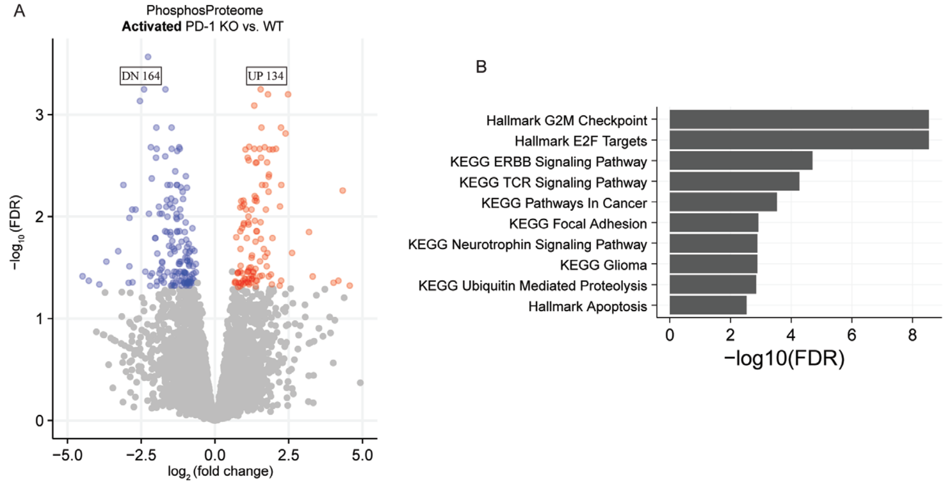


Figure 5. Phosphoproteome response in activated CD8$^+$ T cells. A, Volcano plot indicates significant phosphosites in activated CD8$^+$ T cells (PD-1 KO versus WT in activated T cells). Red dots highlighted

upregulation phosphosites (UP), blue dots highlighted downregulation phosphosites (DN). FDR < 0.05, and 1.5 fold change. B, Functional enrichment with differential expression phosphosites located proteins (FDR < 0.05).

T cell activation. As shown in Figure S5A, we identified 1067 significantly changed (5% FDR and > 1.5-fold change) phosphosites (Table S2B) during WT CD8$^+$ T cell activation. These phosphorylation events are dominated by cell cycle activity (G2M checkpoint, E2F targets, and mitotic spindle), metabolism (MYC target V1 and V2), cytokine signaling (TNFα signaling via NFκB, and IL2 STAT5 signaling) and TCR signaling (Fig. S5B). Silencing of PD-1 resulted in 1392 significantly regulated phosphosites (Table S2C) upon antigen stimulation, which mapped to 884 unique proteins (Fig. S5C). The functional enrichment of these phosphorylated proteins revealed similar pathways as in the WT condition (Fig. S5D).

To get insight into the phosphoproteome response to PD-1 knockout in CD8$^+$ T cells, we performed a differential expression comparison between WT and PD-1 KO in resting and activated, separately. Unexpectedly, we observed 134 up-regulated and 164 down-regulated phosphosites (Table S2D) upon T cell activation, while almost no significant phosphosites were identified between these conditions in resting CD8$^+$ T cells (Fig. 5A and S5E). These results show, that in contrast to expression changes at the proteome level (Fig. S4A), the effect of PD-1 KO is more clearly present at the phosphoproteome level. The observed changes in phosphorylation reveled, besides involvement in the processes described above, enrichment in the ERBB signaling pathway (Fig. 5B).

Phosphopeptide based enrichment signature

To further assess the regulation of signaling involved in T cell activation, a site-centric PTM signature enrichment analysis (PTM-SEA) approach was applied to the phosphoproteome data (49). The PTM-SEA predicted that 7 and 12 signatures were significantly modified upon activation in WT and PD-1 KO CD8$^+$ T cells, respectively (Fig. 6A). Notably, in both WT and PD-1 KO cell populations, we observed a significant increase of activity of the mTOR, CDK1 and CDK5 signature and a dramatic decrease of activity of rapamycin signaling. Where, the mTOR pathway is known to regulate glucose metabolism and glycolysis (50), and modulate the differentiation, and migratory ability of CD8$^+$ cytotoxic T cells (51), rapamycin has been shown to inhibit activation of T cells and B cells by reducing their sensitivity to interleukin-2 (IL-2) through mTOR inhibition (52).

Activation of mTOR signaling leads to phosphorylation of the ribosomal protein S6 (Rps6 or Rs6) at phospho-sites Ser235/236 and Ser240/244 by the S6-kinase (S6K) (53), ultimately promoting cell growth (54). Indeed, we observed a significant increase of ribosomal protein S6 phosphorylation at sites Ser240/244 in WT CD8$^+$ T cells, and at sites Ser236, Ser240/244 in PD-1 KO CD8$^+$ T cells (Fig. 6B). Simultaneously, no change was detected in phosphorylation of the translational repressor 4EBP1 (eukaryotic initiation factor 4E binding protein-1) (Fig. 6B), thereby relieving its translational inhibition. Furthermore, the mammalian autophagy-initiating kinase Ulk1, a key regulator of autophagy whose activity is determined by mTOR

(55), showed increase of phosphorylation at site Ser 622 and Ser 555 in PD-1 KO cells in our data (Fig. 6B). Taking these data together shows that metabolic reprogramming plays a key role in governing T cell activation, and PD-1 KO evokes increased phosphorylation in this program.

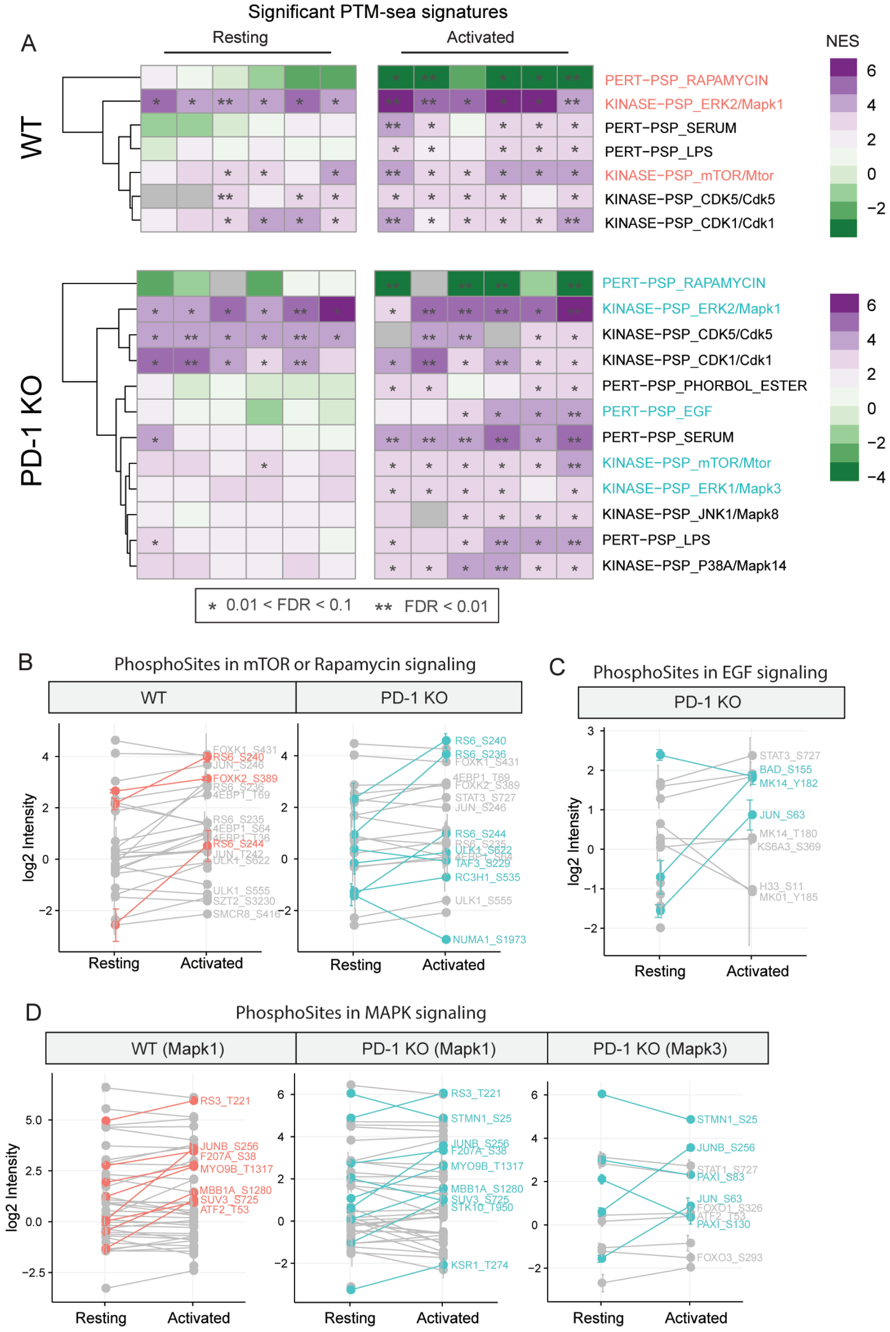


Figure 6. Phosphorylation sites centric signature enrichment. A, Heat-map indicated the significantly enriched signatures in WT and PD-1 KO T cells, colored by the normalized enrichment score (NES). The asterisk highlighted the significance of enrichment (*, 0.01 < FDR < 0.1; **, FDR < 0.01). B-D, Line plots showing phosphosites that enriched in mTOR and Rapamycin, EGF, and MAPK signaling, respectively. Lines colored in red are significant expression phosphosites in WT cells, while lines colored in blue are significant expression phosphosites in PD-1 KO cells, error bar presents ±1 SE.

The epidermal growth factor (EGF) is a protein that stimulates cell growth and differentiation by binding to its receptor, EGFR (56). Our data showed enrichment of the ERBB2 pathway (Fig 5B) and the EGF signature (Fig. 6A and 6C) in PD-1 KO cells upon TCR stimulation. Several key phosphorylation sites were enriched in EGF signaling, including upregulation of Jun Ser63 and Mk14 Tyr182 phosphosites, and downregulation of the Bad Ser155 phosphosite (Fig. 6C). Moreover, we observed strong activation of MAPK pathways, especially in PD-1 KO CD8$^+$ T cells (Fig. 6A and 6C), downstream of EGFR signaling (57, 58). MAPK is demonstrated to coordinately regulate glutamine uptake and metabolism for T cell activation (59). These results indicate that strong activation of MAPK signaling may involve regulating T cell survival, glutamine uptake and metabolism. Silencing of PD-1 activates the epidermal growth factor and corresponding downstream cascade including MAPK1/2 signaling, amplifying the cascade reaction in modulating CD8$^+$ T cell activation.

Conclusion

In this study, we present a proteomics and phosphoproteomics study in wild type and PD-1 knockout CD8$^+$ T cells upon T cell receptor stimulation by anti-CD3. PD-1 was observed to significantly upregulate in CD8$^+$ T cells upon activation. Our quantitative mass spectrometry reveals that T cell receptors stimulated CD8$^+$ T cells reprogram their phospho(proteome) and activated the mTORC1 pathway. Depletion of PD-1 altered the E3 ubiquitin-protein ligases, increased glucose, and lactate transporters. Interestingly, there were more similarities than differences between WT and PD-1 KO CD8$^+$ T cells on proteome level, however, an opposite result was observed on phosphoproteome level. We find that silencing of PD-1 induced more significant phosphorylation sites in regulating mTOR signaling, and activated the epidermal growth factor and corresponding downstream MAPK pathway. Our result reveal how PD-1 influence the proteome and phosphoproteome in CD8$^+$ T cells upon TCR activation, which may contribute to future strategies for checkpoint blockage therapy.

Acknowledgements

This research was part of the Netherlands X-omics Initiative and partially funded by NWO (Project 184.034.019) and the China Scholarship Council (CSC) no. 201606300049.

References

1. Tan, H., et al., Integrative Proteomics and Phosphoproteomics Profiling Reveals Dynamic Signaling Networks and Bioenergetics Pathways Underlying T Cell Activation. Immunity 2017, 46, (3), 488-503.

2. Sharpe, A. H.; Pauken, K. E., The diverse functions of the PD1 inhibitory pathway. Nat Rev Immunol 2018, 18, (3), 153-167.

3. Khoury, S. J.; Sayegh, M. H., The roles of the new negative T cell costimulatory pathways in regulating autoimmunity. Immunity 2004, 20, (5), 529-38.

4. Parry, R. V., et al., CTLA-4 and PD-1 receptors inhibit T-cell activation by distinct mechanisms. Mol Cell Biol 2005, 25, (21), 9543-53.

5. Francisco, L. M.; Sage, P. T.; Sharpe, A. H., The PD-1 pathway in tolerance and autoimmunity. Immunol Rev 2010, 236, 219-42.

6. Nishimura, H.; Minato, N.; Nakano, T.; Honjo, T., Immunological studies on PD-1-deficient mice: implication of PD-1 as a negative regulator for B cell responses. International Immunology 1998, 10, (10), 1563-1572.

7. Nishimura, H., et al., Autoimmune dilated cardiomyopathy in PD-1 receptor-deficient mice. Science 2001, 291, (5502), 319-322.

8. Ribas, A., et al., Association of Pembrolizumab With Tumor Response and Survival Among Patients With Advanced Melanoma. Jama-Journal of the American Medical Association 2016, 315, (15), 1600-1609.

9. Garon, E. B., et al., Pembrolizumab for the treatment of non–small-cell lung cancer. New England Journal of Medicine 2015, 372, (21), 2018-2028.

10. Ribas, A.; Wolchok, J. D., Cancer immunotherapy using checkpoint blockade. Science 2018, 359, (6382), 1350-+.

11. Shinohara, T.; Taniwaki, M.; Ishida, Y.; Kawaichi, M.; Honjo, T., Structure and Chromosomal Localization of the Human Pd-1 Gene (Pdcd1). Genomics 1994, 23, (3), 704-706.

12. Chemnitz, J. M.; Parry, R. V.; Nichols, K. E.; June, C. H.; Riley, J. L., SHP-1 and SHP-2 associate with immunoreceptor tyrosine-based switch motif of programmed death 1 upon primary human T cell stimulation, but only receptor ligation prevents T cell activation. Journal of Immunology 2004, 173, (2), 945-954.

13. Riley, J. L., PD-1 signaling in primary T cells. Immunological Reviews 2009, 229, 114-125.

14. Parry, R. V., et al., CTLA-4 and PD-1 receptors inhibit T-cell activation by distinct mechanisms. Molecular and Cellular Biology 2005, 25, (21), 9543-9553.

15. Patsoukis, N., et al., Selective Effects of PD-1 on Akt and Ras Pathways Regulate Molecular Components of the Cell Cycle and Inhibit T Cell Proliferation. Science Signaling 2012, 5, (230).

16. van der Windt, G. J. W.; Pearce, E. L., Metabolic switching and fuel choice during T-cell differentiation and memory development. Immunological Reviews 2012, 249, 27-42.

17. O'Sullivan, D.; Pearce, E. L., Targeting T cell metabolism for therapy. Trends in Immunology 2015, 36, (2), 71-80.

18. Patsoukis, N., et al., PD-1 alters T-cell metabolic reprogramming by inhibiting glycolysis and promoting lipolysis and fatty acid oxidation. Nature Communications 2015, 6.

19. Jersie-Christensen, R. R.; Sultan, A.; Olsen, J. V., Simple and Reproducible Sample Preparation for Single-Shot Phosphoproteomics with High Sensitivity. Methods Mol Biol 2016,

1355, 251-60.

20. Cox, J., et al., Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol Cell Proteomics 2014, 13, (9), 2513-26.

21. Team, R. C., R: A language and environment for statistical computing. R Foundation for Statistical Computing 2013.

22. van der Maaten, L.; Hinton, G., Visualizing Data using t-SNE. Journal of Machine Learning Research 2008, 9, 2579-2605.

23. Ritchie, M. E., et al., limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 2015, 43, (7).

24. Subramanian, A., et al., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 2005, 102, (43), 15545-15550.

25. Tao, W.; Radstake, T. R.; Pandit, A., RegEnrich: An R package for gene regulator enrichment analysis reveals key role of ETS transcription factor family in interferon signaling. bioRxiv 2021.

26. Langfelder, P.; Horvath, S., WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008, 9, 559.

27. Krug, K., et al., A Curated Resource for Phosphosite-specific Signature Analysis. Mol Cell Proteomics 2019, 18, (3), 576-593.

28. Ahn, E., et al., Role of PD-1 during effector CD8 T cell differentiation. Proceedings of the National Academy of Sciences of the United States of America 2018, 115, (18), 4749-4754.

29. Halstead, E. S.; Mueller, Y. M.; Altman, J. D.; Katsikis, P. D., In vivo stimulation of CD137 broadens primary antiviral CD8(+) T cell responses. Nature Immunology 2002, 3, (6), 536-541.

30. Pollok, K. E., et al., Inducible T-Cell Antigen 4-1bb - Analysis of Expression and Function. Journal of Immunology 1993, 150, (3), 771-781.

31. Mueller, D. L., E3 ubiquitin ligases as T cell anergy factors. Nature Immunology 2004, 5, (9), 883-890.

32. Tran, C. W., et al., Glycogen Synthase Kinase-3 Modulates Cbl-b and Constrains T Cell Activation. Journal of Immunology 2017, 199, (12), 4056-4065.

33. King, C. G., et al., TRAF6 is a T cell-intrinsic negative regulator required for the maintenance of immune homeostasis. Nature Medicine 2006, 12, (9), 1088-1092.

34. Howden, A. J. M., et al., Quantitative analysis of T cell proteomes and environmental sensors during T cell differentiation. Nature Immunology 2019, 20, (11), 1542-+.

35. Langfelder, P.; Luo, R.; Oldham, M. C.; Horvath, S., Is My Network Module Preserved and Reproducible? Plos Computational Biology 2011, 7, (1).

36. Ramirez, R. N., et al., Dynamic Gene Regulatory Networks of Human Myeloid Differentiation. Cell Syst 2017, 4, (4), 416-429 e3.

37. Goode, D. K., et al., Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation. Developmental Cell 2016, 36, (5), 572-587.

38. Man, K., et al., The transcription factor IRF4 is essential for TCR affinity-mediated metabolic programming and clonal expansion of T cells. Nature Immunology 2013, 14, (11), 1155-U79.

39. Murphy, T. L.; Tussiwand, R.; Murphy, K. M., Specificity through cooperation: BATF-IRF interactions control immune-regulatory networks. Nature Reviews Immunology 2013, 13, (7), 499-509.

40. Visekruna, A.; Volkov, A.; Steinhoff, U., A key role for NF-κB transcription factor c-Rel in T-lymphocyte-differentiation and effector functions. Clinical and Developmental Immunology 2012, 2012.

41. Coppola, A., et al., Anti-Inflammatory Action of Heterogeneous Nuclear Ribonucleoprotein A2/B1 in Patients with Autoimmune Endocrine Disorders. Journal of Clinical Medicine 2020, 9, (1).

42. Liu, Y. B., et al., A novel VHL alpha isoform inhibits Warburg effect via modulation of PKM splicing. Tumor Biology 2016, 37, (10), 13649-13657.

43. Christofk, H. R., et al., The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. Nature 2008, 452, (7184), 230-U74.

44. Barberi, T. J.; Dunkle, A.; He, Y. W.; Racioppi, L.; Means, A. R., The prolyl isomerase Pin1 modulates development of CD8+ cDC in mice. PLoS One 2012, 7, (1), e29808.

45. Esnault, S.; Shen, Z. J.; Whitesel, E.; Malter, J. S., The peptidyl-prolyl isomerase Pin1 regulates granulocyte-macrophage colony-stimulating factor mRNA stability in T lymphocytes. Journal of Immunology 2006, 177, (10), 6999-7006.

46. Esnault, S., et al., Pin1 Modulates the Type 1 Immune Response. Plos One 2007, 2, (2).

47. Liu, W. F.; Youn, H. D.; Zhou, X. Z.; Lu, K. P.; Liu, J. O., Binding and regulation of the transcription factor NFAT by the peptidyl prolyl cis-trans isomerase Pin1. Febs Letters 2001, 496, (2-3), 105-108.

48. Koikawa, K., et al., Targeting Pin1 renders pancreatic cancer eradicable by synergizing with immunochemotherapy. Cell 2021, 184, (18), 4753-+.

49. Krug, K., et al., A Curated Resource for Phosphosite- specific Signature Analysis. Molecular & Cellular Proteomics 2019, 18, (3), 576-593.

50. Finlay, D. K., et al., PDK1 regulation of mTOR and hypoxia-inducible factor 1 integrate metabolism and migration of CD8(+) T cells. Journal of Experimental Medicine 2012, 209, (13), 2441-2453.

51. Sinclair, L. V., et al., Phosphatidylinositol-3-OH kinase and nutrient-sensing mTOR pathways control T lymphocyte trafficking. Nature Immunology 2008, 9, (5), 513-521.

52. Mukherjee, S.; Mukherjee, U., A comprehensive review of immunosuppression used for liver transplantation. Journal of transplantation 2009, 2009.

53. Thoreen, C. C., et al., A unifying model for mTORC1-mediated regulation of mRNA translation. Nature 2012, 485, (7396), 109-U142.

54. Harter, P. N., et al., Immunohistochemical Assessment of Phosphorylated mTORC1-Pathway Proteins in Human Brain Tumors. Plos One 2015, 10, (5).

55. Kim, J.; Kundu, M.; Viollet, B.; Guan, K. L., AMPK and mTOR regulate autophagy through direct phosphorylation of Ulk1. Nature Cell Biology 2011, 13, (2), 132-U71.

56. Herbst, R. S., Review of epidermal growth factor receptor biology. International Journal of Radiation Oncology Biology Physics 2004, 59, (2), 21-26.

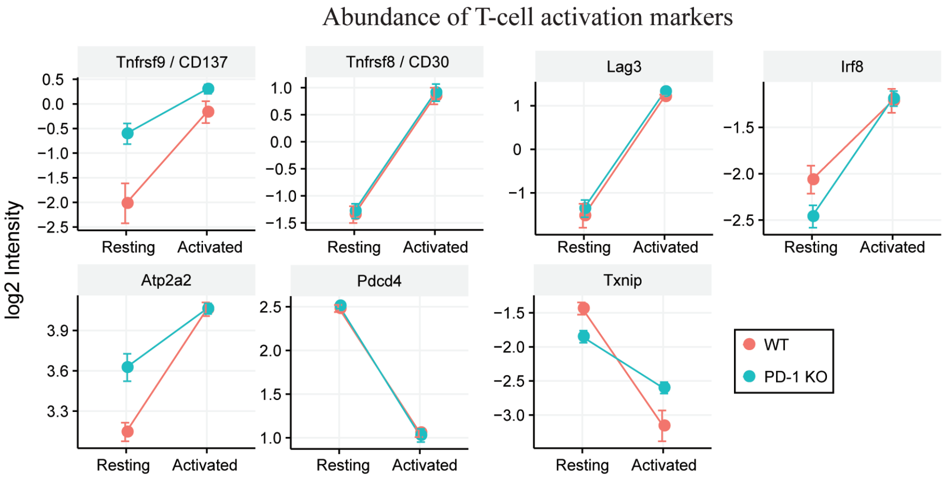
57. Wu, J., et al., Inhibition of the Egf-Activated Map Kinase Signaling Pathway by Adenosine-3',5'-Monophosphate. Science 1993, 262, (5136), 1065-1069.

58. Bunone, G.; Briand, P. A.; Miksicek, R. J.; Picard, D., Activation of the unliganded estrogen receptor by EGF involves the MAP kinase pathway and direct phosphorylation.

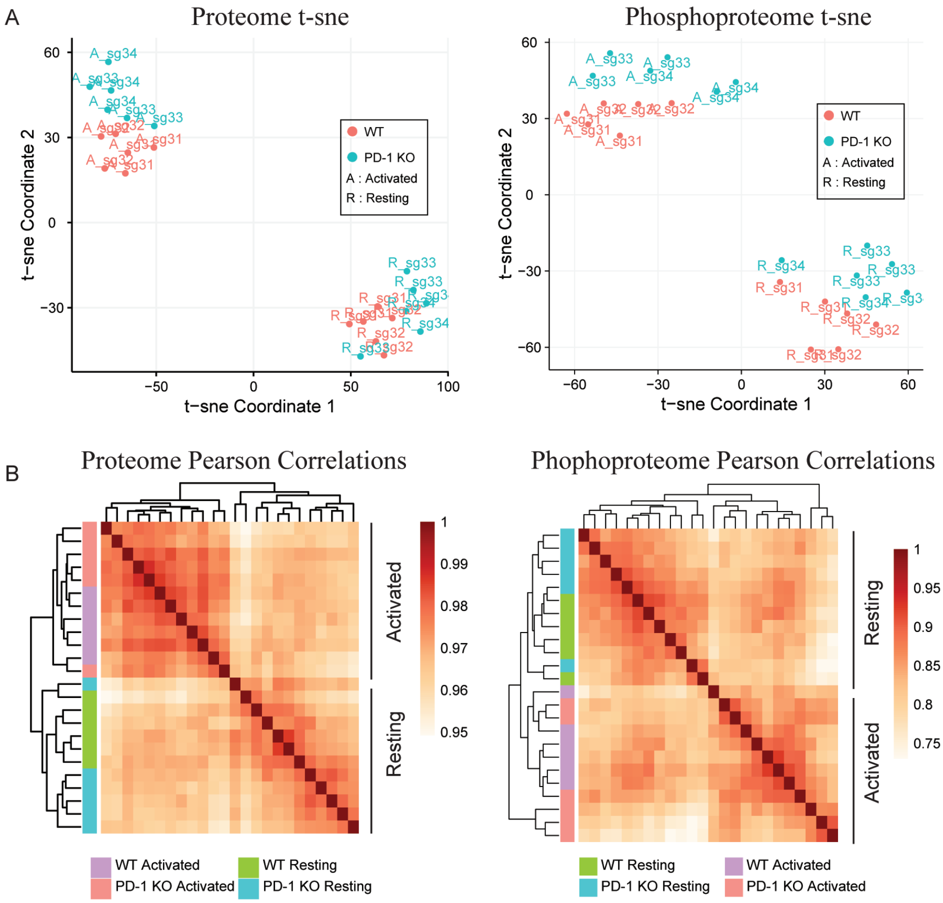
Embo Journal 1996, 15, (9), 2174-2183.

59. Carr, E. L., et al., Glutamine Uptake and Metabolism Are Coordinately Regulated by ERK/MAPK during T Lymphocyte Activation. Journal of Immunology 2010, 185, (2), 1037-1044.

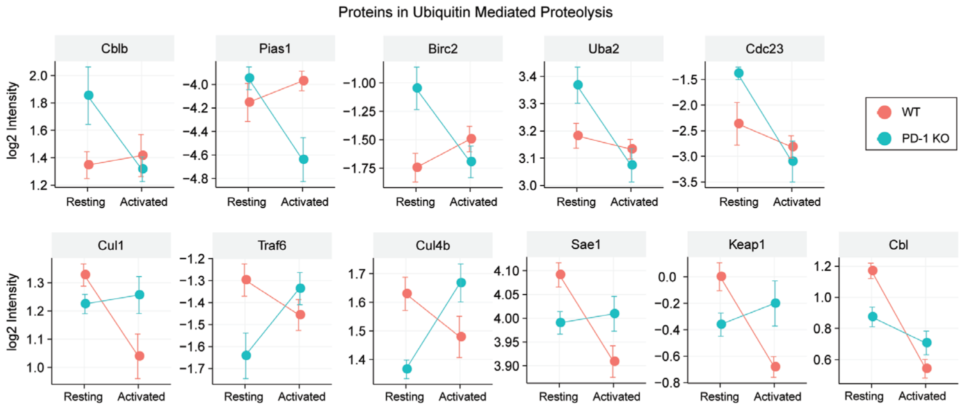
Supporting information for chapter 3



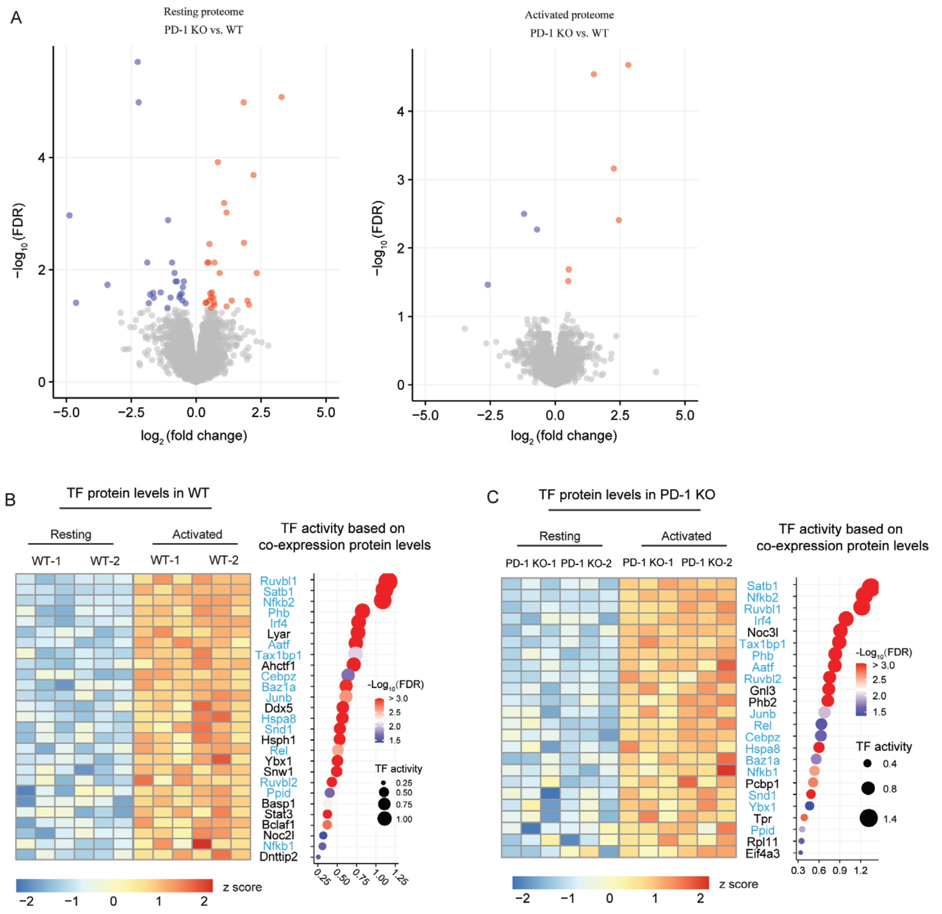
Supplementary Figure 1. Abundance of T-cell activation markers. Line plots for activation markers and important proteins in WT (red) and PD-1 KO (blue), error bar donates ±1 SE.



Supplementary Figure 2. Assessments of the experimental reproducibility. A, t-SNE plot based on proteome profile (left) and phosphoproteome profile (right), showing the trajectory of how WT and PD-1 KO CD8$^+$ T cells respond upon TCR stimulation. B, Heatmap of sample Pearson correlation with proteome profile (left) and phosphoproteome profile (right).

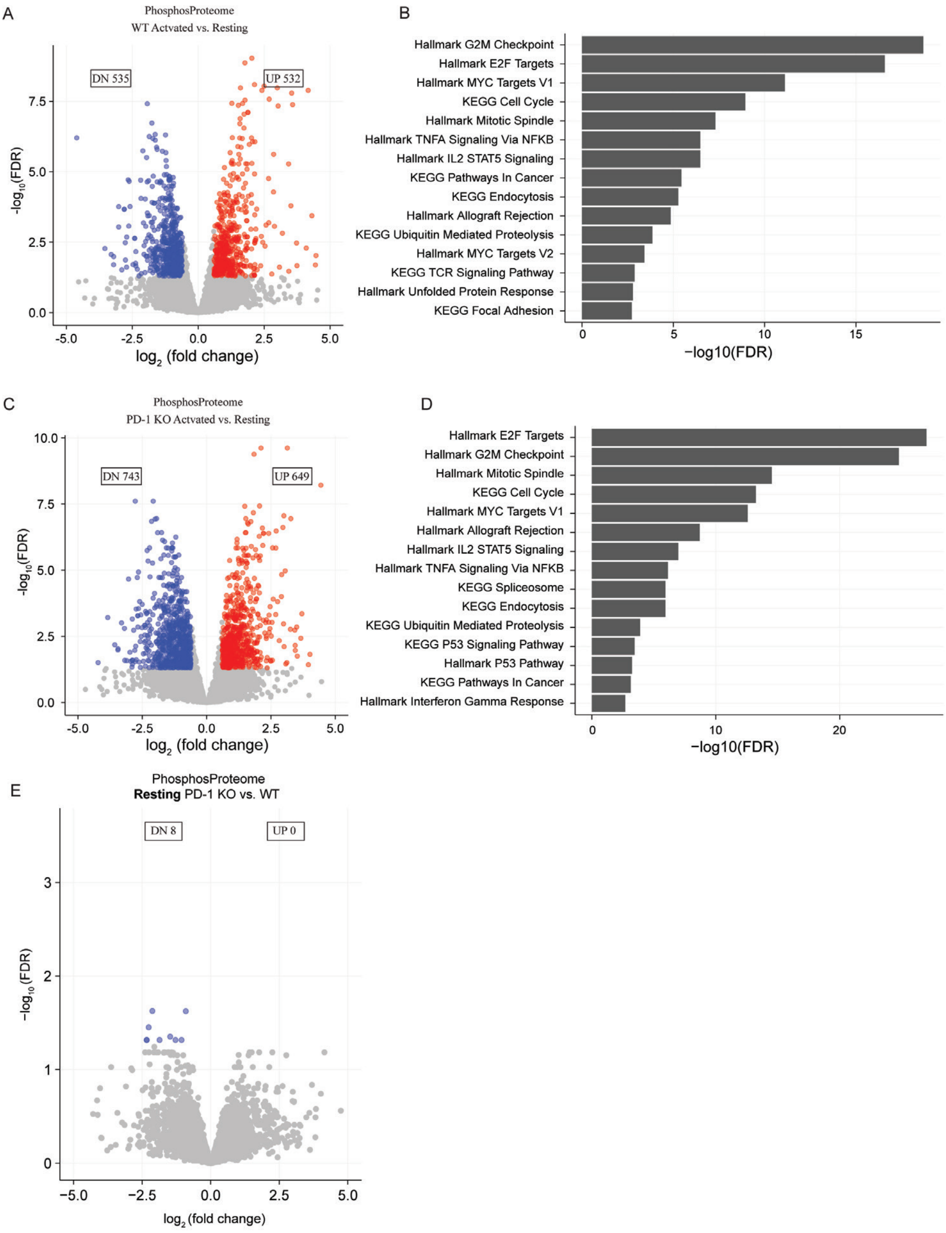


Supplementary Figure 3. Expression of proteins in the ubiquitin mediated proteolysis. Line plots showing the expression of protein ligases in WT (red) and PD-1 KO (blue) cells, error bar donates ±1 SE.



Supplementary Figure 4. Proteome response to PD-1 KO and significant transcription factors (regulators).

A, Volcano plot indicates significant proteins (FDR < 0.05) in resting and activated CD8$^+$ T cells, respectively. Red dots highlighted upregulation proteins, blue dots highlighted downregulation proteins. B & C, (left) Expression (z-scored intensity) of significant transcription factors and regulators enriched in WT and PD-1 KO T cells upon activation, respectively; (right) Enrichment score of transcription factors that shade by log2 fold change of expression intensity, and –log10(FDR) (significance in TF enrich) donates dots size.



Supplementary Figure 5. Phosphoproteome response upon TCR stimulation. A, Volcano plot indicates significant phosphosites in WT CD8$^+$ T cells upon activation. Red dots highlighted upregulation phosphosites (UP), blue dots highlighted downregulation phosphosites (DN). FDR < 0.05, and >1.5 fold change. B, Functional enrichment with differential expression phosphosites located proteins in WT CD8$^+$ T cells (FDR < 0.05). C, Volcano plot shows significant phosphosites in PD-1 KO CD8$^+$ T cells upon activation. Red dots represent upregulation phosphosites (UP), blue dots represent downregulation phosphosites (DN). FDR < 0.05, and >1.5 fold change. D, Functional enrichment with differential expression phosphosites located proteins in PD-1 KO CD8$^+$ T cells (FDR < 0.05). E, Volcano plot for significant phosphosites in resting CD8$^+$ T cells (PD-1 KO versus WT in resting T cells).

Chapter 4

Identification of protein complexes by integrating protein abundance and interaction stoichiometries using a deep learning strategy

Identification of protein complexes by integrating protein abundance and interaction stoichiometries using a deep learning strategy

Bohui Li[1,2], Maarten Altelaar[1,2,3], Bas van Breukelen*[,1,2]

1 Biomolecular Mass Spectrometry and proteomics group, Padualaan 8, 3584 CH Utrecht.

2 Utrecht Institute for Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, the Netherlands

3 Mass Spectrometry and Proteomics Facility, The Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands

Int. J. Mol. Sci. 2023, 24(9), 7884

Abstract

Motivation

Many essential cellular functions are carried out by multi-protein complexes that can be characterized by their protein-protein interactions. The interactions between protein subunits are critically dependent on the strengths of their interactions and their cellular abundances, both of which span orders of magnitude. Despite many efforts devoted to the global discovery of protein complexes, by integrating large-scale protein abundance and interaction stoichiometries, there is still room for improvement.

Results

Here, we integrated >7,000 quantitative proteomic profiles with 3 published affinity purification/co-fractionation mass spectrometry datasets into a deep learning framework to predict protein-protein interactions (PPIs), followed by the identification of protein complexes using a two-stage clustering strategy. Our deep learning technique-based classifier significantly outperformed recently published machine learning prediction models and captured in the process 5,010 complexes containing over 9,000 unique proteins. The vast majority of proteins in our predicted complexes exhibit low or no tissue specificity, which is an indication that the observed complexes tend to be ubiquitously expressed throughout all cell types and tissues. Interestingly, our combined approach has increased the model sensitivity for low abundant proteins, which amongst others allowed us to detect the interaction of MCM10 that connects to the replicative helicase complex via the MCM6 protein. The integration of protein abundances and their interaction stoichiometries using a deep learning approach provided a comprehensive map of protein-protein interactions and a unique perspective on possible novel protein complexes.

Availability

The source code is available at: https://github.com/Bohui2447/ProteinComplex1.

1 Introduction

Protein complexes are multi-protein assemblies that play a crucial role in diverse biological processes, including the control of cellular homeostasis, growth, and proliferation (1). For example, the 26S proteasome, which consists of 31 different subunits, is essential in controlling the cell cycle, cell growth, and apoptosis by degrading obsolete or damaged proteins (2). Elucidating the components and functions of multi-protein complexes is fundamental to understanding cellular processes. Despite tremendous efforts (3-5), it remains a daunting task to identify exactly which human proteins are present in protein complexes on a proteome-wide scale.

To identify protein complexes from protein-protein interactions, several experimental technologies are employed. For instance, yeast two-hybrid assays, which depend on bringing the DNA-binding domain (BD) and transcription activation domain (AD) of a eukaryotic transcription factor in close proximity by a bait-BD fusion protein and a pray-AD fusion protein thereby, enabling identification of protein interactions and protein complexes (6, 7). High-throughput experimental techniques, such as affinity purification-mass spectrometry (AP-MS) (8, 9) and co-fractionation mass spectrometry (CF-MS) (1, 8) have enabled large-scale characterization of protein interactions. The AP-MS approach depends on the expression of a bait protein that is coupled to a matrix, allowing purification of the target proteins (preys) that interact with the bait from a lysate (10). In the CF-MS approach, cellular lysates are extensively fractionated by multiple, non-denaturing biochemical methods which allow for the identification of protein complexes that co-elute (1). Subsequently, a PPI network is represented by the co-elution network, and protein complexes are inferred using correlations of the protein elution profiles (1, 11). These high-throughput techniques have established the identification of large-scale protein interaction networks in humans and other model organisms, dramatically increasing the coverage of the PPI network.

In the past few years, two large-scale studies (BioPlex (5) and Hein et al. (8)) using the AP-MS approach, and one large-scale study by (Wan et al. (12)) using CF-MS have significantly improved the understanding of human PPI networks. However, the interactions identified by these different studies show only limited overlap (13). One possible explanation may be that different experimental methods detect different types of interactions, thereby reporting different subsets of the actual PPI network (14). Thus, Drew et al. integrated these datasets using a support vector machine (SVM) classifier to build a PPI network and ultimately obtained a global map of human protein complexes (13). Besides these largescale studies, many more protein-protein interaction datasets have been deposited into public repositories, such as BioGRID (15), BioPlex (16), and STRING (17). This allows researchers to combine and integrate public datasets using in sillico, e.g. computational approaches.

Proteins in a complex are typically expressed and localized in a spatiotemporal-similar manner, meaning that these proteins are often found in near cellular vicinity simultaneously, and possess similar biological functions (18). Another predictor for protein interactions is to look at co-translation (2). For instance, Shieh et al. showed that proteins LuxA and LuxB are co-translated and assembled into the luciferase enzyme complex in Escherichia coli (19). In addition, studies employing gene co-expression analyses have revealed that the network modules in a co-expression network are related to protein complexes. Examples of these complexes are the spliceosome, ribosome, and RNA polymerase II (20, 21). Besides, Bork and colleagues have constructed the STRING database (17), which incorporates data from multiple sources, including information on protein co-expression, text-mining, and experimental data. This multi-level approach provides a system-wide view of protein-protein interactions (17, 22, 23), thereby showing the strength of data integration in the prediction of PPIs.

Despite many efforts have been devoted to quantifying and classifying the protein complexes, approaches by integrating large-scale protein abundance and interaction features are needed to improve. In this study, the integration of large-scale protein quantification data from multiple human cell samples was combined with AP-MS and CF-MS data to improve the construction of the human PPI network. We have constructed a comprehensive map of human protein complexes via integrating protein interaction and protein abundance stoichiometry features. Briefly, the protein interaction stoichiometry features were obtained from three high-throughput AP-MS/CF-MS datasets (5, 8, 12), comprising 258 parameters describing different protein-protein interaction properties. The protein abundance stoichiometry features were derived from >7000 label-free human protein quantification profiles from the PRoteomics IDEntifications (PRIDE) database (https://www.ebi.ac.uk/pride/). Subsequently, a deep learning (DL) model was built by using these features as input, and ultimately to infer an integrated protein interaction network. Next, a two-step unsupervised clustering procedure was performed to obtain a comprehensive map of human protein complexes. Our approach resulted in a comprehensive overview of protein complexes that also contain low-abundant and poorly characterized proteins, thereby providing a unique perspective on the human interactome.

2 Materials and Methods

2.1 Gold-standard reference set and the training and test protein pairs

A fundamental step in predicting protein complexes is the prediction of protein-protein interactions which is considered a classification task in machine learning and requires a gold-standard reference set comprising a positive and a negative subset. The positive subset is defined by the set of protein pairs that are within the same complex. In contrast, the negative subset is defined by the set of protein pairs within the entire set of protein complexes but that are not in the same complex.

The human protein complexes in the CORUM database (24) form a high confidence set of manually curated protein complexes and therefore can be considered as a gold standard reference set in this study. The training and test sets that contain the gene names of protein pairs were downloaded from the hu.MAP database (13). These training and test protein pairs were generated as described by Drew et al. (13) and were derived from the CORUM database. Briefly, a set of non-redundant complexes were retained by merging the complexes with a large overlap (i.e., Jaccard coefficient >0.6) within the entire CORUM database. This non-redundant dataset was randomly divided into two sets, i.e., a training set and a test set. To not skew the measurements of the performance of our model in subsequent classification steps, complexes with larger than 30 subunits in the test set were removed. The positive and negative subsets of protein interactions were generated for both training and test sets, followed by removing the interactions from the training set that overlapped with those in the test set. The final training set contained 14,186 and 95,802 protein-protein pairs in the positive and negative subsets, respectively. The test set contained 5,781 and 111,055 protein-protein pairs in the positive and negative subsets, respectively.

2.2 Featurization of protein-protein interaction pairs

2.2.1 Protein abundance features

We retrieved 246 independent projects from the PRIDE repository (Table S1), containing a total of 7,330 MS/MS-based proteomics quantification profiles, as MaxQuant outputs, which is the protein abundance stoichiometry data in this study (25). For each protein abundance dataset, reliable proteins were retained by removing potential contaminants and removing proteins that were identified with less than 2 peptides. In addition, the raw intensity of each protein as reported by MaxQuant was maintained, and the raw intensities of each protein were averaged if a protein was reported multiple times. Subsequently, the intensities of proteins were Log10-transformed, resulting in a matrix (M) containing expression data of 17,951 proteins originating from 7,330 different protein abundance stoichiometry profiles.

The protein abundance stoichiometry features ($D_{i,j}$) of a protein pair (proteins i and j) were calculated by

$$D_{i,j} = |M_{i,.} - M_{j,.}| \qquad (1)$$

where $M_{i,.}$ and $M_{j,.}$ are the rows of matrix M, which correspond to the abundance of protein i and protein j across all profiles, respectively. The protein abundance feature matrix was calculated for all protein pairs in the training and test sets, generating a 109,988 (protein pairs) x 7,330 protein abundance feature matrix for the training set and a 116,836 (protein pairs) x 7,330 protein abundance feature matrix for the test sets.

2.2.2 Protein interaction features

For each protein pair in the training and test sets, the AP-MS/CF-MS features comprised 258 features (Table S1) that were generated by integrating over 9,000 mass spectrometry experiments from three published papers (Wan et al (12), BioPlex (5, 26), and Hein et al (8)) (13), which were downloaded from the hu.MAP database (termed protein interaction stoichiometry data in this study) (27). More specifically, these features were collected from the following 6 resources: (1) 220 co-fractionation features, i.e., 4 types of co-fractionation measures (Poisson noise Pearson correlation coefficient, a weighted cross-correlation, a co-apex score, and a MS1 ion intensity distance metric) for each of the 55 fractions in Wan et al (12); (2) nineteen genomic/proteomic/literature features of worm, fly, human, and yeast from HumanNet (28), such as genetic interactions, results of high-throughput yeast 2-hybrid assays, co-citation of genes, et al; (3) two features that describe protein interactions obtained from AP-MS experiments in fruit fly (“ext_Dm_guru”, (29)) and human (“ext_Hs_malo” (30)); (4) Nine features from the BioPlex database, being NWD score, Z score, plate Z score, entropy, unique peptide bins, ratio, total PSMs, ratio total PSMs, and unique to total peptide ratio; (5) Four features from Hein et al, being the Pearson’s correlation coefficient of the intensity profiles of the prey and bait proteins (“prey.bait.correlation”), the number of available quantitative data of the prey (“valid.values”), the log10-transformed stoichiometry of prey to bait protein in the pulldown samples (“log10.prey.bait.ratio”), and the log10-transformed stoichiometry of prey to bait protein in the HeLa proteome samples (“log10.prey.bait.expression.ratio”) (31); and (6) four features (i.e., “neg_ln_pval”, “pair_count”, “hein_neg_ln_pval”, and “hein_pair_count”) generated based on Drew et al’s weighted matrix model interpretation (13) of the AP-MS datasets in BioPlex (5) and Hein et al (31). This results in a 109,988 (14,186 + 95,802) × 258 protein interaction feature matrix and a 116,836 (5,781 + 111,055) × 258 protein interaction feature matrix for training and test sets, respectively. (More details of protein interaction features are in the supplementary materials)

2.3 Deep learning neural network implementation

The neural network model was implemented by using the R interface to Keras (version number: 2.2.5.0), which is a high-level neural network API (32). Our model consists of three densely connected hidden layers with different numbers of neurons and the output layer is aimed at predicting PPIs (Fig. 1). The rectified linear unit (ReLU) activation functions were used for all hidden layers. The sigmoid activation function was applied to the output layer. For each of the hidden layers, a dropout layer was appended to avoid overfitting. The training process was performed for 10 epochs using the ‘RMSProp’ (33) optimizer with binary cross-entropy as the loss function. The optimal combination of 6 hyper-parameters (the number of neurons in each hidden layer and dropout rate in each dropout layer) were tuned by random searching. Briefly, the number of neurons for each hidden layer was generated randomly, ranging from 10 to 750, and the probabilities for the dropout layers followed a uniform distribution over the interval of 0 to 0.5 (Table S2). We have applied this training process on three different feature matrices; i) protein pairs containing only the protein abundance stoichiometry features, ii) protein pairs with only the protein interaction stoichiometry features, and iii) protein pairs that have integrated both abundance and interaction stoichiometry features. Subsequently, the F1-measure (F1-measure = 2*(precision * recall)/(precision + recall)) that represents the harmonic means of precision and recall of the prediction by these models, was used to compare the model performance and select the best model. Finally, the best model was applied to all protein pairs with protein abundance and interaction stoichiometry features to generate the weighted PPI network, in which nodes were proteins, and the weight of the edge was the protein-protein interaction probability predicted by the best model.

To further evaluate the performance of our model, we used the interaction stoichiometry feature matrix as input to train the SVM classifiers. The SVM implementation of the R package ‘e1071’ (version 1.7.2) (34), which is based on the LIBSVM library (35), was applied with function tune.svm. To seek an optimal model, we performed a parameter tuning of the hyperparameters (C and gamma) for the SVM model training using 10-fold cross-validation by tune.control function (parameters were detailed in Table S2). The performance of the SVM models was subsequently evaluated by comparing the F1-measure.

2.4 Evaluation of feature importance

After training the deep learning model, the importance of features in the model is evaluated by the decrease of model performance when randomly shuffling the values of each feature. Firstly, the best deep learning model was applied to the test dataset (T_0) to make the prediction and to calculate F1-measure ($F1_0$). Secondly, only the values of the i_{th} feature in the test dataset were randomly shuffled generating a sudo-test dataset (T_i), which was fed to the model to make prediction and to calculate F1-measure ($F1_i$). Thirdly, repeat the second step N (N = 50) times, and calculate the mean value of F1-measure for the i_{th} feature $\frac{1}{N}\sum_{j}^{N} F1_{i,j}$. Lastly, the importance of i_{th} feature (I_i) is evaluated by

$$Ii = F1_0 - \frac{1}{N}\sum_{j}^{N} F1_{i,j} \qquad (2)$$

2.5 Two-stage clustering to predict protein complexes

The weighted protein interaction network as generated by the final deep learning model using both protein abundance and interaction stoichiometry features was used to derive protein complexes through a two-stage clustering approach. The first clustering method, ClusterONE (Clustering with Overlapping Neighborhood Expansion), is a graph clustering algorithm (36), which starts from a single seed vertex and exploits a greedy procedure that adds or removes vertices to find clusters with high cohesiveness. The parameter “density” was set to determine the complex density. The “overlap” specifies the maximum allowed overlap between two clusters, which determines whether to merge or not merge highly overlapping complexes. The second clustering method is MCL (Markov Cluster Algorithm) (37). This unsupervised cluster algorithm is based on stochastic simulation of flow in networks/graphs and is controlled by the inflation (-I) parameter. Inflation affects the granularity or resolution of the clustering outcome, where low values lead to fewer and larger clusters, and high values lead to more and smaller clusters.

We first sorted the edges in descending order by their weights that were predicted by the deep learning model, resulting in a subnetwork with the top r percent of edges. Here, r (ranging from 1 to 20) is a tuning parameter that needed to be optimized to obtain the best set of complexes in the following steps. In the ClusterOne clustering step (36), a seed method of “nodes” and a minimum size of 2 were applied to each subnetwork ($r = r_i$) to generate a set of intermediate clusters. Here the parameters for the ClusterOne algorithm “density” were tuned in the range of [0.2, 0.25, 0.3, 0.35, 0.4], and “overlap” was tuned in the range of [0.6, 0.7, 0.8]. Since we allowed merging high-overlapping clusters in the ClusterOne process, this may lead to large clusters that are over-merged, i.e., biologically unrelated complexes merged into a single large cluster (12). Therefore, a second clustering stage, based on the Markov Cluster (MCL) algorithm was performed on each cluster generated by ClusterOne to split the over-merged clusters. Here, the parameter inflation (-I) of the MCL algorithm (37) was tuned in the range of [1.2, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 15]. Proteins that did not share any edge with the remaining proteins in the final clusters were removed. This two-stage clustering process was carried out for each combination of parameters, i.e., r, density, overlap, and inflation, followed by a k-clique evaluation (the parameter combinations were detailed in Table S3).

2.6 K-clique method-based accuracy evaluation

To measure the accuracy of the reconstructed complexes, we used the k-clique algorithm for each of the two-stage-clustering results. As described above (13), this approach is based on the matching of cliques within the set of all possible cliques between reconstructed or predicted complexes and benchmark (golden dataset) complexes (here the CORUM complexes). Specifically, the predicted complexes and CORUM complexes were first divided into different subsets according to their clique size k (e.g., k = 2, all pairwise combinations; k = 3, all triplet combinations; etc.). Secondly, we removed the predicted complexes in which all protein members are not in the gold standard set. In other words, we only evaluate the complexes containing proteins that form known complexes to not penalize novel predicted complexes as false positives. Thirdly, for each clique size k, the true positive (TP_k) was defined by the number of common complexes between predicted complex set and gold standard complex set; the false positive (FP_k) was the number of complexes in the predicted complex set but not in the gold standard complex set; the false negative (FN_k) was the number of complexes in the gold standard complex set but not in the predicted complex set. Subsequently, the precision (P_k), recall (R_k) and F-measure (F_k) were defined as follow:

$$P_k = \frac{TP_k}{TP_k + FP_k} \qquad (3)$$

$$R_k = \frac{TP_k}{TP_k + FN_k} \qquad (4)$$

$$F_k = \frac{2 * P_k * R_k}{P_k + R_k} \qquad (5)$$

Finally, a global F-measure (F-Grand, equation 4) was defined as the mean of F_k, iterating over clique sizes of k from 2 to K, where K is the largest cluster size of the predicted complexes set.

$$F_{grand} = \frac{\sum_{K=2}^{K} F_k}{K - 1} \qquad (6)$$

2.7 Enrichment analysis and tissue specificity

We used the g:Profiler web tool (38) to perform protein and pathway enrichment analysis for each predicted complex, with significantly enriched terms (Benjamini-Hochberg FDR < 0.05). For comparing tissue specificity, we mapped our predicted complexes to the tissue-based map of the human proteome from the Human Protein Atlas (39, 40).

3 Results

3.1 Feature matrices constructed by incorporating protein abundance and interaction stoichiometries datasets

In this study, we integrated two recently published AP-MS protein interaction datasets from BioPlex and Hein et al., and one CF-MS protein interaction dataset [5, 8, 12). As shown in Fig. 1, we obtained 241 features from Wan et al.' (12) CF-MS (co-fractionation mass spectrometry) analysis of human proteins and their orthologues, comprising 6,387 fractional MS experiments. Nine affinity purification mass spectrometry (AP-MS) features and two features generated by Drew and collaborators (13) were collected from BioPlex (Version 1) (5), which encompasses 2,594 AP-MS experiments containing over 50,000 interactions from HEK239T cells. Four AP-MS features describing 28,504 interactions were obtained from Hein and colleagues (8).

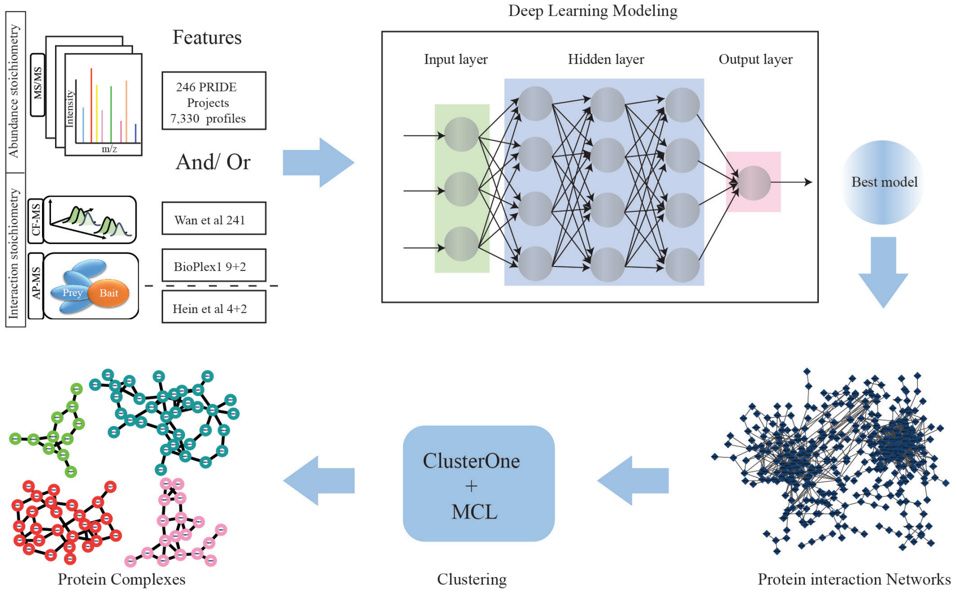


Figure 1. Flowchart for protein complex discovery. Schematic workflow for the discovery of protein complexes by employing feature selection and deep learning algorithms. 7,330 protein abundance samples (MS/MS profiles) and three protein interaction datasets (Wan et al, BioPlex, and Hein et al that contains 258 interaction features) were used as input to train the deep learning (DL) models, the optimal DL model was applied to infer protein-protein interaction scores and ultimately generated a weighted protein interaction network. Two unsupervised clustering algorithms, i.e., ClusterOne and MCL, were subsequently applied to obtain the final protein complexes dataset.

In PPI studies, researchers expect to retrieve subunits of complexes in equimolar amounts after immunoprecipitation (IP) from biological experiments. However, in practice, the range of detected interacting protein abundances spans several orders of magnitude (41). This is caused by the possible involvement of some protein subunits in multiple different complexes with fractions of their total cellular pools, and subunits may behave differently under different states (different tissue or disease states). To reduce the bias caused by the huge span of protein abundances in protein complex identification, we incorporate >7,000 protein abundance profiles from the PRIDE-archive (42). Precisely, the protein abundance dataset consists of 246 quantitative proteomics projects, consisting of 7,330 profiles (Fig. 1). The number of unique proteins detected in each profile ranges from 500 to over 8,500 (Fig. 2A). In total, we incorporated 17,951 proteins from protein abundance profiles, which covers more than 98% of the proteins quantified in interaction datasets (Fig. 2B). Moreover, the protein abundance samples were distributed over 25 different human tissues and organs, indicating a large sample diversity in our dataset (Fig. 2C).

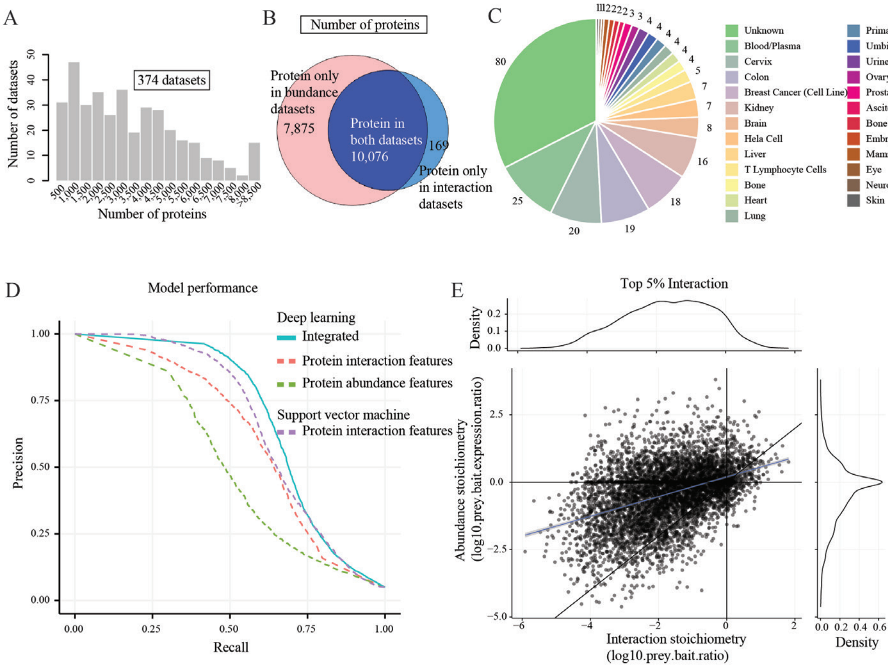


Figure 2. The Integration of protein abundance and interaction stoichiometry substantially improves model performance. A, Histogram plot showing the distribution of the number of proteins in the 374 proteomics datasets. The largest datasets contained over 8500 proteins whereas the smallest datasets only contained about 500 proteins. B, A total of 17,951 and 10,245 proteins were collected from protein abundance profiles and the protein interaction datasets separately, in which over 98% of proteins (10,076 proteins) were observed in both datasets. The pink area represents the number of proteins only in protein abundance datasets, the light blue area represents the number of proteins only in protein interaction datasets, and the dark blue area represents the number of proteins in both datasets. C, A pie chart showing the distribution of sample tissue specificities for the protein abundance profiles. This plot shows that the protein abundance profiles were distributed over more than 25 different tissues or organs, indicating a large sample diversity which in turn improves the robustness of the deep learning model. Different colors indicate the organs, the numbers in the pie chart are the numbers of datasets that are collected in the organs. For those datasets that do not show organ information in the PRIDE database are labeled “unknown.” D, A comparison of Model performance for the deep learning and SVM models based on different data sources. The precision is calculated by true positive / (true positive + false positive), and the recall is calculated by true positive / (true positive + false negative). The harmonic mean of precision and recall, namely F1-measure or F1-score was further used to determine the model performance. The integration of both abundance and interaction features (blue line) outperforms all other single feature based models (dashed lines). E, A scatterplot showing the top 5% of protein interactions. From this plot it can be observed that the predicted protein-protein interactions were greatly overlapping with the Hein et al.’ interaction network and exhibiting similar stoichiometry distributions.

3.2 Model performance comparison

Having established the feature matrices, we next generated the training set and test set by labeling protein pairs based on a gold-standard literature-curated set of human protein complexes: CORUM (43). The positively labeled protein-protein interactions (PPIs) are proteins within the same complex in the CORUM database. The negative protein pairs are those that are observed in the gold standard set but which do not interact with subunits in the CORUM complexes. Protein pairs which were not included in the training process were labeled as ‘unknown’. Next, we implemented a deep learning neuronal network to train three types of models: i) protein abundance feature matrix only, ii) protein interaction feature matrix only, iii) an integrated protein abundance and protein interaction feature matrix (Fig. 2D). Moreover, to compare the performance of our models, we also build an SVM model using the protein interaction feature matrix (44, 45).

To obtain an optimal classifier, we trained our DL models by varying the number of neurons in three densely connected layers and the probabilities in dropout layers (details in Table S2). This training process resulted in 1995 protein interaction feature-based models, where 63 models’ F1-measure > 0.59 (Fig. S1A and Table S2A); 1921 protein abundance feature-based models, where 87 models’ F1-measure > 0.49 (Fig. S1B and Table S2B); and 2338 integrated models (integrated protein abundance and interaction features), with 109 models’ F1-measure > 0.66 (Fig. S1C and Table S2C) (see methods for details on the F1 measure). Moreover, 28 SVM classification models, based on protein interaction features, were obtained using a grid search algorithm (See methods). The precision-recall curve for the best models using the different feature matrices shows that the integrated deep learning model (F1-measure = 0.68) outperformed all other models (F1-measure of protein abundance-DL, protein interaction-DL, and protein interaction-SVM models are 0.51, 0.61, and 0.64, respectively) (Fig. 2D).

The optimal deep learning model contains 350, 140, and 25 neurons in three hidden layers, where dropout rates are 0.438, 0.214, and 0.037, respectively. It takes around 1.8 hours to train the model. This model was further applied to predict the interaction score for all protein pairs characterized in the feature matrix. The optimal model takes ~ 80 seconds to make prediction for 10,000 protein pairs. The interaction score of a protein pair indicates the likelihood of that pair of proteins participating in the same complex. Subsequently, a weighted PPI network was generated, where the weights of edges were defined by the predicted interaction score (Fig. 1). To assess the predicted PPI network, we further compared it with the network generated by Hein et al. (Fig. 2E and Fig. S2A-S2C). Notably, the network formed by the top 5% of predicted interactions showed similar stoichiometries to the Hein et al. network (Fig. 2E) (8). In addition, a weaker interaction stoichiometry was observed when the network was filtered by decreasing the protein interaction confidence, suggesting that an interaction filtering step is required to obtain an optimal PPI network to infer protein complexes (Fig. S2A-S2C).

3. 3 Protein complexes identified by two-stage clustering method

To elucidate the relationships among densely connected regions of the interaction network, a two-stage clustering was employed (13). At the first stage of clustering, the ClusterOne algorithm (36) was employed to derive the intermediate clusters. Due to the appearance of over-merged clusters (merging high-overlapping clusters may lead to biologically unrelated complexes being merged) (12), we applied a second stage of clustering that is based on MCL (37) (see methods) to further break over-merged clusters produced by ClusterOne. To optimize the clustering performance, as described in the methods, we tuned the parameters including the top percentage of interaction edges r, density and overlap parameters in ClusterOne, and inflation in MCL. A set of protein complexes resulting from each combination of parameters was compared to the gold standard CORUM complex set by the k-clique algorithm (13), enabling the evaluation of their similarity and overlap to a benchmark complex set (here the CORUM complexes) on a global level. This two-stage clustering step generated 7605 datasets containing complexes, with their corresponding similarity measurements (F-grand values) as defined by the k-clique algorithm (Fig. 3A). The best parameter combination was edges r: 119,560; density: 0.2; overlap: 0.8; inflation (-I): 5, which resulted an F-grand at 0.46. The optimal set contains 5,010 complexes with 101,818 interactions (5% interaction edges of the full network) among 9,129 human proteins (Table S4). Additionally, in line with the finding by Huttlin et al. (5), a vast majority of complexes contain a limited number of protein members (Fig. 3B).

If a group of proteins can form a protein complex, we assumed that their expression may show a concordance, which can be evaluated by the Manhattan distance of their stoichiometry abundances. To further evaluate the quality of the final predicted complexes, we first evaluated the expression concordance of protein members in complexes by calculating the pair wise Manhattandistance of the proteins using the abundance stoichiometry (label-free quantification) data. Subsequently, we randomly shuffled the protein members amongst the other complexes while maintaining the same complex size, and then calculated the Manhattan distance of proteins within these randomly generated complexes. This shuffling process was repeated 100 times. As indicated in Fig. 3C, our final set displays a shorter Manhattan distance within the complex than the shuffled complex set. Additionally, we annotated proteins with information on their expression in different tissues using the Human Protein Atlas (39). We observed that a considerably high percentage of proteins in our complexes showed a low tissue specificity. For instance, the protein complex 76 (Table S4 line 76) that we predicted was reported as the proteasome (46), the predicted complex 55 (Table S4 line 55) is known as the mediator complex (47), indicating that our complex set can capture many common fundamental processes in human cells (Fig. 3D) (8).

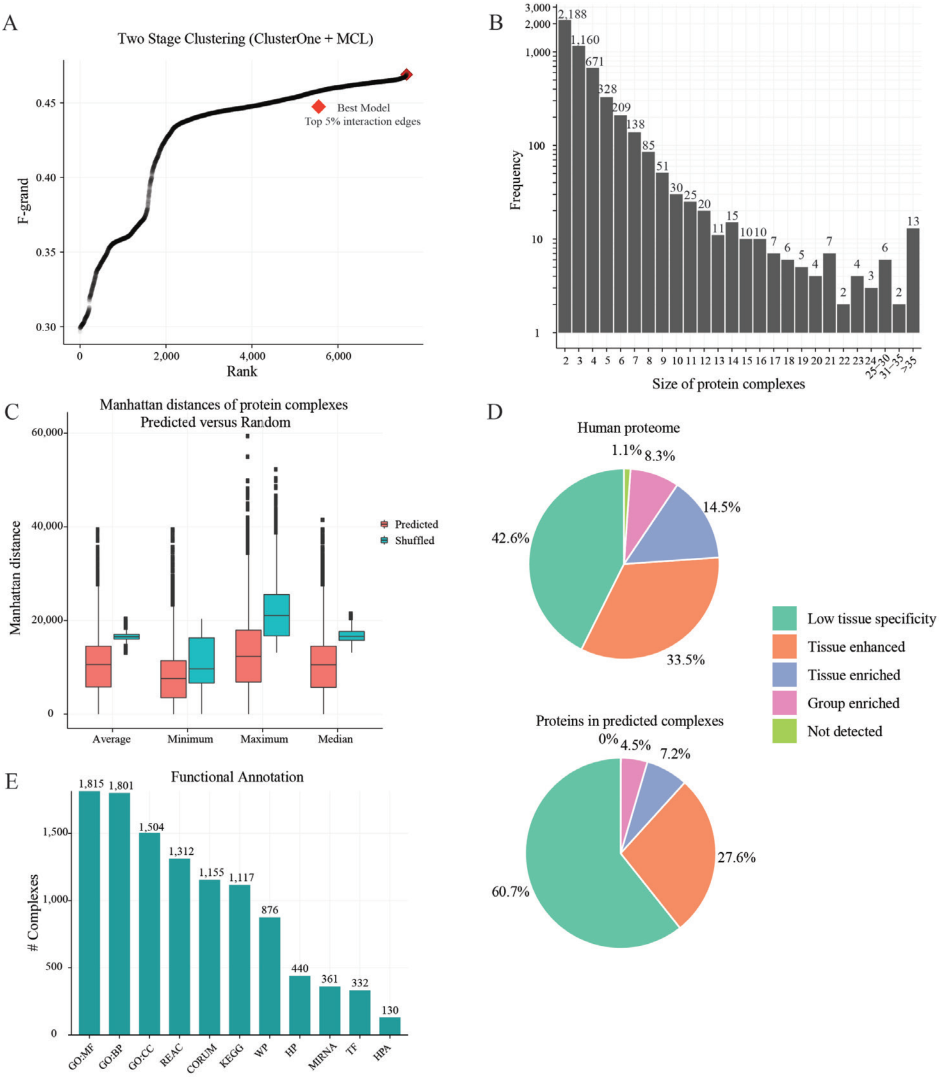


Figure 3. Biological features of predicted human protein complexes. A, Parameter optimization for two-stage clustering (ClusterOne followed MCL) procedures. Each data point indicates one F-Grand measure generated in the clustering step. This two-stage clustering step generates 7605 results with 3187 F-grand (points) over 0.45, indicating the high stability of the protein-protein interaction network. B, Distribution of protein complex sizes in the final interaction map, the vast majority of protein complexes contain a small number of protein members. C, Boxplots showing the average, minimum, maximum and median of the protein complexes' Manhattan distance as calculated based on the abundance stoichiometry of the protein complex subunits. The shuffled protein complex distance (blue) was evaluated by permuting protein members while maintaining the sizes of the protein complexes. It can be seen that the predicted complexes display a shorter Manhattan distance than the shuffled complexes, indicating the credibility of predicted protein-protein interactions. D, Pie charts showing the proportions of proteins with varying tissue expression patterns from the Human Protein Atlas. From this plot 60.7% of proteins in our complexes showed a low tissue specificity, indicating the ubiquitous expression property of the proteins. E, The distribution of number of protein complexes with significantly enriched annotation terms using g:Profiler web tool. Most complexes could be enriched in one or more categories with significant terms, indicating the biological significance of complexes.

Moreover, the functional annotation analyses of the protein complexes showed that a large proportion of predicted complexes could be enriched in functional terms (Fig. 3E). For example, around 36% of the predicted complexes were significantly enriched in GO molecular functions and GO biological processes.

3.4 Protein abundance stoichiometry contributes to capturing novel subunits

Our final dataset achieved a high model performance based on the F1-measure (0.68) and k-clique evaluation (F-grand = 0.46). In addition, approximately 15% (737 of 5,010) of our complexes exhibited a complete or partial overlap with 42% (1,100 of 2,597) of the gold-standard complexes from the CORUM database (Fig. 4A) (8). This high confidence allows us to predict novel interactions on top of known PPIs. For instance, we predict MCM10 as a novel member of the MCM2-7 complex via interacting with MCM6 (Fig. 4B). Furthermore, the protein abundance stoichiometry shows that the core subunits of MCM2-7 complex exhibit a considerably high expression concordance in most of the 7,330 profiles we obtained from Pride (Fig. 4C). Interestingly, MCM10 is detected in less samples, indicating its lower abundance or poor characterization potential by MS-based techniques. Based on these observations, we asked whether the MCM6-MCM10 interaction is detectable in AP-MS experiments. Indeed, the interaction stoichiometry data from Hein et al. (8) (Fig. 4B, bottom) showed that the MCM6-MCM10 interaction was detectable however with a relatively low stoichiometry, suggesting that MCM10 may be a non-obligatory or transient member of the MCM2-7 complex. In addition, Homesley et al. (48) and Douglas et al. (49) reported that MCM10 was required for the initiation of eukaryotic DNA replication and physically interacts with MCM2-7 via subunit MCM6.

3.5 Members of protein complexes exhibit co-expression characteristic

Co-expression characteristics are of biological interest since co-expressed genes usually are controlled by the same transcriptional regulatory program, functionally related, or members of the same protein complex (50). Proteins that are part of the same protein complex often show co-expression properties and clusters of proteins with related functions often exhibit expression patterns that correlate under diverse conditions. For instance, importin-7 (IPO7) and importin beta-1 (KPNB1) are two important proteins for nuclear protein import (51). These two proteins are highly co-expressed in a majority (~ 5,800) of the abundance stoichiometry profiles (Fig. 4E). Moreover, the interaction between these two proteins was also detected in Hein et al.’s interaction network as a stable interaction (Fig. 4D). Jakel et al. (52) reported that importin-7 (IPO7) and importin beta-1 (KPNB1) work as a heterodimer that binds to histone H1. (More examples showing co-expression properties are shown in the supplementary materials)

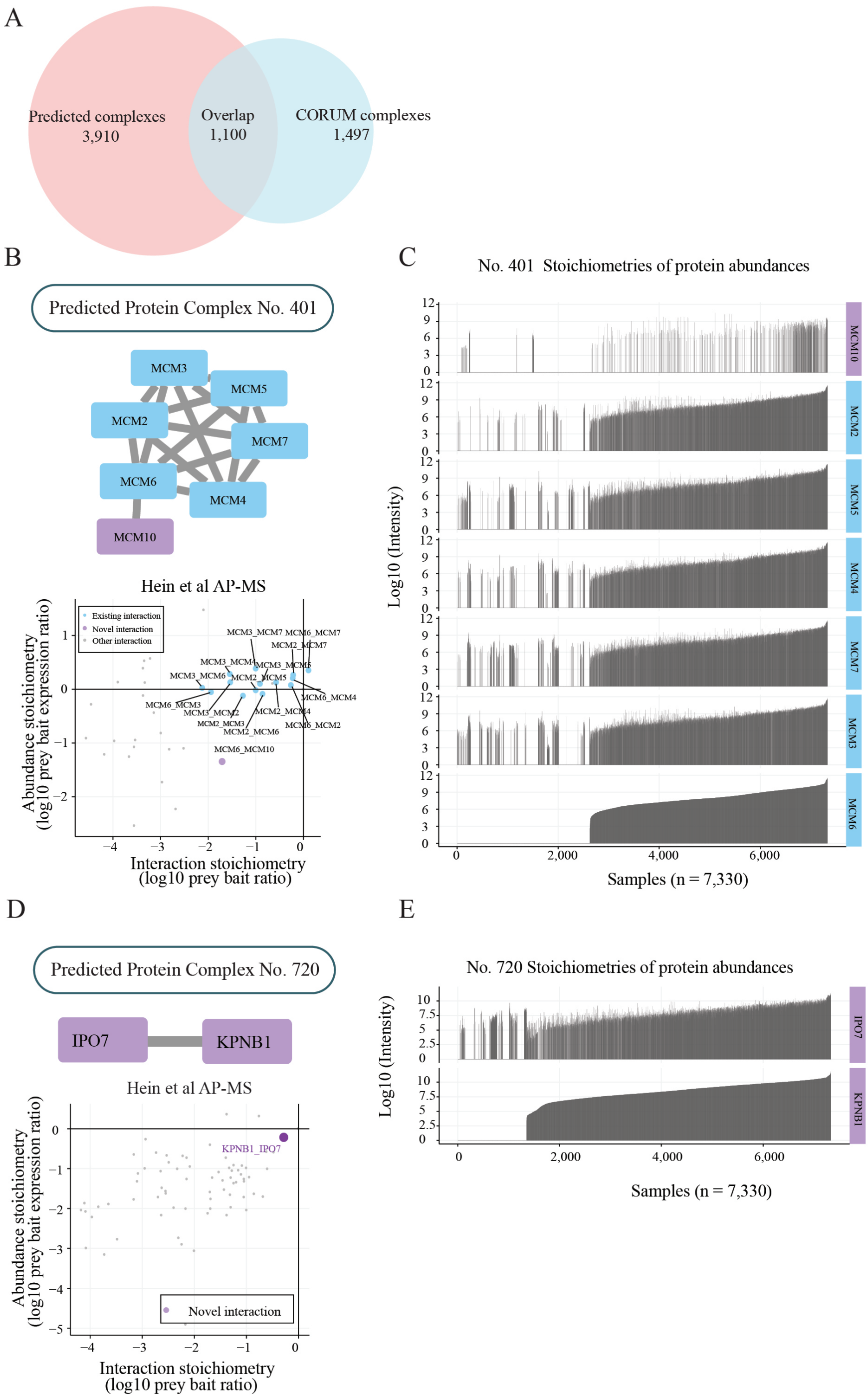


Figure 4. Selected complexes in the map contain novel subunits. A, Venn plot indicating the overlap between the protein complexes predicted by our model (pink circle) and the complexes in CORUM database (blue circle). 1,100 out of 5,010 predicted complexes exhibited a complete or partial overlap

with the gold-standard protein complexes from the CORUM database, showing the potential to predict novel protein-protein interactions. B, top panel; Interaction network of replicative helicase; blue rectangles are known members of the MCM complex; the purple rectangles are novel subunits as predicted by our deep learning model. B bottom panel; Scatter plot with interaction and abundance stoichiometry for the MCM complex from Hein et al' AP-MS experiments. Blue dots are known interactions, the purple dots are novel interactions. Labels for the dots are represented by Bait Prey proteins. It can be observed from the scatterplot that the MCM6-MCM10 interaction follows a similar trend as the other known interactors, indicating that MCM10 could be a transient member of the MCM complex. C, The expression pattern of each subunit within the MCM protein complex. On each row, the X-axis indicates 7,330 profiles collected from PRIDE repository, the Y-axis indicates the log10 transformed intensity of the protein, where missing values are in blanks. It can be observed that MCM10 is detected in fewer samples as compared to the other subunits of the MCM complex, indicating a lower abundance or poor characterization potential by MS-based techniques. D, Top panel; Interaction network of the new complex IPO7-KPNB1. D bottom panel; Interaction-abundance stoichiometry for IPO7-KPNB1complex using Hein et al' AP-MS interaction data. This novel interaction was not observed in the Hein et al' AP-MS interaction network data, demonstrating the sensitivity of our deep learning model. E, The expression pattern of each subunit within the IPO7-KPNB1protein complex. The X-axis indicates 7,330 samples collected from PRIDE repository. The Y-axis indicates the log10 transformed intensity of protein, missing values are in blanks. IPO7 and KPNB1 show significant co-expression in a majority (~ 5,800) of the abundance stoichiometry profiles, indicating a possible protein-protein interaction between IPO7 and KPNB1

4 Discussion

Many vital cellular functions, including DNA replication, RNA transcription, and protein translation and regulation, require the coordination of proteins assembled into complexes. Thus, the analysis of protein complexes and PPI networks are of central importance in biological research. In the past decades, the combination of affinity purification/co-fractionation and mass spectrometry has advanced our understanding of protein complex composition. Increasing efforts have been devoted to generating larger-scale human protein interactions by integrating different AP-MS and CF-MS studies and have established more comprehensive maps of protein complexes (13). Although these protein-protein interaction experiments are very well controlled studies, they are typically performed on certain cell lines/types and may overlook the proteomic abundance differences in human tissues. Here, we present a data integration method, using machine learning and classification algorithms to create a comprehensive map of protein complexes, by integrating protein interaction stoichiometries and large-scale protein abundance stoichiometry profiles.

In this work we developed a deep learning framework that incorporates multiple sources of data to establish a comprehensive human protein complex map. Our results show that a deep learning-based approach, by incorporating multiple sources of features (AP-MS/CF-MS interaction features and MS/MS protein abundance features), outperformed models using either interaction features, or abundance features alone. Besides, these integrated deep learning models exhibit high robustness, not only on the F1-measure but also on the number of outperformed models (section 3.2). We also showed that many complexes, including gold standard and novel complexes, feature a unique characteristic of co-expression patterns in a majority of quantitative proteomics samples. This characteristic enabled us to recapitulate several well-known complexes, for instance the multi-synthetase complex (53) and eukaryotic initiation factor 2B complex (24) (Fig. S4). Moreover, this characteristic also led us to discover highly co-expressed complexes, such as the IPO7-KPNB1 heterodimer complex (Fig. 4D) and the VCP-HSPB90B1 complex (Fig. S3E). These examples indicate that the expression levels of protein complex subunits are generally co-varying (18). Thus, such co-varying characteristic can be used as one of the features for identifying protein-protein interactions and protein complexes. In contrast to other published methods, we did not summarize the concordance of protein expression between proteins as the correlation coefficient, as this may over-simplify the complexity within a large dataset. Instead, we first calculated the expression difference within each protein pair among all 7,330 protein abundance profiles and subsequently used a deep learning algorithm to achieve a high-level featurization after training our model. Here, the state-of-art deep learning algorithm addresses this featurization by computing increasingly more complex features and then taking the results of preceding operations as input (54). Therefore, our model makes full use of not only the protein interaction stoichiometries but also the protein abundance stoichiometry with tissue/sample-level details.

In addition to model performance, the contribution of features is an important aspect of deep learning. Thus, we performed feature importance evaluation by the decrease of model performance using randomly shuffling the values of each feature (see Methods). As expected, the top-ranked features are the interaction features. For example, all of the top 15 features are interaction features, including “hein_neg_ln_pval”, “neg_ln_pval”, “hein_pair_count”, and “prey-bait correlation” (Fig. S5B), which are the most important outcomes in AM-MS experiments. We also found that, within the protein abundance features, the feature importance is positively correlated with the number of proteins (Fig. S5B). In other words, if more proteins are identified in a protein abundance profile, a higher importance that feature (i.e., protein abundance profile) shows. This suggests that the number of proteins in a protein abundance profile could be one of the criteria to improve the quality of the data in future works.

The weak interactions have frequently been overlooked or remained undetected and they have been thought to be less important in large-scale interaction research, even though they are crucial features of networks in general (55, 56). In addition, weaker interactions with low abundant proteins are challenging to detect in AP-MS experiments (8). To detect low-abundant proteins and characterize weak interactions, one possible strategy is to improve the sensitivity and resolution of the mass spectrometer or to remove high-abundant proteins from proteomic samples (57). Another strategy is to increase data diversity via incorporating multiple sources of quantification profiles, as we have taken in this study. The integration of the protein abundance profiles and large-scale AP-MS experimental interaction networks, enable us to fill in the missing features caused by a single AP-MS experiment. For instance, we have observed that the protein MCM10 binds to the MCM2-7 complex via MCM6 in a potential transient manner.

Overall, we observe a good performance of our model, however there is still room for future improvements. Firstly, we included only the protein pairs with both interaction and abundance features to predict the PPI network. Ideally, the number of detected proteins accumulate and would ultimately reflect the total number of proteins in human proteome, when enough proteomic quantification profiles are collected. However, due to technical challenges, there is no such set of protein interaction studies that contain a comprehensive list of the whole proteomic-level baits. The lack of complete datasets limits the comprehensiveness of interaction features of protein pairs. The model based on these incomplete features therefore predicts an incomplete PPI network, which probably results in an incomplete protein complex map. Due to a lower performance of the deep learning model, using only the protein abundance features, an expediency can be as follows: the core PPI network can be predicted using the integrated model, while the peripheral network is populated only using the protein abundance features model. Then the inference of the protein complex map based on this integrated core-peripheral network needs to be further explored. Secondly, wet-lab experiments such as co-immunoprecipitations and more targeted approaches such as knockout studies, need to be performed to further validate and confirm the complexes. However, according to the evaluation metrics of the deep-learning model and protein complex map, we are convinced that the integration of the protein interaction features and the protein abundance features can improve the model's performance, compared to using either type of these features alone. Thus, our work provides a new methodology to improve the reconstruction of PPI interaction and the understanding of protein complexes.

In conclusion, by incorporating interaction stoichiometry and large-scale protein abundance stoichiometry, our deep learning framework serves as a pioneering protein complexes discovery analysis.

Acknowledgements

This work has been supported by EPIC-XS, project number 823839, funded by the Horizon 2020 programme of the European Union and the NWO funded Netherlands Proteomics Centre through the National Road Map for Large-scale Infrastructures program X-Omics, Project 184.034.019. BL was supported by the China Scholarship Council (CSC) no. 201606300049.

Author Contributions:

Bohui Li contributes to the design of the work, acquisition, analysis, interpretation of data, and drafting the manuscript; Bas van Breukelen revise it critically for important intellectual content; Bas van Breukelen and Maarten Altelaar final approval of the version to be published.

References

1. Havugimana, P. C., et al., A census of human soluble protein complexes. Cell 2012, 150, (5), 1068-81.

2. Williams, N. K.; Dichtl, B., Co-translational control of protein complex formation: a fundamental pathway of cellular organization? Biochem Soc Trans 2018, 46, (1), 197-206.

3. Marsh, J. A.; Teichmann, S. A., Structure, dynamics, assembly, and evolution of protein complexes. Annu Rev Biochem 2015, 84, 551-75.

4. Wu, Z.; Liao, Q.; Liu, B., A comprehensive review and evaluation of computational methods for identifying protein complexes from protein-protein interaction networks. Brief Bioinform 2020, 21, (5), 1531-1548.

5. Huttlin, E. L., et al., The BioPlex Network: A Systematic Exploration of the Human Interactome. Cell 2015, 162, (2), 425-440.

6. Paiano, A.; Margiotta, A.; De Luca, M.; Bucci, C., Yeast Two-Hybrid Assay to Identify Interacting Proteins. Curr Protoc Protein Sci 2019, 95, (1), e70.

7. Rual, J. F., et al., Towards a proteome-scale map of the human protein-protein interaction network. Nature 2005, 437, (7062), 1173-8.

8. Hein, M. Y., et al., A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell 2015, 163, (3), 712-23.

9. Huttlin, E. L., et al., Architecture of the human interactome defines protein communities and disease networks. Nature 2017, 545, (7655), 505-509.

10. Liu, X., et al., An AP-MS- and BioID-compatible MAC-tag enables comprehensive mapping of protein interactions and subcellular localizations. Nat Commun 2018, 9, (1), 1188.

11. Drew, K.; Muller, C. L.; Bonneau, R.; Marcotte, E. M., Identifying direct contacts between protein complex subunits from their conditional dependence in proteomics datasets. Plos Computational Biology 2017, 13, (10).

12. Wan, C. H., et al., Panorama of ancient metazoan macromolecular complexes. Nature 2015, 525, (7569), 339-+.

13. Drew, K., et al., Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. Molecular Systems Biology 2017, 13, (6).

14. Sarkar, D.; Saha, S., Machine-learning techniques for the prediction of protein-protein interactions. Journal of Biosciences 2019, 44, (4).

15. Oughtred, R., et al., The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Science 2021, 30, (1), 187-200.

16. Schweppe, D. K.; Huttlin, E. L.; Harper, J. W.; Gygi, S. P., BioPlex Display: An Interactive Suite for Large-Scale AP-MS Protein-Protein Interaction Data. Journal of Proteome Research 2018, 17, (1), 722-726.

17. Szklarczyk, D., et al., The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Research 2021, 49, (D1), D605-D612.

18. Zhang, J. X.; Zhong, C.; Huang, Y. R.; Lin, H. X.; Wang, M., A method for identifying protein complexes with the features of joint co-localization and joint co-expression in static PPI networks. Computers in Biology and Medicine 2019, 111.

19. Shieh, Y. W., et al., Operon structure and cotranslational subunit association direct protein assembly in bacteria. Science 2015, 350, (6261), 678-680.

20. Wu, Y. D., et al., Co-expression of key gene modules and pathways of human breast cancer cell lines. Bioscience Reports 2019, 39.

21. Liu, J. H., et al., Eleven genes associated with progression and prognosis of endometrial cancer (EC) identified by comprehensive bioinformatics analysis. Cancer Cell International 2019, 19.

22. Szklarczyk, D., et al., The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Research 2017, 45, (D1), D362-D368.

23. von Mering, C., et al., STRING: a database of predicted functional associations

between proteins. Nucleic Acids Research 2003, 31, (1), 258-261.

24. Ruepp, A., et al., CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic Acids Res 2010, 38, (Database issue), D497-501.

25. Vizcaino, J. A., et al., 2016 update of the PRIDE database and its related tools. Nucleic Acids Res 2016, 44, (22), 11033.

26. Important facts about cancer. Boston Medical and Surgical Journal 1920, 182, 125-126.

27. Hu.Map database. Available from: http://hu1.proteincomplexes.org/download.

28. Lee, I.; Blom, U. M.; Wang, P. I.; Shim, J. E.; Marcotte, E. M., Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Research 2011, 21, (7), 1109-1121.

29. Guruharsha, K. G., et al., A Protein Complex Network of Drosophila melanogaster. Cell 2011, 147, (3), 690-703.

30. Malovannaya, A., et al., Analysis of the Human Endogenous Coregulator Complexome. Cell 2011, 145, (5), 787-799.

31. Hein, M. Y., et al., A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. Cell 2015, 163, (3), 712-723.

32. R interface of Keras. Available from: https://keras.rstudio.com.

33. Tieleman, T.; Hinton, G., Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning 2012, 4, (2), 26-31.

34. Meyer, D., et al., e1071: Misc functions of the Department of Statistics (e1071), TU Wien. R package version 2014, 1, (3).

35. Chang, C.-C.; Lin, C.-J., LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST) 2011, 2, (3), 1-27.

36. Nepusz, T.; Yu, H. Y.; Paccanaro, A., Detecting overlapping protein complexes in protein-protein interaction networks. Nature Methods 2012, 9, (5), 471-U81.

37. Enright, A. J.; Van Dongen, S.; Ouzounis, C. A., An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 2002, 30, (7), 1575-1584.

38. Raudvere, U., et al., g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Research 2019, 47, (W1), W191-W198.

39. Uhlen, M., et al., Tissue-based map of the human proteome. Science 2015, 347, (6220).

40. Human Protein Atlas. Available from: https://www.proteinatlas.org/about/download, proteinatlas.tsv.zip.

41. Collins, B. C., et al., Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. Nature Methods 2013, 10, (12), 1246-+.

42. Vizcaino, J. A., et al., 2016 update of the PRIDE database and its related tools. Nucleic Acids Research 2016, 44, (D1), D447-D456.

43. Ruepp, A., et al., CORUM: the comprehensive resource of mammalian protein complexes-2009. Nucleic Acids Research 2010, 38, D497-D501.

44. Chen, H. L.; Zhou, H. X., Prediction of interface residues in protein-protein complexes by a consensus neural network method: Test against NMR data. Proteins-Structure Function and Bioinformatics 2005, 61, (1), 21-35.

45. Lage, K., et al., A human phenome-interactome network of protein complexes implicated in genetic disorders. Nature Biotechnology 2007, 25, (3), 309-316.

46. Kopp, F.; Dahlmann, B.; Kuehn, L., Reconstitution of hybrid proteasomes from purified PA700-20 S complexes and PA28 alpha beta activator: Ultrastructure and peptidase activities. Journal of Molecular Biology 2001, 313, (3), 465-471.

47. Sato, S., et al., A set of consensus mammalian Mediator subunits identified by multidimensional protein identification technology. Molecular Cell 2004, 14, (5), 685-691.

48. Homesley, L., et al., Mcm10 and the MCM2-7 complex interact to initiate DNA synthesis and to release replication factors from origins. Genes & Development 2000, 14, (8), 913-926.

49. Douglas, M. E.; Diffley, J. F. X., Recruitment of Mcm10 to Sites of Replication Initiation Requires Direct Binding to the Minichromosome Maintenance (MCM) Complex. Journal of Biological Chemistry 2016, 291, (11), 5879-5888.

50. Stuart, J. M.; Segal, E.; Koller, D.; Kim, S. K., A gene-coexpression network for global discovery of conserved genetic modules. Science 2003, 302, (5643), 249-255.

51. Jakel, S.; Gorlich, D., Importin beta, transportin, RanBP5 and RanBP7 mediate nuclear import of ribosomal proteins in mammalian cells. Embo Journal 1998, 17, (15), 4491-4502.

52. Jakel, S., et al., The importin beta/importin 7 heterodimer is a functional nuclear import receptor for histone H1. Embo Journal 1999, 18, (9), 2411-2423.

53. Wolfe, C. L.; Warrington, J. A.; Treadwell, L.; Norcum, M. T., A three-dimensional working model of the multienzyme complex of aminoacyl-tRNA synthetases based on electron microscopic placements of tRNA and proteins. Journal of Biological Chemistry 2005, 280, (46), 38870-38878.

54. Eraslan, G.; Avsec, Z.; Gagneur, J.; Theis, F. J., Deep learning: new computational modelling techniques for genomics. Nature Reviews Genetics 2019, 20, (7), 389-403.

55. Granovetter, M. S., The strength of weak ties. American journal of sociology 1973, 78, (6), 1360-1380.

56. Csermely, P., Weak links: Stabilizers of complex systems from proteins to social networks. Weak Links: Stabilizers of Complex Systems from Proteins to Social Networks 2006, 37.

57. Anderson, N. L.; Anderson, N. G., The human plasma proteome - History, character, and diagnostic prospects. Molecular & Cellular Proteomics 2002, 1, (11), 845-867.

Supporting information for chapter 4

Supplementary to “Identification of protein complexes by integrating protein abundance and interaction stoichiometries using a deep learning strategy”

Protein abundance stoichiometry contributes to capturing novel subunits: We predicted that MTF2 (also known as PCL2) is a subunit of the polycomb repressive complex 2 (PRC2) (Fig. S3A, B). The PRC2 core complex, consisting of SUZ12, EED, and EZH2 (1, 2), is important in chromatin compaction and catalyzes the methylation of histone H3 at lysine 27 (3, 4). We found that the novel member MTF2 showed a high concordance of abundance stoichiometry with the PRC2 core in a comparable set of profiles (Fig. S3B). It is reported that PCL proteins, including PCL1, PCL2 (MTF2) and PCL3, interact with PRC2 through EZH2, and to some extent through SUZ12 (5). Interestingly, we captured this novel interaction, i.e., MTF2-EED, through the interaction of MTF2 with the PRC2 complex, which is also supported by the interaction stoichiometry, albeit relatively weak, in the study by Hein et al. (Fig. S3A right).

Another example is the complex centralspindlin, reported as a heterotetramer consisting of a dimer of the kinesin KIF23 (also known as MKLP1) and a dimer of the accessory protein RACGAP1 (also known as Cyk4 or MgcRacGAP) (6). The SHC SH2-domain binding protein 1 (SHCBP1) was predicted by our method as a novel subunit that interacts with the centralspindlin complex through RACGAP1 (Fig.S3C, D). These proteins were detected in 30% (~ 2,200) of total abundance stoichiometry profiles (Fig. S3D), indicating their co-expression and co-occurring characteristics. Interestingly, interactions between SHCBP1 and RACGAP1 were detected multiple times in the AP-MS interaction map by Hein et al. (Fig. S3C right). These results suggest that incorporating protein abundance stoichiometry from diverse datasets could improve the prediction of protein complexes and enable the identification of novel interactions with high confidence.

The co-expression feature assists the identification of protein complexes: In addition to the IPO7- KPNB1 complex, we observed that two chaperone proteins VCP (also known as p97) and HSP90B1 (also known as gp96 or GRP94) displayed a high concordance in expression across a great number of the protein abundance stoichiometry profiles (Fig. S3E). It has been reported that VCP cooperates with diverse partner proteins to help process ubiquitin-labelled proteins for recycling or degradation by the proteasome in many cellular contexts (7). Interestingly, the gp96 was also demonstrated governing protein ubiquitination and degradation (8). Thus, we deduce that p97 and gp96 may interact in their regulation of protein ubiquitination and degradation in the endoplasmic reticulum.

References

1. Cao, R., et al., Role of histone H3 lysine 27 methylation in polycomb-group silencing. Science 2002, 298, (5595), 1039-1043.
2. Czermin, B., et al., Drosophila enhancer of Zeste/ESC complexes have a histone H3

methyltransferase activity that marks chromosomal polycomb sites. Cell 2002, 111, (2), 185-196.

3. Margueron, R.; Reinberg, D., The Polycomb complex PRC2 and its mark in life. Nature 2011, 469, (7330), 343-349.

4. Laugesen, A.; Hojfeldt, J. W.; Helin, K., Molecular Mechanisms Directing PRC2 Recruitment and H3K27 Methylation. Molecular Cell 2019, 74, (1), 8-18.

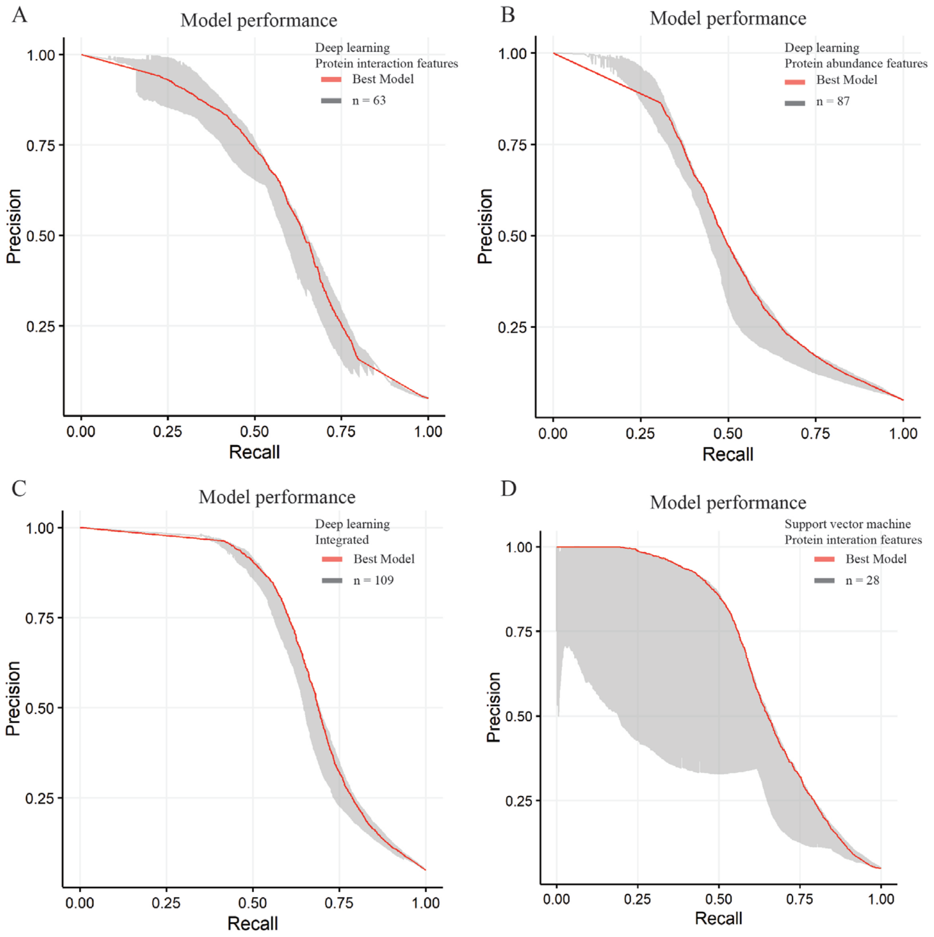
5. Nekrasov, M., et al., Pcl-PRC2 is needed to generate high levels of H3-K27 trimethylation at Polycomb target genes. Embo Journal 2007, 26, (18), 4078-4088.

6. Mishima, M.; Kaitna, S.; Glotzer, M., Central spindle assembly and cytokinesis require a kinesin-like protein/RhoGAP complex with microtubule bundling activity. Developmental Cell 2002, 2, (1), 41-54.

7. Meyer, H.; Bug, M.; Bremer, S., Emerging functions of the VCP/p97 AAA-ATPase in the ubiquitin system. Nature Cell Biology 2012, 14, (2), 117-123.

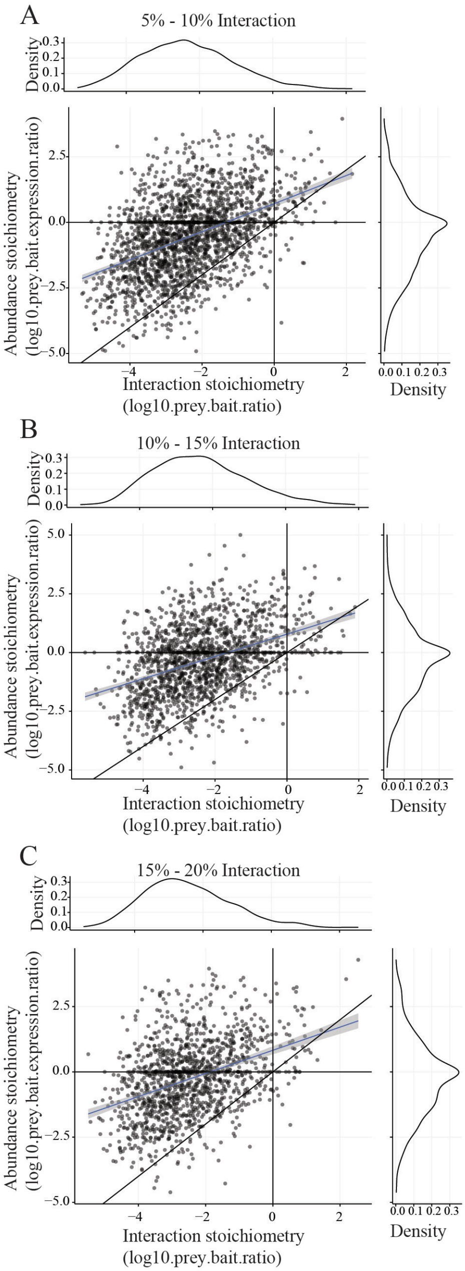
8. Wu, B., et al., Heat shock protein gp96 decreases p53 stability by regulating Mdm2 E3 ligase activity in liver cancer. Cancer Letters 2015, 359, (2), 325-334.

Supplementary figures



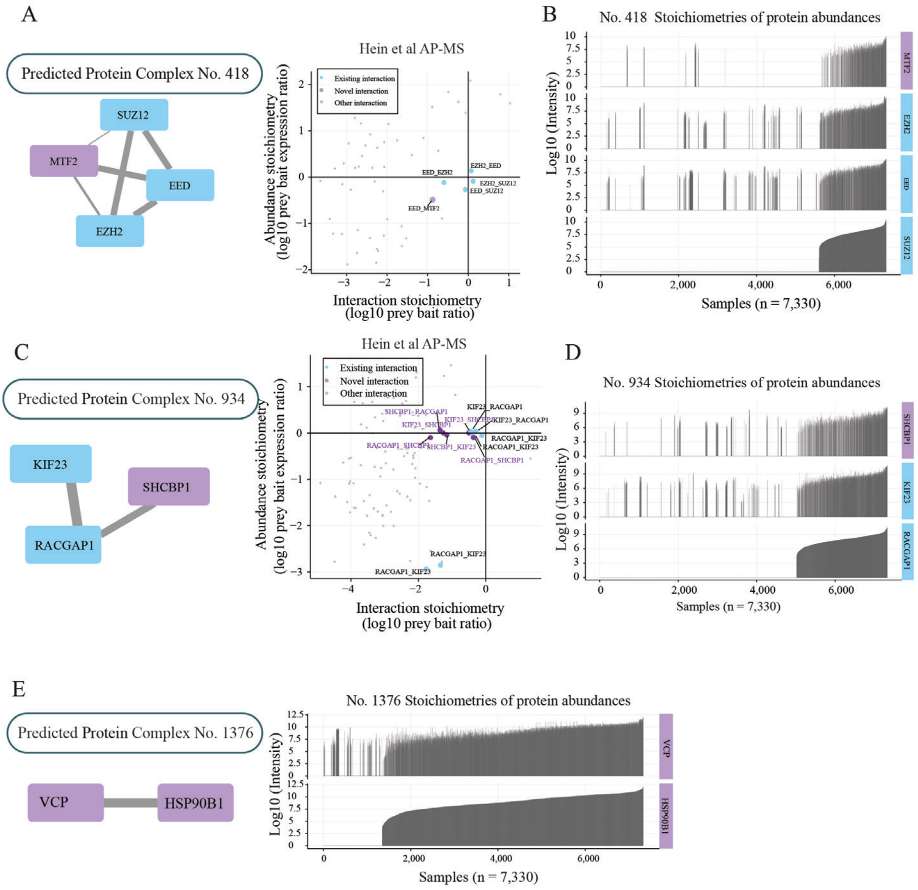
Supplementary Figure 1. Comparison of model performance. The precision is calculated by true positive / (true positive + false positive), and the recall is calculated by true positive / (true positive + false negative). A harmonic mean of precision and recall, namely F1-measure or F1-score was furtherly used to decide the model performance.

A, Precision-recall curve of the 63 protein interaction feature matrix based deep learning models, the curve highlighted in red is the best model with F1-measure at 0.61. B, The training of 87 deep learning models with protein abundance features and an F1-measure > 0.49 were used to represent in the precision-recall plot, the curve highlighted in red is the best model with F1-measure at 0.51. C, Precision-recall curve indicating 109 integrated deep learning models with an F1-measure > 0.66, the outperforming model is shown in red with F1-score at 0.68. D, A total of 28 SVM models were trained with the protein interaction features, the red line shows the best SVM classifier with F1-measure at 0.64.



Supplementary Figure 2. Evaluation of predicted protein-protein interactions

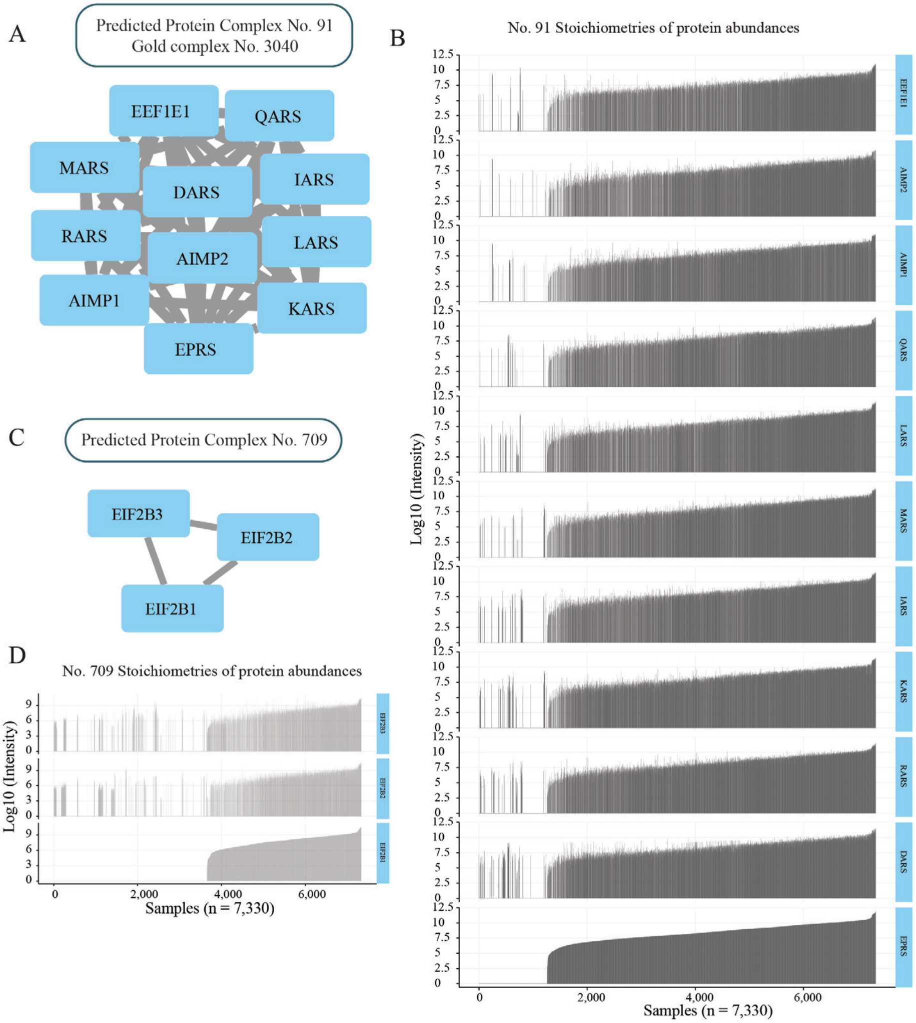
A-C, Scatterplot showing the abundance and interaction stoichiometries as obtained from Hein et al. experiment with different percentage of interaction from our predicted PPI network. A weaker interaction stoichiometry was observed when decreasing the protein interaction confidence, suggesting the importance of this filtering step in obtaining an optimal PPI network.



Supplementary Figure 3. Protein Complexes with novel subunits as well as newly predicted highly co-expressed protein complexes that are predicted by our model

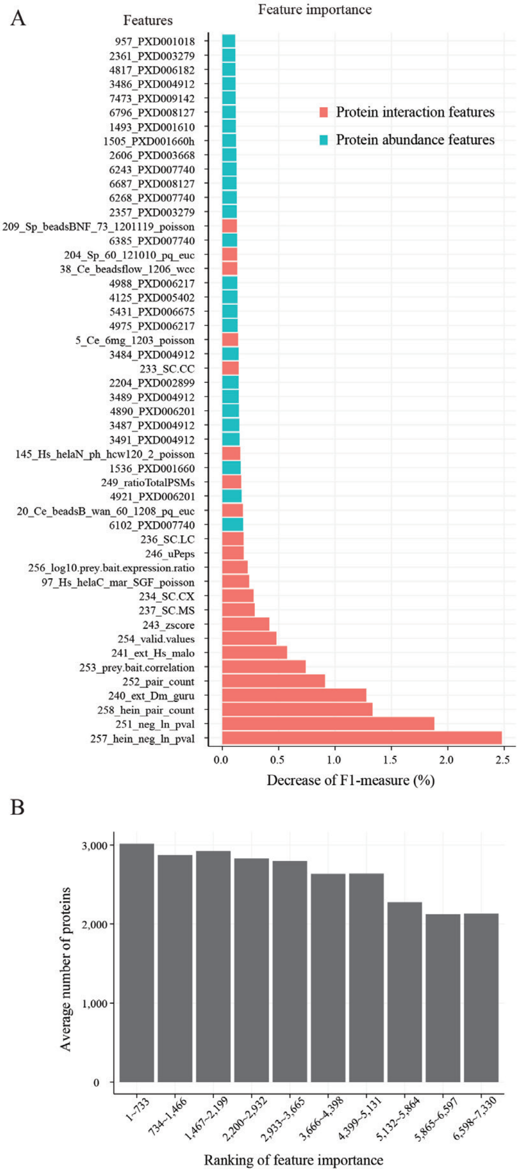
A left panel; Interaction network of the core Polycomb repressive complex 2 (PRC2, highlighted in blue) with a potential novel subunit MTF2 (purple) as predicted by our model, A right panel; Scatter plot with interaction and abundance stoichiometry for the PRC2 complex from Hein et al' AP-MS experiments. Blue dots are known interactions, the purple dots are novel interactions. Labels for the dots are represented by Bait Prey proteins. The predicted EED-MTF2 interaction shows a correlation with the other proteins from the PRC2 complex albeit with lower protein abundance and interaction stoichiometry. Suggesting an interaction between EED-MTF2 albeit a somewhat weaker. B, The expression pattern of each subunit within the PRC2 protein complex. On each row, the X-axis indicates 7,330 samples collected from PRIDE repository and the Y-axis indicates the log10 transformed intensity of corresponding protein, where missing values are in blanks. This plot shows that the expression pattern of MTF2 shares a high concordance with the subunits of the PRC2 complex, suggesting the abundance stoichiometry of proteins could improve the model sensitivity. C left panel; Interaction network of the Centralspindlin complex (core members are in blue and new predicted protein is in purple). C right panel; Interaction-abundance stoichiometries for Centralspindlin complex from Hein et al' AP-MS experiments. Blue dots are known interactions, purple dots are novel interactions. Labels for the dots are represented by Bait_Prey proteins. D, The expression pattern of each subunit within the Centralspindlin protein complex. On each row, the X-axis indicates 7,330 samples collected from the PRIDE repository and the Y-axis indicates the log10 transformed intensity, missing values are in blanks. It can be seen that the novel subunit SHCBP1 shows a high similarity of expression pattern with the subunits of the Centralspindlin complex, suggesting a high possibility of co-occurrance for these proteins. E, Interaction network and protein expression pattern plot for the predicted complex VCP-HSP90B1. These two proteins displayed a high concordance in expression across around

5,700 of the protein abundance stoichiometry profiles, however, it has not been detected in Hein et al' AP-MS experiment also showing the sensitivity of our deep learning model and the importance of integrating the protein abundance stoichiometries in protein complex predictions.



Supplementary Figure 4. Protein complex members show a significant co-expression.

A, Our model predicts the interaction network of the 11 subunits multi-synthetase protein complex, which is also a well-defined protein complex found in the CORUM database. This indicates our deep learning model possesses a high accuracy and robustness. B, The expression pattern of the multi-synthetase protein complex. The X-axis indicates 7,330 samples collected from PRIDE repository. The Y-axis indicates the log10 transformed intensity of protein, missing values are in blanks. This multi-protein complex contains 11 subunits, which are all co-expressed in most of the profiles. C, Interaction network of eukaryotic initiation factor complex with 3 protein subunits. D, The expression pattern of the eukaryotic initiation factor complex. This protein complex contains 3 subunits that are co-expressed in across around 2800 profiles, indicating the importance of the co-expression property in predicting protein complexes.



Supplementary Figure 5. Evaluation of feature importance

A, Bar plot shows the importance of features. The importance is indicated by the decrease of F1 measure using the randomly shuffling the values of each feature (see Methods). Red bars represent the protein interaction features and blue the protein abundance features. B, Bar plot shows the average number of proteins in the protein abundance profiles with different intervals of importance ranking.

Chapter 5

Summary and future outlook

Summary and future outlook

1 summary

In this thesis, I describe the work I did during my PhD using bioinformatics tools and machine learning algorithms to address the increasing scale and complexity of proteomics endeavors, covering the facets of drug resistant mechanisms in melanoma, the role of the PD1 in T-cell activation, and protein-protein interactions and protein complexes prediction.

In **chapter one**, a general introduction to the basic principles and techniques of MS-based proteomics, quantification strategies, and a generalized shotgun proteomics workflow are given. Moreover, I also outline how to analyze proteomics data from a bioinformatics perspective including normalization, dealing with missing values, differential analysis, functional annotation, as well as how to reveal the biology from post-translational modification data. Furthermore, I generalized the basics of machine learning algorithms from the perspective of supervised and unsupervised machine learning, along with that the application of machine learning algorithms to the identification of protein complexes.

In **chapter two**, we are seeking to explore the drug addiction mechanism in melanoma cells that carry BRAF mutation. We present a proteomics and phosphoproteomics study of BRAFi-addicted melanoma cells (i.e., 451Lu cell line) in response to BRAFi withdrawal, in which ERK1, ERK2, and JUNB were genetically silenced separately using CRISPR-Cas9. We show that inactivation of ERK2 and, to a lesser extent, JUNB prevents drug addiction in these melanoma cells, while, conversely, knockout of ERK1 fails to reverse this phenotype, showing a response similar to that of control cells. Our data indicate that ERK2 and JUNB share comparable proteome responses dominated by the reactivation of cell division. Importantly, we find that EMT activation in drug-addicted melanoma cells upon drug withdrawal is affected by silencing ERK2 but not ERK1. Moreover, we reveal that PIR acts as an effector of ERK2 and phosphoproteome analysis reveals that silencing of ERK2 but not ERK1 leads to the amplification of GSK3 kinase activity. Our results depict possible mechanisms of drug addiction in melanoma, which may provide a guide for therapeutic strategies in drug-resistant melanoma.

More recently, immunotherapy shows promising clinical efficacy in the treatment of melanoma. The most successful immunotherapy for melanoma is immune checkpoint inhibiting, including the PD-1/PD-L1 inhibitors (pembrolizumab, pidilizumab, and nivolumab). Thus, in **chapter three**, we are dedicated to exploring the role of PD-1 in T cell activation by comparing the proteome and phosphoproteome profiles in resting and activated CD8+ T cells, in which PD-1 was silenced using CRISPR–Cas9. Our data reveal that the activated T cells reprogrammed their proteome and phosphoproteome marked by activating of mTORC1 pathway. Moreover, we find that silencing of PD-1 altered the expression of E3 ubiquitin--protein ligases, and increased glucose and lactate transporters. On the phosphoproteomics level, it evokes phosphorylation events in the mTORC1 pathway and activates the epidermal growth factor and its downstream MAPK pathway. Therefore, the data presented in this chapter depicts mechanisms of PD-1 in response to TCR stimulation in CD8+ T cells, which may provide a guide in immune homeostasis and immune checkpoint therapy.

In **chapter four**, we construct a comprehensive map of human protein complexes through integration of protein-protein interactions (from 3 published affinity purification/co-fractionation mass spectrometry datasets) and protein abundance (from >7000 quantitative proteomic profiles) features. A deep learning framework was built to predict protein-protein interactions (PPIs), followed by a two-stage clustering to identify protein complexes. Our deep learning technique-based classifier significantly outperformed recently published machine learning prediction models with an F1-measure of 0.68 and captured in the process 5,010 complexes containing over 9,000 unique proteins. Moreover, this deep learning model enables us to capture poorly characterized interactions and the co-expressed protein involved interactions.

2 Nederlandse Samenvatting

In dit proefschrift beschrijf ik de toepassing van bioinformatica en kunstmatige intelligentie om inzichten te verkrijgen in complexe proteomics data. De biologische focus lag hierbij op geneesmiddel resistentie in melanoma, de rol van PD1 in T-cell activering, eiwit-eiwit interacties en het voorspellen van eiwitcomplexen.

Hoofdstuk 1 bevat een algemene introductie over de basisprincipes van massaspectrometrie en proteomics, beschrijft verschillende kwantitatieve proteomics methoden en er wordt een algemene shotgun proteomics workflow beschreven. Daarnaast vermeld ik in dit hoofdstuk hoe proteomics data geanalyseerd dient te worden vanuit een bioinformatica perspectief, waarbij data normalisatie, de werkwijze betreffende ontbrekende data, differentiële analyse, functionele annotatie en de biologische interpretatie van post translationele modificatie data, de revue passeren. Als laatste beschrijf ik de basisprincipes van kunstmatige intelligentie, inclusief gecontroleerde en ongecontroleerde methoden en de toepassing van kunstmatige intelligentie algoritmen voor de identificatie van eiwitcomplexen.

In Hoofdstuk 2 worden de mechanismen van geneesmiddel verslaving in melanoma cellen met een BRAF-mutatie onderzocht. Ik beschrijf hier een proteomics en phosphoproteomics studie naar BRAFi-verslaafde melanoma cellen (451Lu cellijn) na BRAFi onthouding, waarbij ERK1, ERK2 en JUNB individueel genetisch werden stilgelegd door middel van CRISPR-Cas9. Inactiviteit van ERK2 en in mindere mate van JUNB, voorkomt geneesmiddel verslaving in deze melanoma cellijn, terwijl de inactiviteit van ERK1 dit effect niet heeft. Onze data latenzien dat ERK2 en JUNB vergelijkbare effecten hebben op het eiwitpatroon in de cel, dat wordt gedomineerd door heractivering van de celdeling. Een belangrijke bevinding in dit hoofdstuk is dat ERK2 inactiviteit invloed heeft op de EMT activatie in geneesmiddel verslaafde melanoma cellen na geneesmiddel onthouding, terwijl ERK1 dit effect niet heeft. Verder laten we zien dat PIR een effector eiwit is van ERK2 en dat het stilleggen van ERK2 (maar niet van ERK1) leidt tot amplificatie van GSK3 kinase activiteit. De resultaten gepresenteerd in dit hoofdtuk geven inzicht in mogelijke mechanismen van geneesmiddel verslaving in melanoma cellen, wat een richtlijn kan verschaffen voor de therapeutische behandeling van geneesmiddel resistentie in melanoma.

Immunotherapie heeft recentelijk veelbelovende resultaten laten zien in de behandeling van melanoma. De meest succesvolle behandelstrategie in melanoma is het gebruik van checkpointremmers, zoals PD-1/PD-L1 remmers (pembrolizumab, pidilizumab en nivolumab). In Hoofdstuk 3 leggen wij ons toe op het onderzoeken van de rol van PD-1 in T-cel activering door de (phospho)proteome profielen van rustende en geactiveerde CD8+ T-cellen met een PD-1 silencing-mutatie te vergelijken. Onze data laten zien dat de mTORC1 signaalroute geactiveerd wordt in geactiveerde T-cellen en dat het stilleggen van PD-1 de expressie van E3 ubiquitin-eiwit ligases beïnvloedt en leidt tot een toename van het aantal glucose en lactaat transporters. Daarnaast verhoogt PD-1 de mate van fosforylatie van de mTORC1 signaalroute en activeert het de epidermale groeifactor receptor en de bijbehorende MAPK signaalroute. De data die worden gepresenteerd in dit hoofdstuk geven inzicht in de rol van PD-1 na TCR stimulatie in CD8+ T-cellen en dragen daarmee bij aan ons begrip van immuun homeostase en immuun checkpoint therapie.

Veel essentiële cellulaire functies worden uitgevoerd door eiwitcomplexen die bestaan uit meerdere eiwitten die bijeen worden gehouden door eiwit-eiwit interacties. Deze interacties zijn afhankelijk van de sterkte van de individuele eiwitinteracties en de abundantie van de betreffende eiwitten, die beide sterk van grootte kunnen verschillen. Hoofdstuk 4 geeft een uitgebreide beschrijving van humane eiwitcomplexen door de integratie van informatie over eiwitinteracties (van 3 gepubliceerde datasets) en informatie over eiwitabundantie (van >7000 kwantitatieve proteomics profielen). We hebben een deep learning framework opgezet om eiwit-eiwit interacties te voorspellen, gevolgd door een two-stage clustering om eiwitcomplexen te identificeren. Deze methode gaf een significante verbetering ten opzichte van recent gepubliceerde machine learning modellen, had een F1 van 0,68 en identificeerde 5.010 complexen met meer dan 9.000 unieke eiwitten. Dit model stelt ons verder in staat om slecht gekarakteriseerde interacties te identificeren en verschaft inzicht in co-expressie van eiwitten die betrokken zijn bij interacties.

3 Future outlook

Mass spectrometry (MS)-based proteomics has matured into an attractive technology for global analysis of protein expression, composition, modifications, and dynamics. It is an indispensable tool for cellular biology and clinical research and is now being routinely applied for high-throughput identification and quantification of proteins, post-translational modifications, as well as protein-protein interactions (1, 2). However, still plenty of challenges remain, from the need to extract biological function from proteomics data, to reaching a higher accuracy of predicting protein-protein interactions.

3.1 Revealing biological functions from proteomics data

Advances in mass spectrometry (MS)-based proteomics have been enabled the fast growth of proteomics data, which in turn resulted in the development of bioinformatics tools and infrastructures for processing, storing, and analyzing proteomics data. First of all, a lot of open-source or commercial software has been developed for proteomic data processing, such as Maxquant (3), Proteome Discoverer (4), and OpenMS (5). Moreover, to store the increasing amount of proteomics data, several public repositories are available, such as the PRIDE (6), ProteomeXchage (7), and peptide atlas (8). The PTM site information, as well as corresponding kinases, were implemented into several databases, such as PhosphoSitePlus (9), Phospho.ELM(10), and PHOSIDA (11). These public databases have provided access for statisticians to compare their results to the published studies, which will help researchers generalize a more reliable biological conclusion. Moreover, when there is a limited number of samples, especially the patient tissue samples, one can integrate datasets from public databases to make a better analysis of MS data.

To reveal the biological functions, the bioinformatics analysis typically involves the integration of proteome data with annotation databases, such as Metascap (12), Gene Ontology (13), pathway database (KEGG) (14), gene set enrichment analysis GSEA(15), and protein domains (InterPro, PFAM) (16). This type of analysis can directly reflect the functional insights into the data set and is easily achieved using public tools. Besides, software and R packages, such as Limma (17), and WGCNA (18), have been designed to compare differentially expressed proteins between groups (e.g., healthy and disease) and/or obtain co-expression patterns, which offer broader capabilities and flexibility in analysis but require some more programming experience. Moreover, several data visualization tools like Cytoscape (19), and ggplot2 (20) also require users to master basic programming principles. Therefore, there is a urgent need to develop a user-friendly graphical user interface (GUI), which ideally can integrate data mining, functional annotation, and data visualization into an online platform. At the same time, MS-based proteomics data-sets are currently analyzed with algorithms designed for genomics and transcriptomics datasets or other statistical methods. Although those algorithms have been proven to be robust and useful, they have limitations. Thus, an improvement of existing algorithms and/or the development of novel sophisticated data-mining methods for analysis and interpretation can be expected.

3.2 Protein-protein interactions and protein complexes

Proteins do not work independently, many essential cellular functions are carried out by multi-protein complexes that can be characterized by their protein-protein interactions. Investigating protein-protein interactions and protein complexes allows one to place a protein with completely unknown functions into a context given by their interacting partners with already known functions. Thus, one can deduce the novel function for the nearly unknown proteins, or design a reasonable experiment to test its biological function based on interacting partners. Moreover, one protein can have a completely different function when interacting with different partners or in a different biological process. Therefore, an increasing number of large-scale proteomics projects have been conducted in many research institutions to build comprehensive interactome maps.

Recent advances in technologies lead to an increase in protein interaction data and resulting interaction networks and databases (21). The two most frequently used methods in the identification of protein interactions are yeast two-hybrid (Y2H) screening, a well-established genetic in vivo approach, and affinity purification followed by mass spectrometry analysis (AP-MS), an emerging biochemical in vitro technique. So far, a majority of published interactions have been detected using a Y2H screen. For instance, comprehensive protein interaction maps have been established using Y2H in Saccharomyces cerevisiae (22, 23), Drosophila melanogaster (24), and humans (25, 26). Although the Y2H approach is a powerful tool in the characterization of protein-protein interactions, some limitations have become apparent. False negatives (protein-protein interactions which cannot be detected) and false positives (physical interactions detected in the screening which are not reproducible in an independent system) are the most two considerations in the application of Y2H. More recently, affinity purification followed by mass spectrometry (AP-MS) analysis becomes more and more popular in the characterization of comprehensive interactomes in humans and other organisms. For instance, Havugimana (27) and collogues have identified a network of 13,993 high-confidence physical interactions among 3,006 stably associated soluble human proteins, which markedly increases the coverage of protein-protein interactions. Moreover, Huttlin (28, 29) and collogues have established the BioPlex interaction network, which contains 56 553 interactions from 5891 AP-MS experiments.

A prominent concern in the AP-MS protein-protein interaction studies is the identification of true and specific PPIs as opposed to nonspecifically co-purified proteins. These non-specifically interacted proteins are from (i) non-specific interacting proteins that binding to the tag, (ii) the carryover of overexpressed proteins. Several methods have aimed to develop bioinformatics computational tools to classify the true interactions from the vast number of potential interactions. For instance, the Comparative Proteomics Analysis Software Suite (CompPASS) (30) was developed to identify high confidence interactions in AP-MS experiments using mass spectrometry spectral counts. Another computational method, the Significance Analysis of

INTeractome (SAINT) (31), uses the MS quantitative data and generates separate Poisson distributions for true and false interactions to derive the interaction probability. However, these computational pipelines only are suitable for a single AP-MS experimental system. Another possible strategy is to integrate several large-scale protein-protein interaction studies as the data source to generate machine learning models, thus improving the interaction accuracy, at the same time to reduce the non-specific interactions. For instance, Drew and colleagues (32) have constructed a comprehensive map of protein complexes by integrating large-scale mass spectrometry experimental datasets and employing a support vector machine (SVM) classifier to assign the interaction score. In this study, over 9000 mass spectrometry protein interaction datasets from a variety of human and animal cells and tissues were integrated into the analysis. Integrating large-scale protein-protein interaction data is a wise attempt since a protein can be highly dynamic and depend on the growth condition of the cell. Besides, the composition of protein interaction modules or protein complexes changes depending on the cell state and environment. However, at this moment, analyzing protein complexes by integrating different large-scale datasets is only limited to the AP-MS experimental data or the Y2H experimental data. Therefore, in the future, more studies should be carried out to combine information from Y2H and AP-MS experiments to increase the coverage and accuracy of protein-protein interaction and protein complex analysis.

3.3 Machine learning meets proteomics

The continuous improvement of mass spectrometry (MS)-based proteomics technologies has accumulated proteomics data at an astonishing speed and scope. Mining for biological functions and protein-protein interactions in proteomics data has been expanded to the statistical analysis and computational modeling. Fortunately, artificial intelligence (AI) more and more plays critical roles in mining the ‘big data’ including those from biological and medical contexts. As a branch of artificial intelligence, machine learning is to learn rules from data through computational models and algorithms. It is devoted to exploring how to improve the performance of the system itself through computing and using experiences.

Machine learning has already been applied to almost all steps of MS-based proteomics. This ranges from protein digestion, liquid chromatography separation, ionization, ion mobility separation, MS detection, peptide fragmentation, protein identification and quantification, to downstream data analysis and biomarker prediction. For instance, Zhang introduced a mathematical model based on classical kinetics and the mobile proton model of peptide fragmentation, for predicting the low-energy CID spectrum of singly or doubly charged peptides acquired on a quadrupole ion trap mass spectrometer (33). Furthermore, Michael and colleagues developed the Percolator algorithm (34), which employs semi-supervised learning as a postprocessor rather than a fixed discriminant function, to learn a function that consistently ranks the decoy peptide-spectrum matches (PSMs) below a subset of high-confidence target PSMs. Percolator uses an iterative, SVM-based algorithm, initially identifying a small set of high-scoring target PSMs, and then learning to separate these from the decoy PSMs. The educated classifier is applied to the entire set, and if new high-confidence PSMs are identified, then the procedure is repeated. Machine learning-based algorithms can efficiently process large amounts of peptide sequences. However, it is largely affected by feature extraction.

Deep learning (DL), a subset of machine learning (ML), can automatically learn data patterns and abstract features without handcrafted feature engineering. Deep learning is increasingly applied to a variety of proteomics research problems, such as predicting peptide properties (tandem mass spectra, ion mobility, and retention time) (35) from only a primary sequence. Furthermore, deep learning has been applied to improve peptide identification, protein inference, peak detection (36-38), protein structure modeling (39), protein-protein interaction prediction (40, 41), and neoantigen identification and patient classification (42, 43). One limitation of deep learning is that usually tens of thousands of example data points are required to effectively train a neural network. Interpretation of data by deep learning will require the production of thousands of proteome examples, which represents a major barrier given the average throughput of most experiments. Nonetheless, there are still options available to take on this challenge. One potential solution is to obtain proteomics data from public repositories, such as the PRIDE archive and BioPlex Interactome, which store detailed metadata about the sample preparation, individual treatments, and data acquisition. Alternatively, the combination of multiple omics will enlarge the volume of data types if not the data size.

3.4 Integration of multi-omics datasets

The biological networks in cells or tissues involve a highly dynamic and interactive system and are influenced by many environmental factors. Recent technological advancements in high throughput have enabled the measurement of genome, transcriptome, proteome, and metabolome in humans and other organisms (44-46). However, a single layer of “omics” can only provide limited insights into the biological mechanisms. The upcoming trend in "omics" studies is to integrate proteomics data with other layers of biological macromolecules, i.e., genomics, transcriptomics, and metabolomics. Several studies have been conducted to integrate different layers of “omics” data to achieve a more comprehensive understanding of cellular dynamics in biological systems (47-50). For instance, recently one jointly analyzed proteome-transcriptome data of mouse liver samples, which observed a lower level of protein-mRNA correlation (51). Interestingly, only about half of the tested genes show a significant correlation, and little overlap was found between the protein- and transcript-associated loci. Another study integrated metabolomics and genomics data to define prognosis characteristics in human neuroendocrine cancers (52). Researchers used genomics signatures to identify activated portions of the metabolic network, then to guide in silico metabolic reconstructions of neuroendocrine cell metabolism to detect specific metabolic changes, and ultimately identify a molecular signature that is associated with poor-prognosis human neuroendocrine cancers. In addition, one study integrated proteome and transcriptome data in the mouse lens epithelium and fibers and revealed that crystallins showed a high correlation between their mRNA and protein levels (53). However, most integrative transcriptomic and proteomic research have so far either failed to detect a correlation or only a weak correlation. Therefore, the consensus nowadays is that the correlation between transcriptomes and proteomes across large datasets was typically modest, and that mRNA levels could not be consistently relied upon to predict protein abundance (54-56). Given these examples of integrative omics research, we can expect that in the near future, more and more efforts would be devoted to developing sophisticated bioinformatics tools and statistical methods to improve the chances of better integrating multi-omics datasets and thereby reveal new biological insights that are not accessible through one-dimensional datasets.

As technologies continue to rapidly advance concerning throughput and sensitivity, bioinformatics tools must keep pace with large-scale experiments. In the come years, more general bioinformatics pipelines/platforms and user-friendly graphical user interface (GUI), which integrates data quality control, data mining, statistical analysis, functional annotation, pathway enrichment, and data visualization will be designed to help researchers. Such platforms should be free, open-source, and easy to access, which are capable of non-programming scientists and researchers with little bioinformatics knowledge. The generalization and application of more clear bioinformatics platforms will have a profound impact on extracting biological function from mass spectrometry-based proteomics data and further understanding biological and clinical research questions, such as elucidating underlying biological mechanisms in diseases or identifying promising biomarkers and novel drug targets. Moreover, given the high dimensionality and throughput of proteomics datasets, there is a great need for machine learning that will enable us to automatically find expression patterns, clusters, and protein-protein interactions underlying big-data matrix. Furthermore, the application of machine learning methods such as support vector machine and deep learning, can automatically classify disease, classify response types during treatment, and implement real-time diagnosis that will advance precision medicine, personalized treatment, and health management. Finally, to study complex biological processes, it is imperative to take an integrative approach that combines multi-omics data (genomics, transcriptomics, proteomics, phosphorproteomics, and metabolomics) to highlight the interrelationships of the involved biomolecules and their functions. Integration of multi-omics data providing information on biomolecules from different layers will be a promising method to understand the molecular mechanisms, processes, and pathways underlying complex biology and disease. Although the underlying heterogeneity in individual omics data and the large size of datasets make multi-omics data integration a challenging task, we believe that a uniform framework that can effectively process and analyze multi-omics data in an end-to-end manner along with easy and biologist-friendly visualization and interpretation will be developed.

References

1. Aebersold, R.; Mann, M., Mass spectrometry-based proteomics. Nature 2003, 422, (6928), 198-207.

2. Altelaar, A. F.; Munoz, J.; Heck, A. J., Next-generation proteomics: towards an integrative view of proteome dynamics. Nat Rev Genet 2013, 14, (1), 35-48.

3. Tyanova, S.; Temu, T.; Cox, J., The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nature Protocols 2016, 11, (12), 2301-2319.

4. Orsburn, B. C., Proteome Discoverer-A Community Enhanced Data Processing Suite for Protein Informatics. Proteomes 2021, 9, (1).

5. Sturm, M., et al., OpenMS-An open-source software framework for mass spectrometry. Bmc Bioinformatics 2008, 9.

6. Martens, L., et al., PRIDE: The proteomics identifications database. Proteomics 2005, 5, (13), 3537-3545.

7. Vizcaino, J. A., et al., The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Research 2013, 41, (D1), D1063-D1069.

8. Desiere, F., et al., The PeptideAtlas project. Nucleic Acids Research 2006, 34, D655-D658.

9. Hornbeck, P. V., et al., PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Research 2012, 40, (D1), D261-D270.

10. Dinkel, H., et al., Phospho.ELM: a database of phosphorylation sites-update 2011. Nucleic Acids Research 2011, 39, D261-D267.

11. Gnad, F., et al., PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biology 2007, 8, (11).

12. Zhou, Y. Y., et al., Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nature Communications 2019, 10.

13. Ashburner, M., et al., Gene Ontology: tool for the unification of biology. Nature Genetics 2000, 25, (1), 25-29.

14. Kanehisa, M.; Goto, S.; Kawashima, S.; Okuno, Y.; Hattori, M., The KEGG resource for deciphering the genome. Nucleic Acids Research 2004, 32, D277-D280.

15. Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005, 102, (43), 15545-50.

16. Hunter, S., et al., InterPro: the integrative protein signature database. Nucleic Acids Research 2009, 37, D211-D215.

17. Ritchie, M. E., et al., limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 2015, 43, (7).

18. Langfelder, P.; Horvath, S., WGCNA: an R package for weighted correlation network analysis. Bmc Bioinformatics 2008, 9.

19. Shannon, P., et al., Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Research 2003, 13, (11), 2498-2504.

20. Villanueva, R. A. M.; Chen, Z. J., ggplot2: Elegant Graphics for Data Analysis, 2nd edition. Measurement-Interdisciplinary Research and Perspectives 2019, 17, (3), 160-167.

21. Gingras, A. C.; Gstaiger, M.; Raught, B.; Aebersold, R., Analysis of protein complexes using mass spectrometry. Nature Reviews Molecular Cell Biology 2007, 8, (8), 645-654.

22. Ito, T., et al., A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proceedings of the National Academy of Sciences of the United States of America 2001, 98, (8), 4569-4574.

23. Uetz, P., et al., A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 2000, 403, (6770), 623-627.

24. Formstecher, E., et al., Protein interaction mapping: A Drosophila case study (vol 15, pg 376, 2005). Genome Research 2005, 15, (4), 601-601.

25. Stelzl, U., et al., A human protein-protein interaction network: A resource for annotating the proteome. Cell 2005, 122, (6), 957-968.

26. Rual, J. F., et al., Towards a proteome-scale map of the human protein-protein interaction network. Nature 2005, 437, (7062), 1173-1178.

27. Havugimana, P. C., et al., A Census of Human Soluble Protein Complexes. Cell 2012, 150, (5), 1068-1081.

28. Huttlin, E. L., et al., The BioPlex Network: A Systematic Exploration of the Human Interactome. Cell 2015, 162, (2), 425-440.

29. Huttlin, E. L., et al., Architecture of the human interactome defines protein communities and disease networks. Nature 2017, 545, (7655), 505-+.

30. Mellacheruvu, D., et al., The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. Nature Methods 2013, 10, (8), 730-+.

31. Choi, H., et al., SAINT: probabilistic scoring of affinity purification-mass spectrometry data. Nature Methods 2011, 8, (1), 70-U100.

32. Drew, K., et al., Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. Molecular systems biology 2017, 13, (6), 932.

33. Zhang, Z. Q., Prediction of low-energy collision-induced dissociation spectra of peptides. Analytical Chemistry 2004, 76, (14), 3908-3922.

34. Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J., Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nature Methods 2007, 4, (11), 923-925.

35. Guan, S. H.; Moran, M. F.; Ma, B., Prediction of LC-MS/MS Properties of Peptides from Sequence by Deep Learning. Molecular & Cellular Proteomics 2019, 18, (10), 2099-2107.

36. Demichev, V.; Messner, C. B.; Vernardis, S. I.; Lilley, K. S.; Ralser, M., DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. Nature Methods 2020, 17, (1), 41-+.

37. Kim, M.; Eetemadi, A.; Tagkopoulos, I., DeepPep: Deep proteome inference from peptide profiles. Plos Computational Biology 2017, 13, (9).

38. Ma, C., et al., DeepRT: deep learning for peptide retention time prediction in proteomics. arXiv preprint arXiv:1705.05368 2017.

39. Gao, W.; Mahajan, S. P.; Sulam, J.; Gray, J. J., Deep Learning in Protein Structural Modeling and Design. Patterns (N Y) 2020, 1, (9), 100142.

40. Sun, T.; Zhou, B.; Lai, L.; Pei, J., Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinformatics 2017, 18, (1), 277.

41. Patel, S.; Tripathi, R.; Kumari, V.; Varadwaj, P., DeepInteract: Deep Neural Network Based Protein-Protein Interaction Prediction Tool. Current Bioinformatics 2017, 12, (6), 551-557.

42. Bulik-Sullivan, B., et al., Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. Nature Biotechnology 2019, 37, (1), 55-+.

43. Behrmann, J., et al., Deep learning for tumor classification in imaging mass spectrometry. Bioinformatics 2018, 34, (7), 1215-1223.

44. Consortium, G. T., Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015, 348, (6235), 648-60.

45. Kim, M. S., et al., A draft map of the human proteome. Nature 2014, 509, (7502), 575-81.

46. Shin, S. Y., et al., An atlas of genetic influences on human blood metabolites. Nat Genet 2014, 46, (6), 543-550.

47. Zhang, B., et al., Proteogenomic characterization of human colon and rectal cancer. Nature 2014, 513, (7518), 382-7.

48. Ghazalpour, A., et al., Comparative analysis of proteome and transcriptome variation in mouse. PLoS Genet 2011, 7, (6), e1001393.

49. Nagaraj, N., et al., Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol 2011, 7, 548.

50. Joyce, A. R.; Palsson, B. O., The model organism as a system: integrating 'omics' data sets. Nature Reviews Molecular Cell Biology 2006, 7, (3), 198-210.

51. Ghazalpour, A., et al., Comparative Analysis of Proteome and Transcriptome Variation in Mouse. Plos Genetics 2011, 7, (6).

52. Ippolito, J. E., et al., An integrated functional genomics and metabolomics approach for defining poor prognosis in human neuroendocrine cancers. Proceedings of the National Academy of Sciences of the United States of America 2005, 102, (28), 9901-9906.

53. Zhao, Y. L., et al., Proteome-transcriptome analysis and proteome remodeling in mouse lens epithelium and fibers. Experimental Eye Research 2019, 179, 32-46.

54. Gygi, S. P.; Rochon, Y.; Franza, B. R.; Aebersold, R., Correlation between protein and mRNA abundance in yeast. Molecular and Cellular Biology 1999, 19, (3), 1720-1730.

55. Cox, B.; Kislinger, T.; Emili, A., Integrating gene and protein expression data: pattern analysis and profile mining. Methods 2005, 35, (3), 303-314.

56. Mootha, V. K., et al., Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. Cell 2003, 115, (5), 629-640.

Curriculum Vitae

Bohui Li was born in Guizhou, China. She obtained her bachelor's degree from Northwest Agriculture and Forestry University (NWAFU, 西北农林科技大学) in 2013, majoring in Bioscience. While conducting scientific training in Wang’s lab at the Center of Bioinformatics, NWAFU, she developed an interest in bioinformatics. Consequently, she decided to pursue a master's degree in the field of bioinformatics. During her master, Bohui focused on understanding the mechanisms of traditional Chinese herbal medicine using systems pharmacology approaches. In 2015, she was awarded the National Scholarship for Postgraduates, and in 2016, she was honored as an Outstanding Graduate Student (优秀毕业生) of the College of Life Sciences and got her Master of Science degree at NWAFU.

In October 2016, Bohui was awarded a PhD scholarship from the China Scholarship Council (CSC, 国家留学基金管理委员会) to pursue her doctoral study in the Netherlands. This scholarship sponsored her to study as a PhD candidate at the Department of Biomolecular Mass Spectrometry and Proteomics under the supervision of Bas van Breukelen and Maarten Altelaar. During her PhD, Bohui focused on analyzing mass spectrometry proteomics data using bioinformatics methods and machine learning algorithms.

Publications

This thesis

Li B, Kong XJ, Post H, Raaijmakers L, Peeper D, Altelaar M. Proteomics and Phosphoproteomics Profiling of Drug-Addicted BRAFi-Resistant Melanoma Cells. Journal of Proteome Research, 2021, 20, 9, 4381–4392.

Li B, Vredevoogd D, Peeper D, Altelaar M. Exploring the role of PD-1 in CD8+ T cell activation by proteomic and phosphoproteomics profiling. (*Under revision*)

Li B, Altelaar M, Breukelen B. Identification of protein complexes by integrating protein abundance and interaction stoichiometries using a deep learning strategy. International Journal of Molecular Sciences 2023, 24(9), 7884.

Other publications

Rontogianni S, Synadaki E, **Li B,** et al. Proteomic profiling of extracellular vesicles allows for human breast cancer subtyping. Communications biology, 2019, 2(1): 1-13.

Li B#, Xu X#, Wang X, Yu H, Li X, Tao W, Wang Y, Yang L, A systems biology approach to understanding the mechanisms of action of Chinese herbs for treatment of cardiovascular disease. International Journal of Molecular Sciences, 2012, 13(10):13501-20.

Li B#, Tao W#, Zheng C, Shar PA, Huang C, Fu Y, Wang Y. Systems pharmacology-based approach for dissecting the addition and subtraction theory of traditional Chinese medicines: an example using Xiao-Chaihu-decoction and Da-Chaihu-decoction. Computers in Biology and Medicine, 2014 Oct; 53:19-29.

Tao W#, **Li B**#, Gao S, Bai Y, Shar PA, Zhang W, Guo Z, Sun K, Fu Y, Huang C, Zheng C, Mu J, Pei T, Wang Y, Li Y, Wang Y. CancerHSP: anticancer herbs database of systems pharmacology. Scientific Reports, 2015, 5. doi: 10.1038/srep11481.

Tao W, Xu X, Wang X, **Li B**, Wang Y, Li Y, Yang L, Network pharmacology-based prediction of the active ingredients and potential targets of Chinese herbal Radix Curcumae formula for application to cardiovascular disease. Journal of Ethnopharmacology, 2013, 145(1):1-10.

Ru J, Li P, Wang J, Zhou W, **Li B**, Huang C, Li P, Guo Z, Tao W, Yang Y, Xu X, Li Y, Wang Y, Yang L. TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. Journal of Cheminformatics, 2014, 6(1):13.

Wang X, Xu X, Li Y, Li X, Tao W, **Li B**, Wang Y, Yang L, Systems pharmacology uncovers Janus functions of botanical drugs: activation of host defense system and inhibition of influenza virus replication. Integrative Biology, 2013 Jan 28; 5(2):351-71.

Li P, Chen J, Wang J, Zhou W, Wang X, **Li B**, Tao W, Wang W, Wang Y, Yang L. Systems pharmacology strategies for drug discovery and combination with applications to cardiovascular diseases. Journal of Ethnopharmacology, 2014, 151(1), 93-107.

Acknowledgements

First of all, I would like to thank my supervisors, **Maarten Altelaar** and **Bas van Breukelen**. **Maarten**, thank you so much for including me working in your group. I appreciate the knowledge that I have gained. You are always been supportive, helpful, and calm. Your positive feedback always cheers me up, your valuable advice in our discussions always points me in the right direction in science, and thanks for your patience and wise suggestions on reviewing my writing. I also appreciate the relaxed research environment, thanks for giving me enough freedom during my PhD. **Bas**, my great gratitude is given to you as well. Your smile encouraged me to do my PhD abroad in the Netherlands when we first met in 2016, and I enjoyed working with you. You have helped me a lot in the past years, not only in scientific but also in personal issues. You are also very approachable and supportive, whenever I come to you for help, you are always smiling and helping me with your valuable suggestions.

Albert, thank you for the opportunity to work in such a great lab. It was a great pleasure to do my PhD in such an awesome lab surrounding by great colleagues and scientists, I felt super lucky to work and befriend so many talented people.

Corine, you are the first person I know in the Netherlands, thank you for organizing our interview, although it was a cloudy and rainy winter day, I feel super warm in my heart because of your detailed schedule and kind help. Many thanks to you as well for all the help during my PhD, from housing to living. I would like to acknowledge my officemate **Henk**. It was nice to discuss with you on some bioinformatics issues and papers, you also gave me many suggestions at the beginning of my PhD, thank you very much. **Harm**, thanks for your contribution to the drug-resistance project. You are such a nice person, thank you for answering all my questions and helping me with the Dutch documents.

Dominique, I am lucky that you say Hi in my office and invite me to join the lunch team. We had a great time sharing happiness and upset. I hope you enjoy your job and have a happy life.

Many thanks go as well to my lovely Z606 office girls and friends, **Matina**, **Donna**, **Charlotte**, and **Marie**. It was a wise decision to move to this ‘active wing’ and this lovely office, again, thanks Corine for the arrangement. **Matina**, you are so sweet in our office, I sometimes bother you with some naïve questions, but you always encourage me ‘that’s ok, go for it, you will make it’, these encouraged me a lot, thank you so much. **Donna**, we have had a super time sitting behind the screen to chat about our cultural difference, fruits (crunchy vegetables like celery :), and food when we were trying to escape from data analysis or writing. Also, thank you for preparing the ‘baby shower surprise party’, it was a special memory for me and my son. **Charlotte**, I am lucky to work with you, you are so smart and sweet. Every time I disturb you with questions, you always stop your work and explain to me. **Marie**, thanks for your chocolate licorice :), I am not a fan of licorice, but I still remember the first time you encouraged me to taste one, I tried but failed :).

David, a big thanks to you as well. You are a so sweet person. We had an unforgettable time, thank you for the chatting and for listening to all my struggles. I wish you all the best in your country and hope to meet you in life. **Nadine**, how awesome that we can celebrate our birthday together, you are an expert on making the different flavors of cakes and desserts, thanks for the delicious sharing. **Kelly S**., I am lucky to work with you on some parts of your project, you are so smart in catching the key point of a project, and I have learned a lot from you. **Kelly D.**, it was great to work with talented people like you, like the defense board said: "you have made a book". I wish you all the best.

Weiwei, you are such a smart and working hard girl. Your expert skills are not only in experimenting but also in cooking, I doubt that you "steal" the experimental protocol for the cooking :). We have had many great moments in the group, however, the corona crisis has stuck us from home. I hope you enjoy the science work and hope to meet you in China. **Fujia**, I am very lucky to know you before I left the Netherlands, thank you for the cool gifts for Yicheng, I wish you and **Lele** all the best, and enjoy the science work, at the same time have fun in the Netherlands.

During the past years, I have had the opportunity to work with many great people. Thank you **Arjan, Mirjam, Harm, Soenita, Geert, Ceri, Pieter, and Dominique.** Without your great contributions, I never have chance to explore the research questions. Many thanks also go to **Mao, Wei, Franziska, Juan, Suzy, Anna R, Linsey, Tim, Julia, Saar, Simone, Richard, Gadi, Barbara, Vojta, Albert B., Karli, Wouter, Oleg, Sem, Tobi, Jing, Miao, Elmo,** and many other students and postdocs.

在荷兰读博士的生活漫长而又有趣，非常感谢在此期间认识的小伙伴们。我最最亲爱的**刘欣**，意诚的欣妈妈，是什么样的缘分才让我们在荷兰相识，成为好朋友。我们一起谈天说地，一起旅游徒步，一起吐槽枯燥和委屈，也分享开心与喜悦。希望我们友谊长存，待我们白发苍苍也还能一起分享各自的喜悦。**石涛**和**小刚**，非常开心能和你们成为好朋友。我们的每一次相聚都非常愉快，一起谈天说地分享美食，每一次都能聊到很晚才依依不舍的回家。希望将来我们还能在一个城市，那样我们又能串门聊天了。**姜玲蕾**师姐，乐观开朗又坚强勇敢的你总能给我带来正能量，很幸运在博士期间能够和你成为朋友。**静文**，你是我佩服的博士妈妈，非常感谢你给我分享的育儿经验。在一段时间里，你几乎成了我们家的家庭医生，每次有问题都咨询你。希望你早日学成归来。

我的妈妈团们，**单丽玲**和**杨超**团队，很开心能在怀孕期间就认识你们，一起散步和等待小生命的降临。后来又一起分享育儿经验和打折屯尿布消息。希望你们在荷兰的生活越来越好，也希望安东和艾琳兄妹健康成长。**莫翠萍**和**杜杰**团队，非常高兴能认识你们和小乐知，乐知妈妈绝对是超人妈妈，又能上班带娃，又会拍美照做美食，传说中的上得厅堂下得厨房就是你了。**宋阳**和**袁怀洋**夫妇和你们可爱的能能和玥玥，很高兴认识你们，从你们那里学习了很多孩子学习方面的东西，你们家孩子还这么小就读了那么多书，真正的读了万卷书又行了万里路。**范飞娟**和**Danny**夫妇，你们是我认识的唯一的异国夫妇，谢谢你们的美食分享。希望你们的两个小宝宝健康成长，也希望将来有机会能再相聚。

此外，感谢我的好友们对我的鼓励和帮助。**赵蕾**，谢谢你飞过来看我们一家，祝你们一家在瑞典生活越来越幸福。**清涤**，我们也算是有缘人了。大学是老乡，同在西农上大学，读博期间更是同在欧洲求学。非常开心我们的郁金香之旅，希望你事业蒸蒸日上。**文娟**师姐，谢谢你在我读博士期间对我的鼓励，每一次和你聊天，你都能带给我满满的正能量。希望你们一家三口，和将来的一家四口能一直幸福。**三毛**和**思慧**，祝你们一家三口越来越幸福。**俊哥**和**波波**，祝俊哥早日学成，也祝你们一家一直幸福。**余意**，在格罗林根和乌特勒支的相聚都让我记忆深刻，祝你事业蒸蒸日上。

在读博士期间，很高兴认识了很多志同道合的朋友们。翟亚楠和许泉夫妇，祝你们学业成。桂天书，余博，其乐木格，很高兴认识你们，和你们的每一次相聚都非常愉快。王维坊，熊武，张宏，黄润月，陈娜，顾朱杰，武文杨，陈婕，刘小松，谢媛，钟蔚，虞楚箫，胡洁，张杰，邹杨，很高兴认识你们。

此外，我非常感激我的父母和家人。儿行千里母担忧，自从上高中求学开始便离开父母。相距的距离也从高中的几十公里，到大学的一千多公里，再到出国求学的上万公里。父母亲的衰老也从一根白发到数不清的白发，这些白发可能是父母对于子女思念的象征，不善于表达的您们将嘱咐和思念藏在心里，这让我能够更勇敢和大胆的前行。感谢您们吃苦耐劳的精神锻造了我坚毅的性格，正是这股坚毅的劲在我求学遇到困境时才能做到坚持不懈。姐姐姐夫，你们总能做到无条件支持我的选择，从求学到结婚生子，你们都是我最坚强的后盾。你们身上的品质，质朴的、勤劳的、孝顺的，是很多人活了一辈子都没有参透的领悟。希望你们生活幸福，希望孩子们学业有成，健康成长。哥哥嫂嫂，希望你们和珺宇越来越幸福，也祝侄子学业有成。春波，虽然你总让家里人头疼，但是在大是大非的事情上面都处理的非常好，我相信你肯定没问题的，一切都会越来越好。希望你们一家生活幸福，两个小侄女健康成长。芳芳，希望越来越勇敢、坚强的你生活幸福，希望紫嫣学业有成，健康成长。非常感谢哥哥姐姐和弟弟妹妹们在我远嫁外省和求学在外时对父母的陪伴和照顾，无论我走到哪里，你们和父母亲都是我最长情的牵挂。

我肯定我是幸运和幸福的，因为我遇到了世界上最好的公婆。感谢我亲爱的公婆培养了这么优秀的卫阳，结婚以来，公公婆婆待我像亲生女儿一样，照顾我们的生活无微不至，同时还能做到毫无怨言。在我生孩子时，公婆和妹妹远赴荷兰照顾了产后的我和刚出生的意诚三个月，期间妹妹还担任了采购员，踩着自行车去各个超市给我们采购生活用品。您们的恩情是我无以为报的，希望公公婆婆身体健康。昭阳，希望你和王锴工作顺利，生活幸福，也祝意洵健康快乐成长。叔叔婶婶，谢谢您们对卫阳的指导和照顾，我们结婚后，谢谢叔叔婶婶对我们小家的照顾和爱护。希望您们工作顺利，生活幸福，祝子阳弟弟学业有成。此外，希望爷爷的老年活动（看体育比赛、和老友切磋牌技、种菜）能更加丰富，祝您身体健康。

卫阳（Weiyang）,缘分让我们相遇、相知、相爱。你总是督促着我进步，很开心能和你一起努力，一起进步，把生活过得越来越好。希望即将赴美的你能披荆斩棘，大展宏图，但是一定记得，累了烦了，我和意诚都会伸开双臂给你一个大大的拥抱，我们永远是你的后盾和港湾。意诚（Yicheng）,我最最亲爱的天使宝宝，你的到来给我带来了数不尽的欢乐和笑声。虽然在读博期间，压力很大，但是每次看到你天真的笑脸和小小的进步，都能吹散我所有的阴霾。妈妈希望你能保持童真和好奇，聪明活泼，健康快乐的成长。

最后，祝大家身体健康、工作顺利、万事如意!