BIOCHEMICAL PROFILE-BASED COMPUTATIONAL INFERENCE OF PROTEIN COMPLEXES

by

Zhongming (Lucas) Hu

A thesis submitted in conformity with the requirements

for the degree of Doctor of Philosophy

Department of Molecular Genetics

University of Toronto

Biochemical Profile-based Computational Inference of Protein Complexes

Zhongming (Lucas) Hu

Doctor of Philosophy

Graduate Department of Molecular Genetics

University of Toronto

2020

Abstract

Protein complexes are key macromolecular machines of the cell, but their description remains incomplete. Our group and others previously reported an experimental strategy for global characterization of native protein assemblies based on chromatographic fractionation of biological extracts coupled to precision mass spectrometry analysis (chromatographic fractionation–mass spectrometry, CF–MS), but the resulting data are challenging to process and interpret. In this thesis, I describe EPIC (elution profile-based inference of complexes), a software toolkit for automated scoring of large-scale CF–MS data to define high-confidence multi-component macromolecules from diverse biological specimens. The software toolkit EPIC is "plug-and-play", connects to public repositories for automatic data processing, and can be adopted productively to explore the network biology of any model system with little computational expertise required. The optimized CF-MS pipeline and EPIC data analysis workflows described in this thesis can be used to study different biological specimens, including diverse model organisms, to chart protein complexes on a global scale to expand our knowledge of macromolecular networks and their association with physiology, development, evolution and disease. Beyond providing a powerful framework to interpret CF-MS data, as a case study, I used EPIC to map the global interactome of *Caenorhabditis elegans* (WormMap), an important genetic model, comprising 612 putative multi-protein complexes linked to diverse biological processes. These encompassed new subunits for previously annotated protein complexes as well as novel assemblies seemingly unique to nematodes that we verified using stringent benchmarking criteria as well as by independent orthogonal affinity-purification mass

spectrometry validation experiments. To my knowledge, this is the first biochemically-based large-scale map of nematode protein complexes, which provides a rich platform for hypothesis-driven mechanistic investigations of animal biology. The major two outcomes of this dissertation consist of a tool (EPIC) and a knowledgebase (WormMap), which should serve as lasting resources for the broader biological research community.

Acknowledgments

Just as the saying “Pass the parcel. Take it, feel it and pass it on” states, the civilization of human beings is all about passing the torch of knowledge on to following generations and the continuous development of science needs efforts and input from many generations of scientists.

I would like to thank my supervisors Dr. Andrew Emili and Dr. Gary Bader for their generous support and guidance during my PhD study. I still remember, six years ago I joined their groups as a chemistry undergraduate student with very limited knowledge of coding and mathematics. Undoubtedly, it was extremely challenging for me to pursue system or computational biology research. During the past six years, my supervisors gave me so much freedom to learn new knowledge and apply these new skills into my own research. The whole journey is frustrating and tiring with numerous nights and weekends sitting in Gerstein library to try to figure out all the mathematics and algorithms, but the journey is rewarding in the end. I grew up from a chemistry student hardly solves any multi-dimensional integration to an interdisciplinary computational biologist can confidently read NIPS, ICML or many other statistics/AI journal papers. My supervisors lead me into this new world with so many exciting ideas and cutting edge research going on. I also would like to express my thankfulness to Dr. John Parkinson, who served as my supervisory committee member for tracking through my entire PhD study in the past six years and providing critical suggestions during every committee meeting.

The thesis work contains efforts from many collaborators. I would like to thanks Dr. Florian Goebels for his expertise and input on the software engineering. I would also like to thank Dr. June Tan from Dr. Andy Fraser lab for her expertise on worm transgenic line generation, which helps push part of this thesis to get published in Nature Methods. I would like to thank all the members from Emili and Bader lab, especially Eric Wolf for his help on AP-MS experiment and being such a good lab mate/friend. Thanks Sadhna Phanse, Uros Kuzmanov, Ruth Isserlin, John Giogi and Changjiang Xu for their technical assistance and many insightful conversations.

Finally, I would like to thank my parents for everything they did for me, both emotionally and financially. I wouldn’t be me without their tremendous and unconditional support.

I was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC)

Mass Spectrometry-Enabled Science and Engineering (MS-ESE) program since my second year of my graduate school.

Yes, I felt the parcel and hopefully can pass it on in the future.

Table of contents

List of Tables

List of Figures

List of Abbreviations

Accuracy (Acc)

Artificial intelligence (AI)

Affinity-purification mass spectrometry (AP-MS)

Area under the PR curve (auPR)

Area under the ROC curve (auROC)

Calmodulin-binding peptide (CBP)

Co-fractionation-mass spectrometry (CF-MS)

Cystic fibrosis transmembrane receptor (CFTR)

Collision induced dissociation (CID)

Co-immunoprecipitation (Co-IP)

Data-dependent acquisition (DDA)

Data-independent acquisition (DIA)

Disease ontology identifier (DOID)

Expected Improvement (EI)

Electrospray ionization (ESI)

Expectation Maximization (EM)

False-discovery-rates (FDRs)

Gene Ontology (GO)

High performance liquid chromatography (HPLC)

Ion exchange high-performance chromatography (IEX-HPLC)

Interquartile range (IQR)

Infrared (IR)

Karush-Kuhn-Tucker (KKT)

Liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS)

Matrix-assisted laser desorption ionization (MALDI)

Mass-to-charge ratios (m/z)

Mutual information (MI)

Multilayer perceptron (MLP)

Maximum matching ratio (MMR)

Tandem mass spectrometry (MS/MS)

Nematode growth media (NGM)

Nature Language Processing (NLP)

Nuclear Magnetic Resonance (NMR)

Online Mendelian Inheritance in Man (OMIM)

Pearson correlation coefficient (PCC)

Pearson correlation coefficient plus noise (PCCN)

Protein-protein interaction (PPI)

Positive predictive value (PPV)

Precision-recall (PR)

Protein A of *Staphylococcus aureus* (ProtA)

Post-translational modifications (PTMs)

Rectified linear unit (ReLU)

Random forest (RF)

Receiver-operating-characteristic (ROC)

Stable-isotope labeling with amino acids in cell culture (SILAC)

Sequential model-based global optimization (SMBO)

Sensitivity (Sn)

Support vector machine (SVM)

Tandem affinity purification (TAP)

Tobacco Etch Virus (TEV)

Tandem mass tags (TMT)

Parzen Estimator Approach (TPE)

Weighted Cross Correlation (WCC)

Extracted ion currents (XICs)

Yeast two-hybrid (Y2H)

CHAPTER 1

Introduction

1 Introduction

1.1 Proteomics experiments generate fruitful data for charting molecular networks

In cells, proteins rarely carry out biological functions on their own, but rather physically interact with each other to form multi-protein complexes to facilitate or catalyze cellular processes. In the past century, biologists usually focused on one or a few genes/proteins in one experiment, which undoubtedly led to many important scientific discoveries. However, the big picture of biological systems was usually bypassed, thus the derived molecular mechanism models were often incomplete. Unbiased elucidation of global protein-protein interaction (PPI) networks, or protein complex maps, is an alternate strategy to understand biological systems, and has become a major research focus of the proteomics research community. The term "proteomics" refers to "the large-scale study of proteins". Thanks to recent mass spectrometry-based technology development in proteomics, by our group and many others, nowadays thousands of interacting proteins can be identified and quantified in a single experiment, providing unprecedented opportunities to map the intermolecular connectivity that underlies biological systems. In the following sections of Chapter 1, I provide a snapshot of relevant technical and conceptual advances in the functional proteomics field. I focus on two traditional large-scale experimental approaches to map protein-protein physical interactions in biological systems. In closing, I discuss how modern machine learning techniques can be integrated with a powerful proteomics pipeline to make large-scale molecular network charting more efficient and effective.

1.1.1 The development of mass spectrometry and its application in proteomics

Protein mass spectrometry was first introduced as a follow-up step of 2-D gel electrophoresis for performing protein identification by peptide mass fingerprinting (Celis et al., 1996). Since the availability of gene and genome sequence databases and the development of increasingly sophisticated protein identification algorithms, mass spectrometry has become an effective strategy

for analyzing complex proteomics samples. Below I describe the basic workflows commonly used for charactering polypeptide mixtures, namely "top-down" and "bottom-up" proteomics.

Top-down proteomics analyses aim to characterize intact proteins, which has great potential for studying protein post-translational modifications (PTMs) and protein isoforms (Aebersold et al., 2018). Coupled with four-dimensional chromatographic separation, a single top-down proteomics experiment can successfully identify around 1,000 proteins, encompassing previously undetected protein isoforms (Tran et al., 2011). On the other hand, bottom-up proteomics aims to identify proteins from peptides released after proteolytic digestion, usually by the use of the enzyme trypsin. Unlike top-down proteomics, this bottom-up 'shotgun' approach aims to characterize proteins in an indirect way by detecting the digested peptides and then mapping them back to their cognate mother proteins. Top-down method is still limited in terms of sensitivity and throughput, because separation, ionization and fragmentation of entire proteins are naturally much more challenging than with smaller peptides. Thus, bottom-up proteomics has been significantly more widely adopted for protein analysis by the international proteomics community.

Regardless of which proteomics approach is utilized, mass spectrometry represents a state-of-art "workhorse" for protein analysis in both academia and industry. In order to carry out such experiments, peptide/protein analytes must first be ionized and then measured in the gas phase. Electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI) are two efficient ionization techniques used in proteomics studies. MALDI and ESI are considered soft ionization methods, because they do not fragment macromolecules into charged particles. The advantage of gentle ionization makes the detection of complex biological molecules tractable. MALDI uses laser pulses to ionize samples from a dry matrix, and was historically favored for less complex peptide samples, such as excised gel bands containing a single polypeptide. For more complex samples, ESI is normally the preferred choice, as ESI-MS can be relatively easily coupled to a liquid chromatography-based sample separation system. The entire liquid-chromatography ESI-MS workflow is called "LC-MS". Since ESI-LC/MS has dominated proteomics research in recent years, I will only focus on this system in the following discussion.

After ionization, mass-to-charge ratios (m/z) of ions eluting from the LC system are recorded by a mass analyzer, which measures MS1 intensities. Ions from this step can be further selected,

fragmented and measured by a second mass analyzer, in an cyclic process called tandem mass spectrometry (MS/MS). In MS/MS mode, the ESI process is often coupled to an ion trap MS device to select precursor ions according to a chosen mass window, and collision induced dissociation (CID) is then performed afterwards to fragment precursor ions, then the resulting fragmented ion spectra produced by CID of selected precursor ions are subsequently recorded by a mass detector (Michalski et al., 2011). For example, in the popular Q-Exactive mass spectrometry instrument line, a Quadrupole plays the role of mass filter to select precursor ions for fragmentation and analysis by a high-field Orbitrap mass analyzer (Michalski et al., 2011). This whole process is called "LC-MS/MS". In addition to measuring mass-charge ratio, recording protein or peptide fragmentation spectra provides critical sequence information. Based on the sequence-dependent fragmentation behavior, MS/MS spectra can often be easily mapped in silico against protein sequence databases to identify proteins that encode these peptides using different computational algorithms. For instance, in the SEQUEST algorithm (Eng et al., 1994), theoretically fragmented mass spectra are first generated computationally from a reference protein FASTA formatted sequence database, and a similarity metric is then used to calculate the overlap between experimentally derived tandem mass spectra and theoretical spectra to figure out the most likely (best-matched) sequence. The identified peptides are then mapped back to the corresponding protein sequences. Nowadays, with the continuing development of precision mass spectrometry and improvements in database search algorithms, scientists can routinely and confidently identify almost 4,000 proteins expressed by the yeast proteome within one hour using an Orbitrap/Q-Exactive mass spectrometer (Hebert et al., 2014).

Besides protein identification, protein quantification is also important in proteomics research. There are two major approaches for determining protein abundance. The first approach involves stable isotope labeling, in which pairs of chemically identical analytes can be distinguished based on their different masses due to the incorporation of different isotopes. The mass spectrometry signal usually reflects the relative abundance ratio of two chemically identical analytes, which allows changes in protein expression in one or more test samples to be determined relative to the background level in a reference control specimen.

Categorically speaking, there are three commonly used types of stable isotope labeling techniques. The first one to gain wide adoption is to introduce isotope labels metabolically by culturing cells in a

medium that contains isotopically enriched nutrients. For example, in the seminal method of "stable-isotope labeling with amino acids in cell culture" (SILAC), cells are cultured in ^{13}C enriched 'heavy' medium, such that ^{13}C is metabolically incorporated into all newly synthesized proteins/peptides (Ong et al., 2002). Molecules labeled with exogenous ^{13}C are readily discerned based on their mass differential or shift relative to the natural, native ^{12}C isotopologue.

The second approach incorporates stable isotopes through enzymatic reaction. For example, the heavy ^{18}O isotope can be incorporated from ^{18}O-labeled water during proteolysis (Miyagi and Rao, 2007; White et al., 2009).

The third labeling approach is to introduce heavy isotopes through a chemical reaction. In this strategy, specific sites (e.g. amino acid side chains) or functional groups on the proteins/peptides are labeled by reactive isotopically-encoded reagents. For example, tandem mass tags (TMT) are commercially available multiplexing reagents for labeling amino groups on peptides. By incorporating different neutron tags that produce readily discernible reporter ions and advanced mass spectrometry, TMT multiplexing allows relative quantification up to 10 or 11 different conditions in a single mass spectrometry run (Werner et al., 2014). Due to the simplicity of this experimental set-up and multiplexing advantage, TMT labeling has become the most popular isotope reagent among all. Using labeling approaches, especially the advantage of multiplexing, mass spectrometry machine time can be dramatically reduced. Although in this thesis work only the label-free approach is utilized, labeling techniques based multiplexing experiments are definitely a promising future direction that can be incorporated into this thesis work.

Besides isotope labeling based quantification, "label-free" quantification is commonly performed in proteomics studies and is suitable for different purposes, including mapping protein interaction networks and complexes. There are two "label-free" metrics for protein quantification: the summed-up mass spectrometry precursor ion intensity of peptides (MS1 intensity) and the number of MS/MS spectral counts (MS2 spectral counts) acquired for a protein (Nahnsen et al., 2013). Research has shown that protein quantification determined by MS2 spectral counts generally agrees well with MS1 intensity measurement and is proportional to protein abundance (and other factors, such as protein length) (Park et al., 2008), while the sensitivity of the label-free quantification is only

slightly less reliable compared with that achieved by isotopic labeling quantification methods (Old et al., 2005).

In summary, the overall workflow of using bottom-up proteomics to analyze protein mixtures could be summarized into seven parts (Zhang et al., 2013): sample preparation, protein fractionation, protein digestion, peptide separation, LC-MS/MS, database searching, and quantification. In the next section, I discuss how various methods devised to identify PPIs using proteomics based experiments.

1.1.2 Proteomics data elucidates PPI network

Since the possibility of systematically identifying and quantifying proteins from biological systems through proteomics experiments, researchers have devised methods focused on generating global proteome maps of biological systems (Florens et al., 2002; Washburn et al., 2001). The task of mapping the protein architecture of biological systems itself remains undoubtedly important and technically challenging, but it is conceptually simple and leaves many biological questions unanswered. For instance, global proteome mapping is not sufficient to answer the molecular mechanisms of cellular functions, which is the most important for biology research. It is well known that most proteins physically interact with other proteins to carry out their functions within cells. Furthermore, many proteins associate stably to form multi-protein complexes, such as core components of cellular replicative machinery, to regulate or mediate cellular biological functions (Alberts, 1998). For instance, the human nuclear pore complex, composed of 34 different protein subunits, is considered as the largest macromolecular assembly in a cell that is responsible for regulating component exchange between nucleus and cytoplasm while also helping to stabilize the nuclear envelope (Rout et al., 2000).

Apart from playing an architectural role within a cell, many protein complexes function within biological systems through the coordinated activity of multiple enzymatic activities. For example, RNA polymerase II is a well-characterized multi-protein enzyme complex that is responsible for catalyzing the synthesis of mRNA precursors through the transcription of chromosomal DNA templates in eukaryotic cell systems (Gnatt et al., 2001; Orphanides et al., 1996). Many, but not all,

of these complexes are widely conserved across evolution, whereas others show a more restricted lineage distribution.

Although the importance of studying protein complexes is widely appreciated, knowledge of their composition is still lacking for many species, including human, and many protein assemblies remain to be mapped. How to utilize mass spectrometry-based experimental techniques to study PPIs or to map protein complexes in a high-throughput but unbiased manner remains a major challenging task for the proteomics community.

Traditionally, low-throughput experimental approaches have been developed to study specific PPIs. For instance, one of the most commonly used experimental techniques is co-immunoprecipitation (Co-IP), where the antibody is used to capture a target protein and its interacting cellular partners. Afterwards, washing steps are typically performed to eliminate unspecific contaminants. Independent validation experiments often consist of gel-separation followed by Western blotting to probe putative interactors using specific antibodies (Markham et al., 2007). Such low-throughput assays are useful for characterizing small systems, but are difficult to adopt as platforms for large-scale discovery projects due to the high cost of experiments.

Two widely applied high-throughput strategies for mapping PPIs on a large-scale are the yeast two-hybrid (Y2H) assay (Fields and Song, 1989) and affinity-purification mass spectrometry (AP-MS) (Rigaut et al., 1999). In the following sections, I will discuss methodologies of these two high-throughput PPI mapping approaches and their applications in biology studies.

1.1.2.1 Yeast two-hybrid assay

The yeast two-hybrid (Y2H) assay, a molecular genetic test first introduced in 1989, was designed to identify direct binary interactions between two proteins based on the measurement of reporter gene transcriptional activity (Fields and Song, 1989). Briefly speaking, in this assay, genes encoding two potentially interacting proteins are fused to two halves of a transcriptional activator (DNA binding domain, and a transactivation domain). The interaction of the two proteins reconstitutes the two domains to activate the transcription of a reporter construct, leading to the preferential growth of the

yeast strain or expression of some colorimetric enzyme (**Fig. 1-1**). The Y2H system provides a versatile platform to screen hundreds of binary PPIs, and has been applied to map global PPI networks in different species (Arabidopsis Interactome Mapping, 2011; Rolland et al., 2014; Rozenblatt-Rosen et al., 2012; Simonis et al., 2009). However, the limitations of this assay exist; for example, fusion to another protein might alter the biochemical properties of the two proteins and two proteins interacting in the yeast nucleus does not necessarily mean they interact in their native cellular environments. Meanwhile, the yeast two-hybrid assay provides limited information about the composition of multi-protein complexes since it fails to capture indirect PPIs. In addition to false negatives, a major criticism of the Y2H assay is that it suffers a high false positive rate, which was estimated to be up to 50% using manually curated protein complexes as the reference dataset (von Mering et al., 2002). It has been suggested that the Y2H method provides reliable information about binary interactions involved in transient signaling and inter-complexes interactions rather than intra-complex interactions (Yu et al., 2008). (In the work presented in this thesis, I focus on the characterization of protein complexes, in other words, intra-complex PPIs; nevertheless, the terms of intra-complex (co-complex) PPIs and protein complexes will be used interchangeably.)

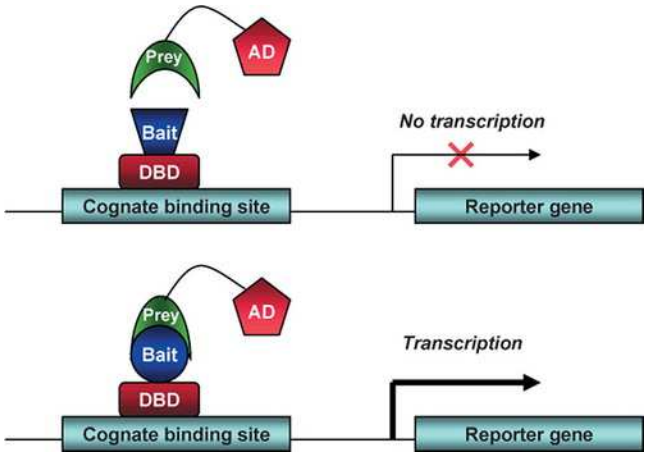

Figure 1-1: Yeast two-hybrid (Y2H) assay. *Adopted from (Ratushny and Golemis, 2008)*

1.1.2.2 Affinity-purification mass spectrometry (AP-MS)

AP-MS is a proven, and widely used high-throughput experimental approach to systematically isolate and characterize protein complexes and PPIs on a large-scale. The first demonstration of the power of APMS was introduced in yeast in 1999, which established its utility for detecting both known and unexpected PPIs in biological samples (Rigaut et al., 1999). Since then, this biochemistry-based PPI detection technique has been adopted by many research groups to map global protein complexes across different model organisms and biological systems (Butland et al., 2005; Guruharsha et al., 2011; Krogan et al., 2006). Lots of interesting biology can be drawn from these interactome maps. For example, in the yeast protein complex map, the authors noticed the protein Iwr1 co-purifies with RNA polymerase II, initiation factor TFIIF and transcription elongation factors Spt4/Spt5/Dst1. Further investigations confirm that Iwr1 is a conserved transcription factor.

The original AP-MS system used a dual tag system consisting of two IgG-binding units from protein A of *Staphylococcus aureus* (ProtA) and a calmodulin-binding peptide (CBP) linked by a Tobacco Etch Virus (TEV) endopeptidase cleavage sequence. Genetic engineering is required to fuse the dual tag system to target protein C-terminals, which can often be achieved by targeted integration of the tagging cassette into the native chromosomal context without disturbing endogenous expression levels. Cell extract or lysate containing the tagged target protein, together with associated interacting proteins, is first incubated with an IgG matrix (usually agarose or magnetic beads) that binds to ProtA. Then TEV protease was added into the sample to release the bound proteins. The eluate is then incubated with Calmodulin-coated beads that capture the CBP fusion protein in the presence of calcium ions (**Fig. 1-2**). The whole process is called tandem affinity purification (TAP) as two washing steps are involved (Rigaut et al., 1999). The eluted proteins can later on be analyzed by mass spectrometry. As described before, the consistent biochemical nature of the AP-MS approach is well suited to the study of soluble protein macromolecules. Indeed, in this thesis work, AP-MS was used as an independent orthogonal approach by which to validate protein complexes detected using an alternate workflow. However, the disadvantages of AP-MS are also obvious: adding tags to target proteins through genetic engineering is laborious, while altering the stability, folding or expression of target proteins might affect PPIs, and the extensive bead washing might remove weakly interacting partner proteins, limiting its screening coverage.

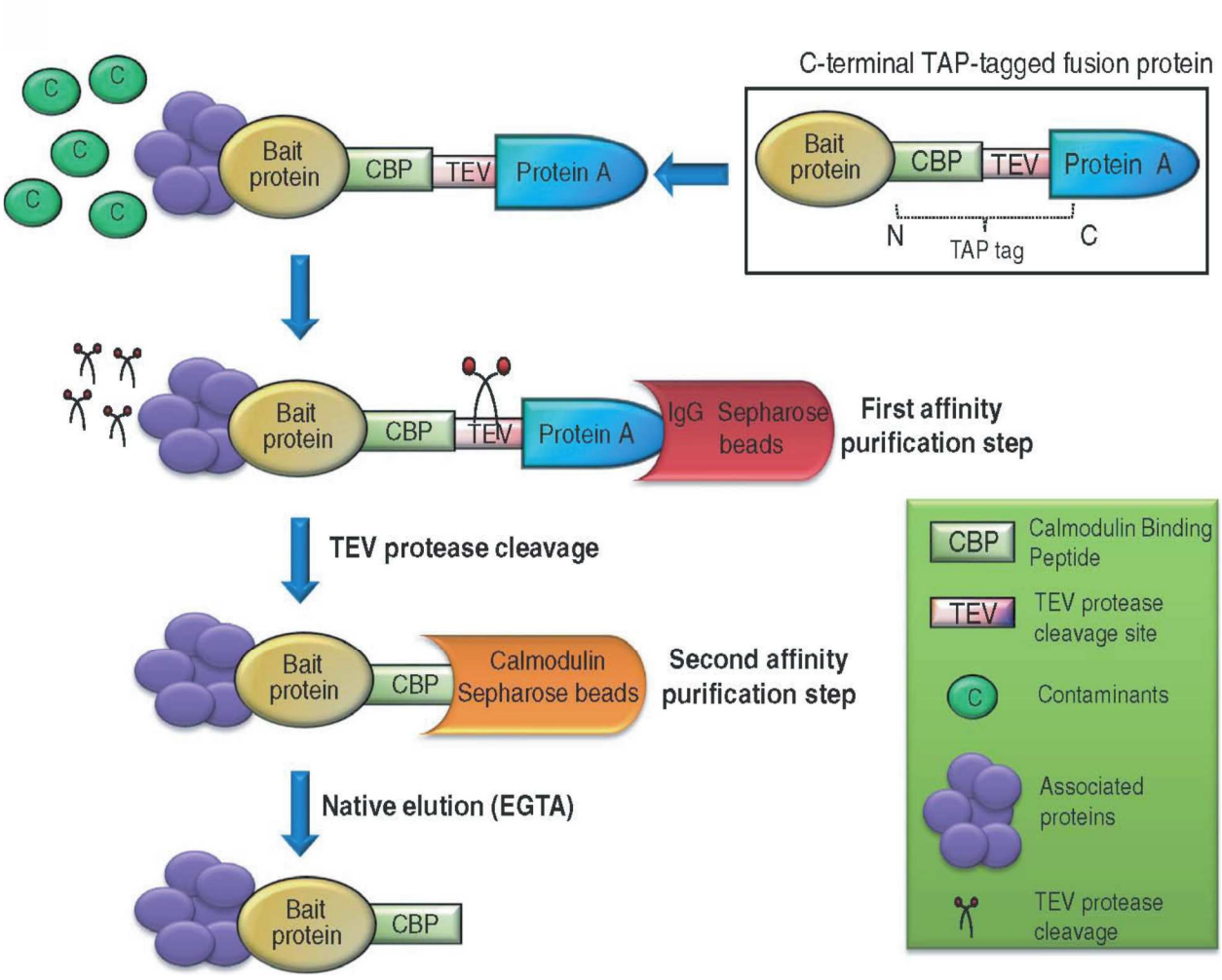

Figure 1-2: Tandem affinity purification (TAP). *Adopted from (Babu et al., 2009)*

1.1.2.2.1 Previous large-scale studies of protein complexes mapped using AP-MS

Since my thesis work is focused on mapping protein complexes, and since AP-MS has been used as an orthogonal approach to validate novel protein complexes from my predicted protein complexes map, I provide a concise review of previous efforts of utilizing large-scale AP-MS to map protein interaction networks in literature.

The large-scale AP-MS-based discovery of protein complexes, or global PPIs surveys, started from pioneering work in unicellular model organisms. The first bacterial protein complex map charted by AP-MS was generated for *Escherichia coli (E. coli)* (Butland et al., 2005). In this work, 857 *E. coli* proteins were tagged and purified. The final filtered network was found to contain many complexes that were predicted to be broadly conserved across prokaryotes. A later paper published in 2009 extended the original work to cover virtually the entire soluble proteome of *E. coli*, encompassing 5,993 PPIs (Hu et al., 2009). In addition to providing a global PPI map, this study also focused on making use of this physical interaction map to assign functions to un-annotated gene products based on the principle of guilt-by-association. Independent experiments were performed to verify the biological significance of certain predicted functions of previously unannotated gene products, ranging from biofilm formation and envelope assembly to protein synthesis and biomolecular replication. Apart from prokaryotes, AP-MS has been used to map protein interactions in eukaryotic organisms. To date, the most comprehensive PPI datasets in eukaryotic system community have been generated by applying AP-MS to the genetically tractable budding yeast *Saccharomyces cerevisiae* (Gavin et al., 2006; Krogan et al., 2006). In two papers co-published in 2006, researchers used tandem affinity purification to process 3,206 and 4,562 tagged proteins, from which 491 and 429 protein complexes were identified, respectively. However, the two maps are surprisingly different on both PPI and complex levels (Goll and Uetz, 2006). Further analyses demonstrate the differences of the two studies mainly come from two sources. First, these two studies sampled different sub-proteomes from yeast with only 1,152 proteins overlapping. Considering 3,033 and 2,701 proteins sampled in each study, the overlap on the sampled proteins is small (Goll and Uetz, 2006). Secondly, the two papers used different computational approaches on all steps from data pre-processing, PPIs scoring to complexes clustering (Goll and Uetz, 2006; Hart et al., 2007). This illustrates the importance of careful analysis and benchmark of large-scale proteomics data.

These microbial AP-MS mapping efforts have also been extended to study macromolecular assembly formation within cellular membranes. Membrane protein complexes are difficult to study due to the unique biochemical properties of cellular membrane proteins that hinder their extraction and affinity purification using traditional protocols. It was shown that using non-denaturing detergent, membrane system associated protein complexes can be identified by AP-

MS, and this strategy has been applied to both eukaryotic and prokaryotic envelope systems (Babu et al., 2018; Babu et al., 2012).

Later on, large scale AP-MS based PPI surveys have been expanded to multi-cellular model organisms. Notably, in 2011, the first global scale protein complexes map in *Drosophila melanogaster* was published, in which, 4,927 *Drosophila* proteins were tagged by FLAG-HA epitope and 556 protein complexes were identified (Guruharsha et al., 2011).

More recently, two large-scale AP-MS based studies of human the PPI network in cell culture models were published (Hein et al., 2015; Huttlin et al., 2015b). These studies used different human cell lines and tagging strategies (HeLa cell line vs. HEK293T cells; GFP tagged vs. FLAG-HA tagged) to define 28,504 and 23,744 PPIs, respectively. However, the two papers show limited overlap: among all 51,468 reported PPIs, only 758 PPIs are reported in both studies, which accounts for only 1.47% of all the reported PPIs from both papers. Apart from different cell lines and tagging strategies having been utilized in both studies, different computational pipelines might be the major reason for the discrepancy. Indeed, future detailed investigation is required before making any solid conclusion. The results of one of the projects, the BioPlex Network, was further expanded in a second recent update, in which more than 25% of protein-coding genes from the human genome were reportedly tagged and purified, forming the largest AP-MS based human PPI network to date (Huttlin et al., 2017).

Besides using AP-MS to map the global protein-protein network within model organism systems, scientists have applied AP-MS to probe specific biological systems. For example, AP-MS has been used to map protein networks centered on the human deubiquitinating enzyme and Cullin-RING ubiquitin ligase systems to study their mechanistic role in protein degradation in eukaryotic cells (Bennett et al., 2010; Sowa et al., 2009).

AP-MS has also been adapted to study disease modified protein networks. A nice example is the use of targeted AP-MS to map the impact of the ΔF508 mutation (a major cause of cystic fibrosis) on the interaction partners of the cystic fibrosis transmembrane receptor (CFTR) (Pankow et al., 2015).

Host-pathogen PPI networks have also been documented by AP-MS. For instance, a HIV-human protein complex map was first charted in 2012, in which all 18 proteins from HIV-1 were tagged and purified from two human cell lines (HEK293 and Jurkat). A putative network of 497 HIV-human PPIs were identified in this work (Jager et al., 2011). In a related work published more recently, scientists applied AP-MS to study PPIs between effector proteins expressed by mosquito-borne flaviviruses (dengue and Zika virus) and their host (human or mosquito) protein targets. This work identified many novel host-pathogen protein interactions, some of which were found to be important for virus infection (Shah et al., 2018).

AP-MS can be coupled with other quantitative mass-spectrometry techniques to answer more difficult biological questions. For example, using a sensitive data independent acquisition mass spectrometry strategy (sequential window acquisition of all theoretical spectra, or SWATH, AP-MS has been adopted to compare protein network dynamics under different conditions, potentially opening a new window to tackle differential protein network problems in biological systems (Collins et al., 2013; Lambert et al., 2013).

1.1.2.3 PPI detection through co-fractionation experiments

Mapping PPI network of a biological system from a global scale is important and informative. However, the current standard approaches for screening PPIs are tedious and laborious, which require tagging proteins individually (AP-MS) or in combination (yeast 2-hybrid). How to simultaneously isolate native protein assemblies to detect thousands of PPIs in a more physiological context without genetic engineering is a key goal for systems biologists. Historically, biochemists have used liquid chromatography to isolate and purify endogenous protein complexes (Boekema et al., 2001). The concept is simple; protein complexes consisting of strongly associated components are expected to co-elute with a similar retention time during liquid chromatographic separation. While traditionally most of the protein complex purification works have been focused on isolating individual protein complexes to near homogeneity, in 2012, two papers extended this idea to high-throughput study of macromolecules by collecting and analyzing multiple, relatively impure fractions using liquid chromatography separation systems (Havugimana et al., 2012; Kristensen et al., 2012).

In a proof of principle paper, human PPIs were identified by correlating protein elution profiles across a large set of biochemical fractions: a strong correlation supports a possible physical interaction between two or more proteins. Later on, this approach was extended to study multiple metazoan species to identify conserved protein complexes (Wan et al., 2015).

A key requirement of this co-fractionation-mass spectrometry (CF-MS) approach was to develop a rigorous computational strategy to reliably and confidently detect PPIs. The major focus of this thesis work was to develop a stringent computational platform to automatically score, predict and evaluate protein complexes using data collected from CF-MS experiments. Since the details of this method (both experimental and computational parts) have now been published and are reported at length in the following data chapters, the general approach to PPI scoring and filtering will only be described briefly here.

However, it is worthy to mention some existing tools have been developed by other groups to analyze this type of data. For instance, PrInCE written in MATLAB was developed by the Foster group (Stacey et al., 2017), in which the authors used the Expectation Maximization (EM) algorithm to fit a Gaussian mixture distribution to chromatographic peaks observed from elution profile. Then they used a Naïve Bayes based machine learning approach to predict protein-protein interactions from pre-processed elution profiles. The advantage of this approach is that by fitting Gaussian mixture distributions to chromatographic peaks, most noise from the elution profile could be eliminated. However, the fitting process requires at least five data points (five fractions), which eliminates weaker, but still useful peak signals detected by the CF-MS approach. Also, chromatographic peaks with skewed tails are not well modeled by Gaussian distributions. As a result, the Gaussian mixture modeling procedure is too stringent that might remove many real within-complex protein interactions. Another paper published by the Marcotte group (Drew et al., 2017) adapts a sophisticated sparse graphical model to infer direct protein interactions using the co-variation pattern of the protein abundances. They tend to skip the protein elution profile correlation to directly predict protein complexes. But this method lacks a software implementation and requires further experimental validation. Further, the lack of automation limits the usage of all existing tools: users need to manually curate multiple datasets from various sources and perform data pre-processing, which would be challenging for many proteomics labs.

1.1.3 Computational challenges in proteomics research

From identifying proteins using mass spectra data collected from high performance LC-MS/MS systems to scoring PPIs based on the results from high-throughput experimental techniques, proteomics researchers have stepped into the 'big data' era. Thus, the demand for better computational algorithms and robust statistical methods for interpreting interactomics data has been increasing. Computational improvements in the proteomics field mainly focus on two aspects. The first concerns how to improve the quality of data acquired using mass spectrometry instruments. Better protein searching algorithms can make protein identification and quantification faster and more accurate. Efforts are also centered on building more effective new algorithms and software tools to enhance instrument performance through technology development. Another angle is how to efficiently or smartly apply innovative computational approaches to extract meaningful biological information from the large amount of proteomics data collected by mass spectrometry. After stepping into the big data era, many different artificial intelligence (AI) techniques, especially many machine learning algorithms, have been injecting fresh ideas into proteomics research help us better understand biology from proteomics data in a highly efficient way. But before discussing some exciting machine learning approaches and their applications in proteomics, I briefly review some important computational protein searching algorithms/tools for proteomics, and some of them are used in this thesis work.

1.1.3.1 Computational tools in proteomics data acquisition

Proteomics research relies on mass spectrometry technology. However, a mass spectrometry machine outputs a spectrum of mass-charge-ratios (m/z) obtained from the injected samples. These need to be interpreted to convert them into molecule identifications. For simplified or purified samples, an experienced mass spectrometry analytical chemist can manually assign chemical structures based on the fragmentation pattern and isotopic evidence. Even so, this work is not an easy task: chemists sometimes require other evidence, usually Infrared (IR) and Nuclear Magnetic Resonance (NMR) spectroscopy, to determine functional groups and finalize chemical

structures. The chemical diversity of polypeptides is much simpler than the chemical space of small molecules, as there are only 21 proteinogenic amino acids and the peptide bonds are strictly formed by the carboxyl group and the amino group from two different amino acids. As a result, the fragmentation pattern for a polypeptide is relatively easy to predict, even with the extra post-modifications on peptides.

The first protein-searching algorithm is SEQUEST that was developed in 1994 (Eng et al., 1994). The idea of SEQUEST is that a cross-correlation function is defined to measure the similarity between the mass-to-charge ratios between fragment ions theoretically predicted from protein sequences and the one experimentally obtained through a tandem mass spectrometry experiment. The matched peptide is the theoretically predicted peptide that receives the highest similarity score. Since then, SEQUEST has been improved by several groups: to be more efficient (Diament and Noble, 2011; Eng et al., 2008), consider post-translation modification on peptides during protein search (Yates et al., 1995) and introduce single nucleotide polymorphism to nucleotide databases for amino acid variant recognition (Gatlin et al., 2000). Overall, SEQUEST has been a very impactful tool in proteomics community that greatly facilitated proteomics research, and many other useful tools are built on SEQUEST.

MaxQuant is a newer generation popular proteomics searching platform. MaxQuant was first introduced as a tool specifically designed for protein identification and quantification in stable amino acid isotope (SILAC) samples (Cox and Mann, 2008). Later on, this platform has been expanded to cover many new functions including the control of false-discovery-rates (FDRs), the analysis of post-modifications (PTMs), label-free protein quantification and incorporating more isotope labeling strategies (Tyanova et al., 2016). The new version of MaxQuant uses Andromeda as the peptide searching engine, although the general framework is very similar to SEQUEST, the detailed peptide matching function and the ways of generating theoretical fragment masses are different (Cox et al., 2011). A major advantage of using the MaxQuant platform to perform protein label-free quantification is that it provides both MS1 intensity and MS2 spectral count measures. To more accurately perform MS1 intensity quantification, the MaxQuant intrinsic label-free quantification algorithm MaxLFQ uses the extracted ion currents (XICs) ratio to select peptides with good quality for further MS1 protein quantification (Cox et al., 2014).

Another interesting computational improvement in protein searching is to combine results from multiple different search engines. The idea is simple: with good statistics, the protein identification coverage should be increased and the protein quantification should be more accurate if multiple search engines are used. MSblender is one of these platforms can statistically integrate results from multiple search engines (Kwon et al., 2011). MSblender use mixture of multivariate distributions to account for the correlation between different database search scores and convert them into one probability score for every possible peptide spectrum match. In this thesis work, we used MSblender as the major protein-search platform. Meanwhile, we compared the performance of identifying co-complex PPIs from CF-MS data using different search and quantification approaches from several other protein-searching tools.

Apart from the efforts mentioned above, many computational tools are also being developed to accompany new proteomics technology developments. For example, data-independent acquisition (DIA) is an emerging mass spectrometry technique, in which, all peptide precursors within predefined mass-to-charge ratio and retention time range are fragmented and all the tandem mass spectral information is stored for further analysis. Compared with traditional data-dependent acquisition (DDA) approach, DIA is considered to be more unbiased and reproducible, but its tandem mass spectra are much more complex with multiple co-fragmenting precursors, thus harder to analyze. Many new software tools have been designed to specifically process DIA mass spectrometry data (Navarro et al., 2016; Rost et al., 2016; Ting et al., 2017; Tsou et al., 2015), which has made DIA based proteomics more popular over the last few years.

1.1.3.2 Computationally analyzing proteomic data is challenging but rewarding

As discussed before, the development of many different new algorithms, software tools and novel technologies have made mass spectrometry based proteomics a robust, powerful and popular analytical science. From a pure analytical chemistry point view, robustness, sensitivity and efficiency are the major goals of proteomics. But on the biology side, once a large amount of data is collected, how to translate the data to help answer biology or clinical questions requires effort from both computational and experimental scientists. Many challenging and diverse questions are still left in this direction in the proteomics community. For example, scientists are

studying clinical cancer samples using proteomics and genomics hoping to use correlation and differences between genomic and proteomic data to prioritize the tumor driver genes (Sinha et al., 2019; Zhang et al., 2014). Other efforts in proteomics research include spatial proteomics mapping (Kislinger et al., 2006; Thul et al., 2017), mitochondrial proteomics dissection (Stefely et al., 2016; Williams et al., 2016; Wu et al., 2014), and proteome-scale thermo-stability analysis (Leuenberger et al., 2017). Many computational tools have been developed to analyze the proteomics data analysis in these studies. Since the major focus of my thesis work is to study the protein physical interaction network, I will give a more detailed review of how researchers have used different algorithms and models to study the protein network. One of the earliest computationally predicted protein networks was published in 1999 (Marcotte et al., 1999). In this paper, scientists noticed that many pairs of known interacting proteins have homologs in another organism fused into one single protein. Based on this observation, 6,809 and 45,502 putative PPIs were predicted in *E. coli* and yeast, respectively. The computational approach used in this work is pioneering as it demonstrates that protein networks can be predicted using gene sequence data. Unfortunately, it suffers from a high false positive rate. Another important computational PPI prediction method was published in 2003 (Jansen et al., 2003). The authors built a Bayesian network to integrate genomic features and previously experimentally determined PPIs to generate a more comprehensive protein network in yeast. This work demonstrates that data integration can be used to improve the accuracy and coverage of protein networks.

In addition to studying physical PPIs, scientists also build functional interaction networks that represent if genes or proteins are functionally related to each other. The first gene functional network was constructed in yeast (Lee et al., 2004). In this work, functional genomics data were integrated using a Bayesian approach. Each evidence source was scored by evaluating its ability to re-construct known biological pathways. The gene pairs from the known pathway are assigned a score calculated by the log-likelihood, conditioned on the functional evidence source. Then within each category of functional data (e.g. mRNA co-expression data), all experiments are integrated by a rank-weighted sum of log-likelihood score. After that, all functional genomic categories, such as mRNA co-expression, co-citation, AP-MS based protein network, are integrated by the rank-weighted approach again to derive the final gene functional network. This network can be used to predict gene function and for functional module detection. Functional

interaction networks have been developed for many other model organisms and continue to be updated as more functional genomic data becomes available (Lee et al., 2010a; Lee et al., 2011; Lee et al., 2008b; Lee et al., 2010b). Functional interaction networks are easy to access from various web search engines. For example, GeneMANIA is a software platform that uses ridge regression to integrate multiple sources of input networks to create an integrated functional interaction network (Mostafavi et al., 2008). Most functional interaction networks are created as a static database. The GeneMANIA algorithm is fast enough to perform the integration at query time, such that users can supply their own functional interaction data as input to the algorithm. STRING is another integrated network search engine with different evidence sources compared to GeneMANIA and coverage of a much wider range of organisms (von Mering et al., 2005). STRING uses naïve Bayes to integrate functional genomic data to predict PPIs trained on known biological pathways from KEGG. An advantage of GeneMANIA and STRING is they are applied to multiple species and all the functional genomic data of these organisms are stored in their databases and available for download. The EPIC software developed in this thesis work takes advantage of this feature to automatically download functional genomic data from these databases and integrate them with co-fractionation data to help predict PPIs (discussed in Chapter 2).

1.2 Machine learning meets proteomics in the AI era

As discussed above, proteomics technology introduced a new approach to study proteins in a large scale, and many protein interaction assays provide us powerful platforms to perform PPI screening in biological systems. Computationally how to infer, integrate and score PPIs using data collected from these assays is a challenge that machine learning can address. In this thesis, I used machine-learning classifiers to integrate functional genomic evidence with experimentally derived co-fractionation MS data to improve the prediction performance of 'within complex' PPIs. I also tried to expand the work by using deep neural network to predict PPI directly from raw co-fractionation data without performing feature engineering. Thus, in this section, I will give a brief introduction to the topic of machine learning, focusing on machine-learning algorithms used in this work.

1.2.1 Popular machine learning algorithms

In the following sections, I provide a short introduction to machine learning algorithms used in this thesis. I start with introduction of linear regression, the simplest supervised learning machine learning algorithm. Then I cover support vector machine and random forest methods, which I used in this thesis to predict PPIs. The algorithms of linear regression, support vector machine and feed-forward neural network are summarized from the book of *Pattern Recognition and Machine Learning* (Bishop, 2006). The algorithm of random forest is summarized from the book of *Machine Learning: A Probabilistic Perspective* (Murphy, 2012).

1.2.1.1 Linear Regression

Linear regression is one of the simplest supervised machine-learning algorithms, which is the foundation of many other complicated machine learning algorithms. The goal of it is to predict the value of one or more continuous target variables given the values of input variables stored in an *n* dimensional vector. A simple form of linear regression is a linear function of input variables, for example, polynomial fitting is a typical linear regression problem. To account for the non-linear nature of data, we can apply nonlinear functions (basis functions) to the input variables to transform the input data. The simplest linear regression model can be formulated as the linear combination of the input variables:

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + w_1 x_1 + \cdots + w_D x_D$$

In the formula above, $\boldsymbol{x}$ is the input data, which can be seen as the input vectors of training features. The key property of this equation is the parameter vector of $\boldsymbol{w}$, which can be obtained from training the model. The simple linear regression can be extended to consider non-linearity in input data by applying nonlinear functions to the input variables. Then linear regression can be formulated in the following way:

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \Sigma_{j=1}^{M-1} w_j \boldsymbol{\phi}_j(x_j)$$

In the equation above, w_0 is a fixed offset called the bias parameter of the model. Φ_j is the nonlinear function applied to the data that is known as basis function. To solve for $\boldsymbol{w}$, an error function that measures the difference between the value of function $y(\boldsymbol{x},\boldsymbol{w})$ and the target value t from the training dataset is minimized. So the goal is to minimize the function:

$$\frac{1}{2}\Sigma_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2$$

In this equation, x_n is the features of nth sample in the training set and t_n is the target value of the nth sample in the training set. The same result derived from minimizing the above function can also be achieved using a maximum likelihood approach under an assumed Gaussian noise model. The derived $\boldsymbol{w}$ is:

$$\boldsymbol{W_{ML}} = (\boldsymbol{\phi}^T\boldsymbol{\phi})^{-1}\boldsymbol{\phi}^T\boldsymbol{t}$$

In the formula above, $\boldsymbol{\Phi}$ is the $N\times M$ design matrix that can be expanded as:

$$\boldsymbol{\phi} = \begin{pmatrix} \phi_0(\boldsymbol{x}_1) & \cdots & \phi_{M-1}(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_0(\boldsymbol{x}_N) & \cdots & \phi_{M-1}(\boldsymbol{x}_N) \end{pmatrix}$$

The result derived above has a perfect fitting with the training data, but it suffers from poor prediction power on data not taken from the training set. This problem is called overfitting in machine learning. To overcome overfitting in linear regression, a regularization term is usually added to the error function. Then the error function becomes:

$$\frac{1}{2}\Sigma_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\lambda}{2}\boldsymbol{w^T w}$$

λ is the coefficient of regularization that controls the importance of the regularization term. The new cost function with added regularization term can be minimized to obtain the following result:

$$\boldsymbol{W_{ML}} = (\lambda\boldsymbol{I} + \boldsymbol{\phi}^T\boldsymbol{\phi})^{-1}\boldsymbol{\phi}^T\boldsymbol{t}$$

1.2.1.2 Random forest

Random forest is another popular machine learning algorithm used in this thesis work, which is known to have good prediction accuracy over a wide range of applications. The basic unit of a random forest is a decision tree (another machine learning method) that is defined by partitioning the input space recursively to define a local model of each individual region after partitioning. To represent the whole partitioning process efficiently, a tree data structure is used, by which the leaf of the tree structure represents the defined region after partitioning. Given the input data, how to optimally partition the data to grow a tree structure is NP-complete. The common practice of this process is to use a greedy procedure for local optimal maximum likelihood estimation. There are several popular ways to implement this, and they all use a split function to choose the best feature and the best value for that feature. The process can be formulated in the following equation:

$$(j^*, t^*) = \mathbf{arg} \min_{j \in \{1,\dots,D\}} \min_{t \in \mathcal{T}_j} cost\big(\{\boldsymbol{x}_i, y_i : x_{ij} \le t\} + cost\{\boldsymbol{x}_i, y_i : x_{ij} > t\}\big)$$

If we assume all inputs are real values, then we can compare a feature x_{ij} to a numeric value t. The thresholds $\mathcal{T}_j$ for feature j is usually defined by sorting unique values of x_{ij}. During the splitting process, multi-way splits (non-binary trees) are usually avoided, as this might result in too little data falling into each sub-tree that could lead to overfitting. Several stopping criteria can be used to determine whether a node will be split or not: the cost function reduction is too small; the depth of the tree exceeds the maximal value; the distributions of two sub-trees are sufficiently homogeneous; the number of data points in each sub-tree is too small. It is important to specify a cost function that helps decide a proposed node splitting and the choice of the cost function depends on whether the goal is regression or classification. In the regression set-up, the cost function is usually defined as:

$$cost(\mathcal{D}) = \sum_{i \in \mathcal{D}} (y_i - \bar{y})^2$$

In the formula above, $\bar{y} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} y_i$ stands for the mean response of the variables in the specified set of data. The other set-up for random forest is classification, in which, there are

several ways to measure the splitting quality. For the data in the leaf satisfying a test condition (i.e. $x_j < t$), a multinoulli model could be fit to estimate the class-conditional probabilities:

$$\hat{\pi}_C = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} \mathrm{I}(y_i = c)$$

In the formula above, $\mathcal{D}$ is the data in the leaf. Several different metrics (e.g. misclassification rate) can be used to measure the common error during tree partitioning. In this case, the most probable class label is defined as:

$$\hat{y}_C = argmax_c\, \hat{\pi}_c$$

Using the result above, the error rate can be calculated as:

$$\frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} \mathrm{I}(y_i \neq \hat{y}) = 1 - \hat{\pi}_{\hat{y}}$$

Another popular error measure is entropy, which can be defined as:

$$\mathrm{H}(\hat{\boldsymbol{\pi}}) = -\sum_{c=1}^{C} \hat{\pi}_c\, log\hat{\pi}_c$$

Minimizing the entropy is same as maximizing the information gain. In other words, given a splitting test $x_j < t$ and class label Y, the information gain can be defined as:

$$\begin{aligned}\mathrm{infoGain}(x_j < t) &= H(Y) - H(Y|x_j < t) \\ &= \left(-\sum_c p(y=c)\log p(y=c)\right) + \left(\sum_c p(y=c|x_j<t)\log p(c|x_j<t)\right)\end{aligned}$$

The Gini index can also be used to measure the error, which is defined as:

$$\sum_{c=1}^{C} (\hat{\pi}_c)(1-\hat{\pi}_c)$$

The Gini index is basically the expected error rate, as $\hat{\pi}_c$ is the probability that a random value in a leaf belongs to class c, and $1 - \hat{\pi}_c$ is the probability that misclassification happens.

Once a full tree is grown, governed by the splitting error measurements defined above, pruning is usually performed to prune the branches contribute the least error increase to prevent overfitting. The common practice of reducing variance of an estimate is to perform many estimates and compute the average. This is how random forest developed based on decision trees. In this case, many different trees on different subsets of the data are trained to get an ensemble, where the subsets of the data are chosen randomly with replacement. The ensemble can be calculated using the following formula:

$$f(\boldsymbol{x}) = \sum_{m=1}^{M} \frac{1}{M} f_m(\boldsymbol{x})$$

In the formula above, f_m is the mth individual tree. One issue of the ensemble approach is that predictors might be highly correlated with each other after re-running many times on different subsets of the input data, thus the initial goal of reducing variance is limited. The random forest model is implemented to eliminate this weakness by de-correlating the base learners. It tries to learn trees by using randomly chosen subset of input variables and randomly chosen subsets of data cases.

1.2.1.3 Support vector machine

Support vector machine (SVM) is a type of kernel-based machine learning. This machine-learning algorithm is attractive due to its explainable mathematics, as the model training process can be treated as a convex optimization problem, thus the local optimum is always the global solution. It is also good at handling sparse, high dimensional data. The mathematics of SVM is elegant with an analytical solution. To understand how this algorithm works, let's first define a two-class classification problem using a linear model:

$$y(\boldsymbol{x}) = \boldsymbol{w}^T \Phi(\boldsymbol{x}) + b$$

In the formula above, $\boldsymbol{w}$ is the parameter vector, $\boldsymbol{x}$ is the input vector, Φ is the transformation in the feature space and b is the bias parameter. The training matrix has input vectors: $x_1, \ldots, x_N$, and each vector has associated target values: $t_1, \ldots, t_N$, in which, the value of t_i is either 1 or -1. A new data point can be assigned to a class based on the sign of $y(\boldsymbol{x})$. To explain how the SVM algorithm works, let's first assume all data points in the training data set are linearly separable in the feature space, thus there exists at least a set of values of parameter $\boldsymbol{w}$ and b to make all training data points clearly distinguished into two groups, which means $y(\boldsymbol{x}_n) > 0$ for all points with the target value is 1 and $y(\boldsymbol{x}_n) < 0$ for all points with the target value is -1. There may be many values of parameter $\boldsymbol{w}$ and b that can make the separation possible, in which case the optimal solution would be the ones that give the smallest generalization error. SVM uses the concept of a margin to approach this classification problem. The margin is defined as the smallest distance between a decision boundary and any sample in the training set. The SVM algorithm aims to find the decision boundary that maximizes the value of margin. On the decision boundary, $y(\boldsymbol{x}) = 0$, thus this decision boundary is also called a hyper-plane. If we set the linear model to have the form of $y(\boldsymbol{x})$, the perpendicular distance of a point $\boldsymbol{x}$ from the hyper-plane is given by $|y(\boldsymbol{x})| / ||\boldsymbol{w}||$. For all correctly classified points, $t_n y(\boldsymbol{x}_n) > 0$ as discussed before. Thus, the distance of a point $\boldsymbol{x}_n$ to the hyper-plane is:

$$\frac{t_n y(\boldsymbol{x_n})}{\|\boldsymbol{w}\|} = \frac{t_n(\boldsymbol{w}^T \Phi(\boldsymbol{x_n}) + b)}{\|\boldsymbol{w}\|}$$

The margin is defined as the perpendicular distance to the closest data point $\boldsymbol{x}_n$ from the training data set. The goal of the SVM algorithm is to maximize the distance by optimizing the value of $\boldsymbol{w}$ and b. This can be represented as:

$$\arg\max_{\boldsymbol{w},b} \left\{ \frac{1}{\|\boldsymbol{w}\|} \min_n [t_n(\boldsymbol{w}^T \Phi(\boldsymbol{x_n}) + b)] \right\}$$

To directly solve this optimization problem is rather complicated. We could first rescale the $\boldsymbol{w}$ and b without changing the distance from any point $\boldsymbol{x}_n$ to the hyper-plane (as the perpendicular distance of a point $\boldsymbol{x}$ from the hyper-plane is given by $|y(\boldsymbol{x})| / ||\boldsymbol{w}||$). We could first set the point that is closest to the hyper-plane to have the following constraint:

$$t_n(\boldsymbol{w}^T \Phi(\boldsymbol{x_n}) + b) = 1$$

In other words, all data points in the training set will have the following property:

$$t_n(\boldsymbol{w}^T \Phi(\boldsymbol{x_n}) + b) \geq 1$$

This re-scaling procedure is known as canonical representation of the decision hyper-plane. If the equality is held, the constraints are called to be active, otherwise, they are said to be inactive. Clearly there is at least one active constraint, since there is always at least a closest point. Once the maximized margin is achieved, there are two active constraints (as there are two target values). Then the SVM optimization problem has been transformed to maximize $\|\boldsymbol{w}\|^{-1}$ that is equivalent to the optimization problem of:

$$\arg\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2$$

This is a typical optimization problem of minimizing a quadratic function subject to a set of linear inequality constraints. To solve this, Lagrange multipliers (a_n) are introduced to give the Lagrangian function as below:

$$\mathrm{L}(\boldsymbol{w}, b, \boldsymbol{a}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{n=1}^{N} a_n\{t_n(\boldsymbol{w}^T\Phi(\boldsymbol{x}_n) + b) - 1\}$$

By setting the derivatives of the above formula with respect to $\boldsymbol{w}$ and b to zero, we obtain the following equations:

$$\boldsymbol{w} = \sum_{n=1}^{N} a_n t_n \Phi(\boldsymbol{x}_n)$$

$$\boldsymbol{0} = \sum_{n=1}^{N} a_n t_n$$

By substituting the above equations into the Lagrangian function, we then get the following equation to maximize:

$$\tilde{\mathrm{L}}(\boldsymbol{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\boldsymbol{x}_n, \boldsymbol{x}_m)$$

In which, $\boldsymbol{a}$ is subjected to the constraints:

$$a_n \geq 0$$

$$\sum_{n=1}^{N} a_n t_n = 0$$

The kernel function is defined as:

$$k(x, x') = \phi(x)^T\phi(x')$$

It is evident that the kernel function is positive definite, so the Lagrangian function $\tilde{L}(\boldsymbol{a})$ is bounded below. According to Karush-Kuhn-Tucker (KKT) condition, the following three properties hold:

$$a_n \geq 0$$
$$t_n y(\boldsymbol{x}_n) - 1 \geq 0$$
$$a_n\{t_n y(\boldsymbol{x}_n) - 1\} \geq 0$$

Also, by substituting $\boldsymbol{w}$ obtained from Lagrangian function into the original linear model, we get:

$$y(\boldsymbol{x}) = \sum_{n=1}^{N} a_n t_n k(\boldsymbol{x}, \boldsymbol{x}_n) + b$$

The third property from the KKT condition tells us that for each data point, it is either $a_n = 0$ or $t_n y(\boldsymbol{x}_n) = 1$. But the above formula indicates that $a_n = 0$ will not participate in making a decision for new data points. All the data points associated with $a_n \neq 0$ are called support vectors and lie on the maximum margin hyperplane in the feature space as they have the property: $t_n y(\boldsymbol{x}_n) = 1$. Up to now, we have solved the optimization problem and found the value for $\boldsymbol{a}$. However, at the beginning of this section, we have assumed that all data points in the training space could be separated in the feature space. If misclassifications are not avoidable in the training data set, we introduce slack variables $\xi_n = |t_n - y(\boldsymbol{x}_n)|$ into the optimization problem, so we try to minimize the following term instead:

$$C \sum_{n=1}^{N} \xi_n + \frac{1}{2}\|w\|^2$$

In the formula above, $C > 0$ can be set to control the weight of slack variables. The following update procedures to solve this optimization are similar to the ones we described before.

1.2.1.4 Deep Learning

According to a review paper written by Y. LeCun and his co-workers, deep learning refers to "computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction" (LeCun et al., 2015). For many decades, traditional machine learning approaches required careful feature engineering to extract informative features from the raw data, and then the machine-learning models could learn, detect and classify patterns

based on the carefully engineered features. Deep learning approaches overcome this problem by learning representations directly from raw data without extensive feature engineering. With multiple layers of representation, deep learning models take the raw data as the input layer and compose simple but nonlinear modules to transform the raw data into successively more abstract representation in each level. Highly complex functions can be learned after enough such transformations. Deep learning based approaches have demonstrated superior performance that beat conventional machine learning methods in many areas including speech recognition (Hinton et al., 2012) and image recognition (Alex et al., 2012). In recent years, deep learning methods have been introduced into biology studies and achieved great success including the predictions of sequence specificities of RNA- and DNA- binding proteins (Alipanahi et al., 2015), noncoding-variant effects (Zhou and Troyanskaya, 2015) and splicing patterns in human tissues (Xiong et al., 2015). Recently, deep learning based approaches have been successfully introduced into proteomics research to predict tandem mass spectra for peptides (Gessulat et al., 2019; Tiwary et al., 2019).

For most deep learning methods, feed-forward neural network is the fundamental architecture. The simplest unit of the feed-forward neural network architecture is the linear combination of fixed nonlinear basis functions $\phi_j(x)$:

$$y(\boldsymbol{x}, \boldsymbol{w}) = f\left(\sum_{j=1}^{M} w_j \phi_j(x)\right)$$

In the above formula, *f* stands for a nonlinear activation function and $\{w_i\}$ are the coefficients or weights to be adjusted during training. In the feed-forward neural network, each basis function $\phi_j(x)$ is a linear combination of the inputs. In other words, the basic neural network could be described as a series of transformations. From the input data, M different linear combinations could be built by linear combinations of the input data points ($x_1, \ldots, x_D$) using the following form:

$$a_j = \sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}$$

In the formula above, $j = 1, \ldots, M$, which means there are M units or neurons in the second layer. The superscript (1) indicates that the coefficients are in the first layer of the neural network. In many neural network literature, $w_{ji}^{(1)}$ is referred as weights, thus we will follow this nomenclature in the following discussions. The term a_j is called activation. Usually all activations will be transformed by a differentiable and nonlinear activation functions $h(.)$:

$$z_j = h(a_j)$$

z_j is called a hidden unit in neural network. There are many options for the nonlinear activation function. Among those, the most popular one nowadays is the rectified linear unit (ReLU):

$$f(a_j) = max(a_j, 0)$$

An advantage of using ReLU compared with other nonlinear activation functions is that ReLU always have a high gradient that makes training neural network with many layers much faster than other activation functions. The hidden unit values after the transformation of nonlinear activation functions are linearly combined again to generate output unit activations:

$$a_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + w_{k0}^{(2)}$$

Since this transformation corresponds to the second layer of the neural network, $k = 1, \ldots, K$ refer to the units or neurons in the third layer in the neural network. Again this activation a_k is transformed by the activation function $h(.)$:

$$z_k = h(a_k)$$

The layers of neural network can be further expanded by the process shown above to construct a very deep neural network. To make the discussion simpler, we assume there are only four layers with two hidden layers in the neural network architecture. Following the third layer, the hidden units are linearly combined again to give the output unit activation:

$$a_l = \sum_{k=1}^{K} w_{lk}^{(3)} z_k + w_{l0}^{(3)}$$

In the formula above, $l = 1, \ldots, L$, which refer to the output units. The activation functions of the output layer can be varied according to the nature of data and different scientific questions. For

example, regression problem could take the identity function $y_l = a_l$ as the activation function. For binary classification problem, the usual choice is the logistic sigmoid function:

$$y_l = \sigma(a_l) = \frac{1}{1 + e^{-a_l}}$$

In the problem of multi-class classification, the softmax activation function could be used:

$$y_l = \sigma(a_l) = \frac{e^{a_l}}{\sum_{p=1}^{L} e^{a_p}}$$

All the layers can be combined to give the overall neural network that could be summarized by the following formula:

$$y_l(\boldsymbol{x}, \boldsymbol{w}) = \sigma\left(\sum_{k=1}^{K} w_{lk}^{(3)} h\left(\sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right) + w_{l0}^{(3)}\right)$$

It is noted that in the formula above all the weights and bias parameters can be grouped together by a vector $\boldsymbol{w}$, and all the input data points are grouped into a vector $\boldsymbol{x}$. In other words, neural network is a nonlinear function based transformation from an input vector $\boldsymbol{x}$ to an output vector $\boldsymbol{y}$ parameterized by an adjustable weight vector $\boldsymbol{w}$. Like most other machine learning methods, the learning process of feed-forward neural network is an iterative procedure to minimize an error function by optimizing parameters within the model through a series of steps. In each step, there are two stages of computation. The first one is about evaluating the derivatives of the error function with respect to weights. Backpropagation is the major technique for computing the gradient of the error function for a multi-layered neural network. The second step is using the computed gradient to adjust the weights in the multi-layered neural network. Backpropagation is considered as one of the most important algorithms developed in the deep learning community (Rumelhart et al., 1986), which enables efficient evaluation of derivatives within multi-layered neural networks. Let's define an error function constructed by the sum of error terms, in which each error term comes from corresponding data point in the training set:

$$E(\boldsymbol{W}) = \sum_{n=1}^{N} E_n(\boldsymbol{W})$$

The backpropagation algorithm basically helps evaluate $\nabla E_n(\boldsymbol{W})$ in the above formula. Recall from discussion in the section of linear regression, in the case of simple linear model, the value of output y_k is basically the linear combination of input data point set $\{x_i\}$:

$$y_k = \sum_i w_{ki} x_i$$

For a particular input pattern n and the target value t, the error function can be defined by the following way:

$$E_n = \frac{1}{2} \sum_k (y_{nk} - t_{nk})^2$$

Thus, the derivative of this error function with respect to a weight w_{ji} is:

$$\frac{\partial E_n}{\partial w_{ji}} = (y_{nj} - t_{nj}) x_{ni}$$

In the case of feed-forward network, each neuron in the network calculates the weighted sum by:

$$a_j = \sum_i w_{ji} z_i$$

And then the weighted sum a_j is transformed by a nonlinear activation function:

$$z_j = h(a_j)$$

Now, we try to evaluate the gradient of E_n with respect to a weight w_{ij}. It is noted that E_n is dependent on w_{ij} through the summed a_j to unit j. Chain rule can be used here to get the partial derivative:

$$\frac{\partial E_n}{\partial w_{ji}} = \frac{\partial E_n}{\partial a_j} \frac{\partial a_j}{\partial w_{ji}}$$

From the formula of weighted sum, it is noted that:

$$\frac{\partial a_j}{\partial w_{ji}} = z_i$$

For simplicity, we can write:

$$\delta_j = \frac{\partial E_n}{\partial a_j}$$

Thus, it is easy to see:

$$\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i$$

In the last layer, for the output units, we can see:

$$\delta_k = y_k - t_k$$

The evaluation of δ for each hidden unit is based on partial derivatives again:

$$\delta_j = \frac{\partial E_n}{\partial a_j} = \sum_k \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial a_j}$$

In the formula above, units k_s are all the units that unit j sends connections to, in other words, these units are located on the outer layer of unit j in the neuron network. This formula also indicates that the variations in the error function contributed by a_j are only through variables a_k. To combine all the equations above, we can derive the final backpropagation formula:

$$\delta_j = h'(a_j)\sum_k w_{kj}\delta_k$$

The value of δ can be obtained by the above formula by propagating δ backwards from units in outer layer in the network. Since the values of δ for output layer are always known from training data ($\delta_k = y_k - t_k$), we can apply this formula to evaluate δ for all the hidden units in a feed-forward neural network recursively. Once the derivative of error function is computed by backpropagation algorithm, we move to the second stage to adjust the weights $\boldsymbol{w}$. There are many different optimization algorithms have been developed for this stage. A simple but efficient algorithm called stochastic gradient descent that updates the weights vector $\boldsymbol{w}$ using one training point at a time:

$$w^{(\tau+1)} = w^{(\tau)} - \eta\nabla E\left(w^{(\tau)}\right)$$

In which, $\eta > 0$ is the learning rate. The update is usually performed by randomly selecting points from the data and repeated for many times. Recently, a method built upon stochastic gradient descent named Adam has been became popular in deep learning community and reported to outperform other optimization algorithms in many papers (Kingma and Ba, 2014).

1.3 Aims and objectives

As discussed above, a large amount of effort has been invested to map physical interactions in biological systems. The traditional AP-MS and yeast two-hybrid approaches require tremendous effort on genetic engineering for individual genes and the introduced tag or fusion to another protein may affect protein structure, which could affect PPIs. The tagless co-fractionation approach is a new approach to map PPIs that is more efficient than traditional ones in terms of setup work required to run the experiment. However, the co-fractionation approach to detect PPIs requires much computational effort for post experimental analysis, which could be challenging

for most biology labs. There is no state-of-art and standard protocol or software to guide how to perform co-fractionation experiments and analyze the data. We close this gap by developing a software workflow, named EPIC, to automatically analyze co-fractionation based proteomics data to predict PPIs and protein complexes and automate the whole process from raw data scoring to visualizing the prediction results (Chapter 2). To demonstrate the practicability of EPIC, we performed co-fractionation experiments on a popular model organism *Caenorhabditis elegans* and applied EPIC to predict its protein complexes map (WormMap, Chapter 3). We validated both known complexes with novel components and totally novel complexes from WormMap using an independent orthogonal approach (AP-MS). WormMap is the first global scale biochemistry based protein complex map in nematode species and will serve as a useful resource for the worm research community. I also explored a relatively new machine learning approach (deep learning) to predict protein complexes using co-fractionation raw data without performing any feature engineering, in an effort to improve on the EPIC workflow (Chapter 4).

CHAPTER 2

EPIC: a software toolkit for elution profile-based inference of protein complexes

A paper has been published in Nature Methods (Hu et al., 2019), partially based on the content of this chapter. The work presented in this chapter was done by a close collaboration with previous postdoctoral fellow Dr. Florian Goebels in Emili and Bader labs. I wrote the initial Perl scripts. In EPIC, I wrote the training set collection, functional evidence integration and results benchmark parts. The co-fractionation data was collected by previous postdoctoral fellow Dr. Cuihong Wan and myself. Prof. Andrew Emili and Prof. Gary Bader co-supervised the project.

2 EPIC: a software toolkit for elution profile-based inference of protein complexes

2.1 Introduction

Systematic mapping of multi-protein complexes formed by PPIs is critical to understand the mechanistic basis of cellular processes. Affinity purification coupled to mass spectrometry (AP/MS) (Rigaut et al., 1999) is a powerful method for identifying such assemblies and has been applied widely (Babu et al., 2012; Gavin et al., 2006; Gavin et al., 2002; Hein et al., 2015; Ho et al., 2002; Hu et al., 2009; Huttlin et al., 2015a; Krogan et al., 2006), but is difficult to scale up or apply to non-model organisms. Biochemical co-fractionation coupled to mass spectrometry (CF/MS) is a more efficient and flexible alternate strategy for examining native macromolecules on a global scale (Havugimana et al., 2012; Wan et al., 2015). CF/MS is based on biophysical (typically chromatographic) co-purification of stable-associated proteins starting from cell-free mixtures (e.g. tissue lysates). However, sophisticated data processing is needed to define genuine interactions, which can be challenging to implement.

To facilitate such studies, a simplified, standardized and fully automated CF/MS data analysis software toolkit, EPIC, was developed, which enables routine scoring and interpretation of large-scale CF/MS data regardless of sample source. Using supervised machine learning EPIC integrates experimentally derived CF profiles and complementary functional evidence from public databases to create probabilistic PPI networks, which are then clustered to define high-confidence complexes. In the following sections of Chapter 2, I will first describe how to experimentally generate co-fractionation data using a coupled system of high performance liquid chromatography (HPLC) and liquid chromatography tandem mass spectrometry. And then I will discuss the details of how to computationally score PPIs and predict protein complexes using the generated co-fractionation data.

2.2 Experimentally generate co-fractionation data

CF/MS is based on extensive experimental separation of native macromolecular mixtures under non-denaturing conditions. While there is no universally optimal protocol, ion exchange high-

performance chromatography (IEX-HPLC) is efficient at resolving stable endogenous complexes. The entire experimental workflow of CF/MS is shown as in **Fig. 2-1** below:

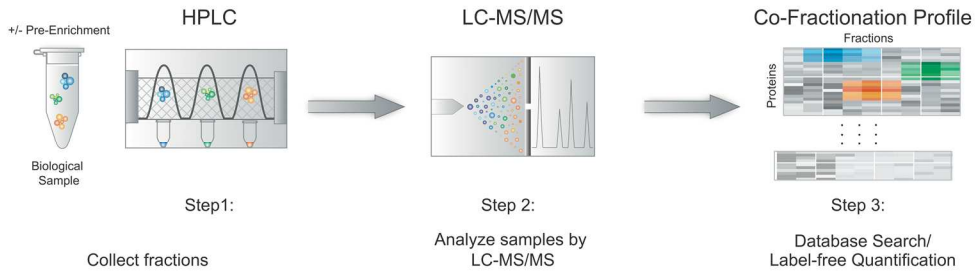

Figure 2-1: CF/MS experiments have three main steps: biochemical fractionation, MS analysis, and protein profile scoring.

In the following sections, I will describe each experimental step of generating co-fractionation data in detail.

2.2.1 Materials and methods

In this thesis work, we applied CF/MS experiments and the EPIC software tool to chart PPIs in *C. elegans*. In the following sections, *C. elegans* will be used as an example to demonstrate how to perform CF/MS experiments.

2.2.1.1 Protein extract preparation

Mixed-staged N2 strain *C. elegans* were collected in M9 buffer (standard recipe (Stiernagle)), and re-suspended into lysis buffer (50 mM HEPES pH7.4, 1 mM $MgCl_2$, 1 mM EGTA, 100 mM KCl) plus protease inhibitor cocktail (Roche). Worms were lysed by 3 rounds of 10 sec sonication on ice (Branson Sonifer 450, output 6.0, duty cycle 60%). Soluble protein lysate [~2

mg/ml] was collected by filter centrifugation (Ultrafree®-MC-HV, 0.45 μm). Bradford assay was used to determine protein concentration.

2.2.1.2 Pre-enrichment before HPLC fractionation

Commercial differential affinity capture beads (NuGel *PRO*spector; BSG) were used to pre-enrich the worm lysate according to the manufacturer's protocol. After removal of lipids and insoluble biomass, extract incubated with different reagent beads (*PRO*-A, *PRO*-B, *PRO*-C, *PRO*-L, *PRO*-N, *PRO*-R). The suspensions were mixed for 10 min at 4 °C, centrifuged using Spin-X filters, and the filtrate was collected as 'flow-through' fractions. Bound proteins were eluted with 200 ul elution buffer (0.2 M Tris, 0.5 M NaCl, pH 9.0). The buffer was exchanged for HPLC loading buffer by Zeba desalt spin column (Thermo) before HPLC fractionation.

2.2.1.3 HPLC separation

C. elegans lysate and affinity enriched eluates (plus flow-through fractions) were individually fractionated by ion-exchange liquid chromatography using a quaternary pump 1100 HPLC system (Agilent Technologies). Whole proteome lysate was resolved into 120 fractions on a PolyCATWAX mixed-bed ion exchange column (200 x 4.6 mm id, 12 μm, 1500 A) over a 240 min salt gradient (0.15 to 1.5 M NaCl). Enriched eluates were separated on a PolyCATWAX mixed-bed ion exchange column (200 x 4.6mm id, 5 μm, 1000A) into 60 fractions using a 120 min salt gradient (0.15 to 1.5 M NaCl).

2.2.1.4 LC-MS/MS analysis

Proteins from the HPLC fractions were acid precipitated, re-dissolved and digested by sequencing grade trypsin overnight at 37 °C. The resulting peptides were dried and solubilized in 5% formic acid. Data-dependent LC-MS/MS was performed using a nano-flow HPLC System

(EASY-nLC, Proxeon, Odense, Denmark) coupled to an LTQ Orbitrap Velos Mass Spectrometer (Thermo Fisher). After loading onto a 2.5 cm C18 trap column (75 mm inner diameter) packed with 100A Luna 5u C18 beads (Phenomenex) using an auto-sampler, peptides were separated on a 10 cm analytical column (75 mm i.d.) packed with 2 mm Zorbax 80XDB C18 reverse phase beads (Agilent). A 60 min gradient consisting from 5% to 35% ACN in water (with 1% formic acid) was used to elute peptides. Electro-spray ionization was performed using at 2.5kV spray voltage, and the instrument was operated in a data dependent mode (one full MS1 ion survey scan directing consecutive MS2 acquisition scans on the top 10 most prominent precursor ions). Collision induced dissociation (CID) directed peptide fragmentation was performed by 35% normalized collision energy.

2.2.1.5 Protein identification and label-free quantification

Raw spectral files were converted into mzXML format using the ReAdW software. A canonical FASTA file for protein searching was downloaded from the UniProt database and appended with common contaminants and reverse decoy sequences to assess the false-discovery rate (FDR). The peptide-spectrum matches from three different searching engines (comet, MSGF+ and X!Tandem) were integrated probabilistically using MSblender (Kwon et al., 2011), setting the FDR to less than 1% for peptide and protein identifications. Parameter settings and detailed search protocols are available online (http://www.marcottelab.org/index.php/MSblender). MaxQuant (Cox and Mann, 2008) (Version 1.6.0.16) search was performed at a fragment ion mass tolerance of 20 pp., maximum missed cleavage of 2 and a 1% false discovery level (controlled by target/decoy approach). SEQUEST (Version 2.7) search was performed at 20 pp. fragment ion mass tolerance and one missed cleavage allowance. The STATQUEST (Kislinger et al., 2003) model was used to assign confidence scores to all putative matches of peptides and proteins and a false discovery rate was controlled at 1% for all identifications.

2.2.2 Results

Applying the experimental strategies described, in total 1,380 HPLX worm lysate fractions were collected from an HPLC machine. All the fractions were processed by the LC-MS/MS system, in which 10,525 worm proteins were identified. Pre-enrichment benefited the process by increasing the coverage of detected proteome and helping detect more low-abundant proteins that results in more diverse GO term representation (**Fig. 2-2**). The results from a CF/MS experiment can be summarized as a matrix of biochemical fractions *versus* protein identities containing MS-derived protein amounts for each fraction (e.g. summed precursor ion intensities or spectral counts; an example of CF/MS derived result matrix is shown in **Fig. 2-3**). The collection of these CF/MS matrices could be used to infer protein complex membership using sophisticated computational approaches as documented in the following sections.

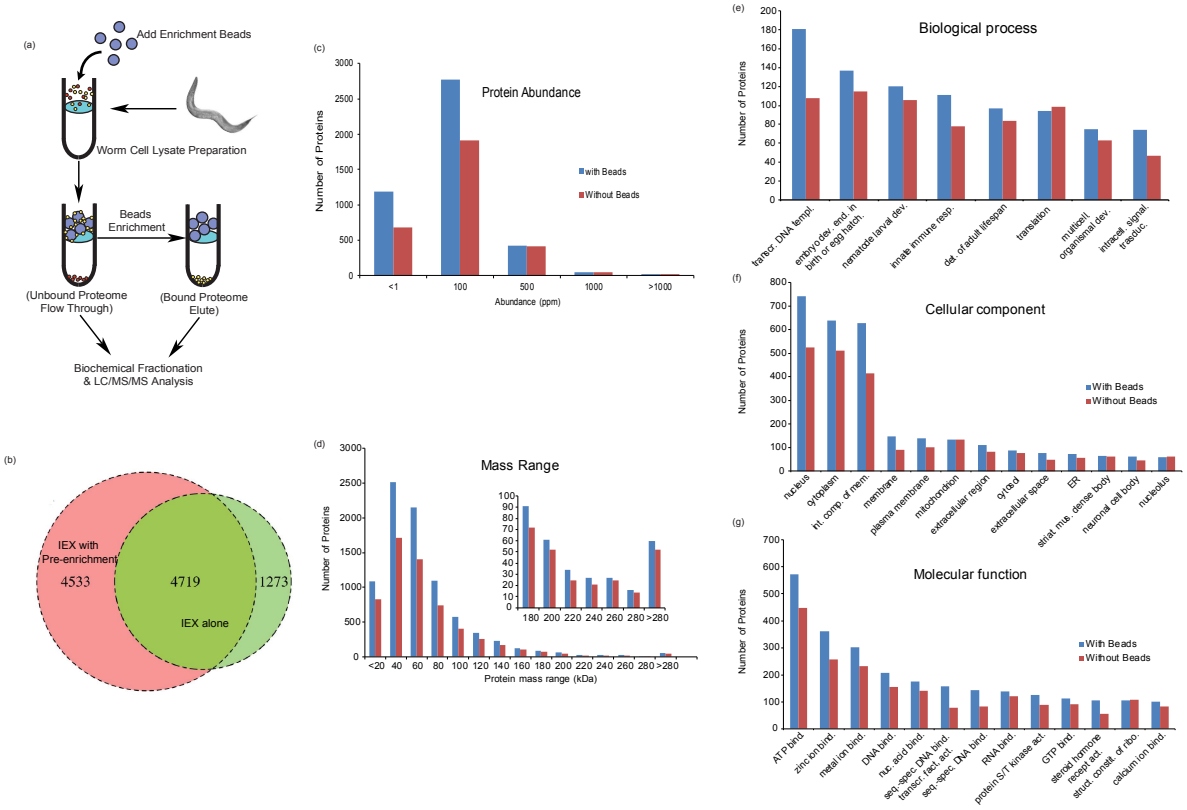

Figure 2-2: Pre-enrichment improves the dynamic range of CF/MS studies. a) Schematic workflow of bead-based sample pre-enrichment. b) Venn diagram showing improved proteome coverage by pre-enrichment. c) Bar chart showing improved detection of low abundance proteins. d) Bar chart showing improved detection of small (low molecular mass) proteins. e) Bar chart showing the distribution of identified proteins across top 8 biological processes in GO. f) Bar chart showing the distribution of identified proteins across top 13 cellular localizations in GO. g) Bar chart showing distribution of identified proteins across top 13 molecular functions in GO.

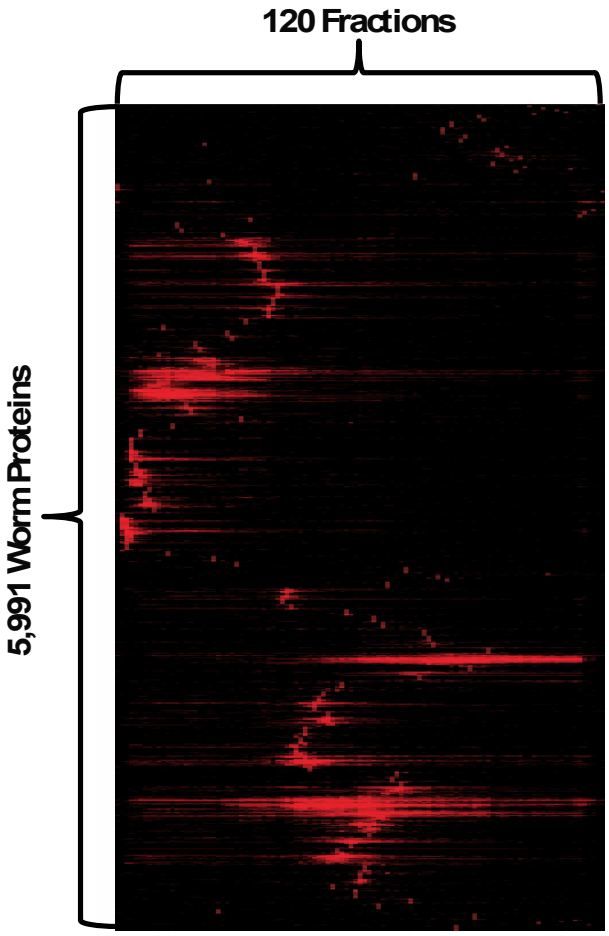

Figure 2-3: An example of real co-elution data from one of the co-CF/MS experiment. In total, 120 fractions were collected and 5,991 proteins were identified and quantified.

2.3 Computationally predict protein complexes from co-fractionation data

The computational workflow of EPIC can be summarized into three major parts: performing feature engineering to calculate different correlation coefficients between pairwise protein elution profiles; using machine learning classifier to integrate co-fractionation data and predict within-complex PPIs; segmenting the resulting protein network to generate protein complexes. The workflow can be summarized in **Fig. 2-4**.

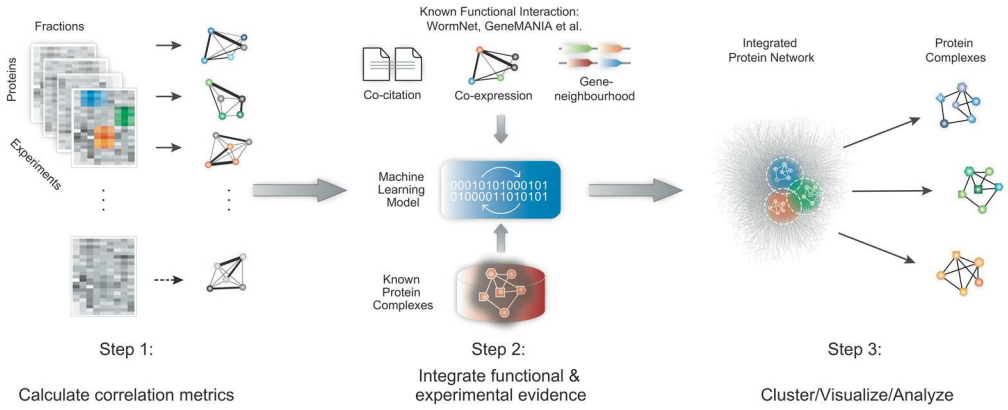

Figure 2-4: Automated computational analysis using EPIC takes CF/MS data as input and consists of three main steps: (i) calculation of co-elution profile similarity using correlation metrics; (ii) co-complex PPI scoring using machine learning-based integration of experimental and functional evidence; (iii) prediction, clustering, and benchmarking of derived complexes.

In the following sections, I describe the computational components of the EPIC workflow that use machine learning methods for prediction with the goal of identifying as many PPIs as possible, while minimizing the 'chance co-elution' problem using CF/MS based protein profiles.

2.3.1 Material and methods

2.3.1.1 EPIC software environment

EPIC employs python scripts to score CF/MS data, with modules to (i) process protein co-elution profiles, (ii) optionally download supporting functional association information from public databases (CORUM (Ruepp et al., 2010), UniProt (UniProt, 2015), IntAct (Orchard et al., 2014), GO (The Gene Ontology, 2017), GeneMANIA (Zuberi et al., 2013), STRING (Szklarczyk et al., 2017), InParanoid (Sonnhammer and Ostlund, 2015)), (iii) predict and benchmark predicted associations versus curated reference assemblies (CORUM, IntAct and GO, **Fig. 2-5**), and (iv) cluster and visualize the resulting PPI network using Cytoscape (Shannon et al., 2003). Given suitable experimental CF/MS data and a standard taxonomy identifier for the organism under study, the software collects required information from online sources and automates all data processing from raw data scoring to visualizing the results.

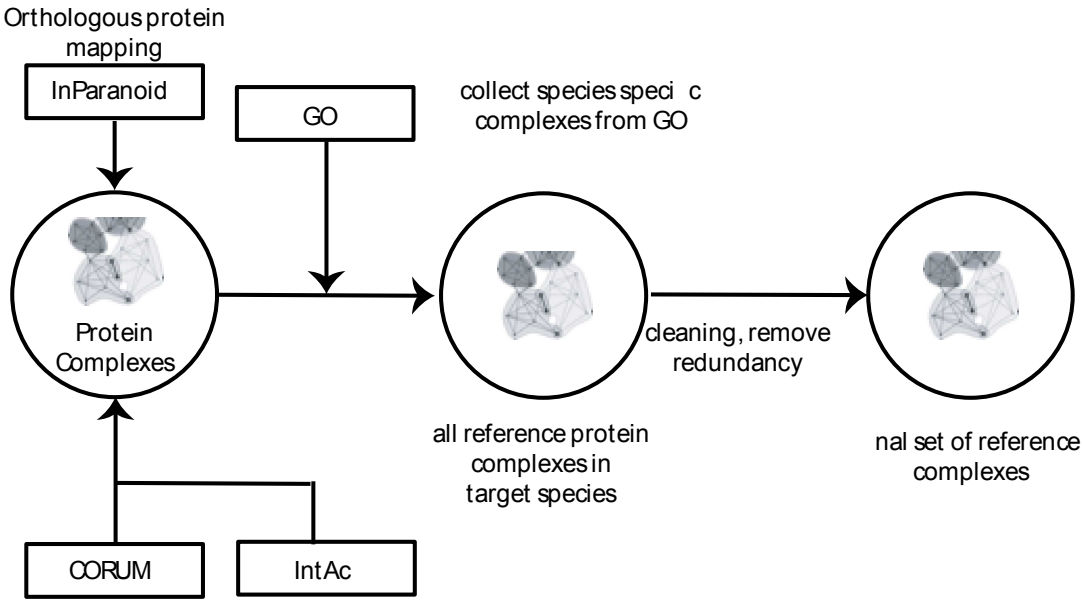

Figure 2-5: Schematic workflow for generating training set of macromolecules. Previously reported protein complexes, collected from the CORUM, GO and Intact curation databases, are first mapped to a target species protein complexes based on InParanoid orthology predictions. Redundancy is minimized to generate a final set of reference assemblies.

2.3.1.1.1 Reference dataset

Our goal is to make EPIC a generic tool for surveying protein complexes in different species. To facilitate standardization, we decided to use the CORUM database (Ruepp et al., 2010) as the source of the gold standard set, as it is the largest manually curated protein complex database available. EPIC utilizes human protein complexes for generating the necessary reference data, since protein complex information is typically sparse for the majority of species and as CORUM itself mainly curates human protein complex information. EPIC automatically downloads the current CORUM version and retains only those complexes that are annotated for human or mammals. Further, only protein complexes defined based on biochemical approaches are retained in the reference dataset, as protein complexes defined based on non-biochemical

methods might not be expected to co-elute by chromatographic separation. As an added set, EPIC downloads all human protein complexes from the IntAct database, for which again only complexes detected by biochemical methods are retained. Additionally, EPIC automatically downloads a set of curated protein complexes in the Gene Ontology (GO) database, annotated based on biochemical evidence for relevant target species (e.g. *C. elegans*).

We then generate an extracted set of positive and negative PPIs for both the training and holdout protein complexes, respectively. PPIs are defined as positive if they are observed in the same protein complex. If proteins exist in the protein complex dataset but never appear in the same protein complex, then these two proteins are defined as negative PPIs.

For mapping human proteins to the input species (test sample), we integrated the InParanoid database, which is also automatically downloaded for each EPIC run. We only consider one-to-one orthologous protein mappings between human and the test species with an InParanoid confidence score of 100%. In this manner, curated human protein complexes are projected on to corresponding orthologous protein complexes in a target species of interest. To avoid bias, protein complexes with less than three members and large assemblies with more than 50 proteins are removed, because these would dominate the machine learning process. Further, to remove redundancy in our data set, highly overlapping protein complexes (high fraction of shared components) are merged. We evaluate the overlap of two complexes A and B as follows, where $|A|$ denotes the number of proteins in A:

$$overlap(A, B) = \frac{|A \cap B|^2}{|A| * |B|}$$

Protein complexes are merged if they have an overlap score of at least 0.8. This automatic process for generation of reference data set currently only supports UniProt identifiers because they are used by GO, IntAct, InParanoid and CORUM.

2.3.1.2 Data processing

Once the protein elution data was acquired from CF/MS experiments. Several steps are required to pre-process the raw mass spectrometry co-elution table in order to improve the quality of the predicted network. Data processing and machine learning prediction (discussed in the next section) are the two cores of EPIC. The details of these two parts are summarized in the figure below.

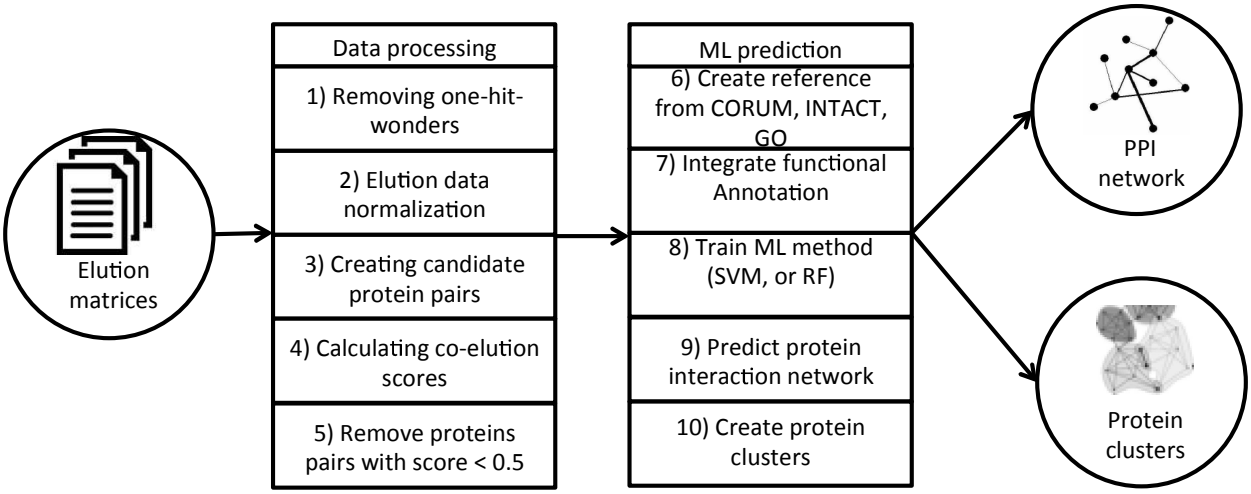

Figure 2-6: Detailed overview of the EPIC computational pipeline of data processing and machine learning prediction.

2.3.1.2.1 Removing "one-hit-wonders"

EPIC is based on the guilt-by-association principle, which posits that proteins that are physically associated tend to elute at the same time. However, to meaningfully evaluate fractionation data, EPIC requires the proteins to be present across multiple biochemical fractions within the same experiment. Thus, proteins measured in exactly one fraction are deemed 'one-hit-wonders' and removed from further analysis. The reason for discarding such proteins is not because we assume they were falsely measured, but rather that EPIC measures co-elution profile similarities based on correlation metrics that evaluate similarity over the entire elution profile, which is not effective for singletons. Some proteins may be identified in only one fraction in multiple

experiments. However, if we predict PPIs in this way, overall performance is markedly decreased (data not shown). Hence, each experiment is processed individually in EPIC, followed by merging/concatenating all the resulting co-elution correlation metric scores into a single unified matrix for machine learning. From the initial raw MS data, we observed that MSblender is highly sensitive and identifies the largest amount of peptides of which many are one-hit-wonders. However, even after removing one-hit-wonders, MSblender still has the largest amount of identified peptides compared with single search engines, resulting in the highest predicted quality protein complexes (**Fig. 2-7**).

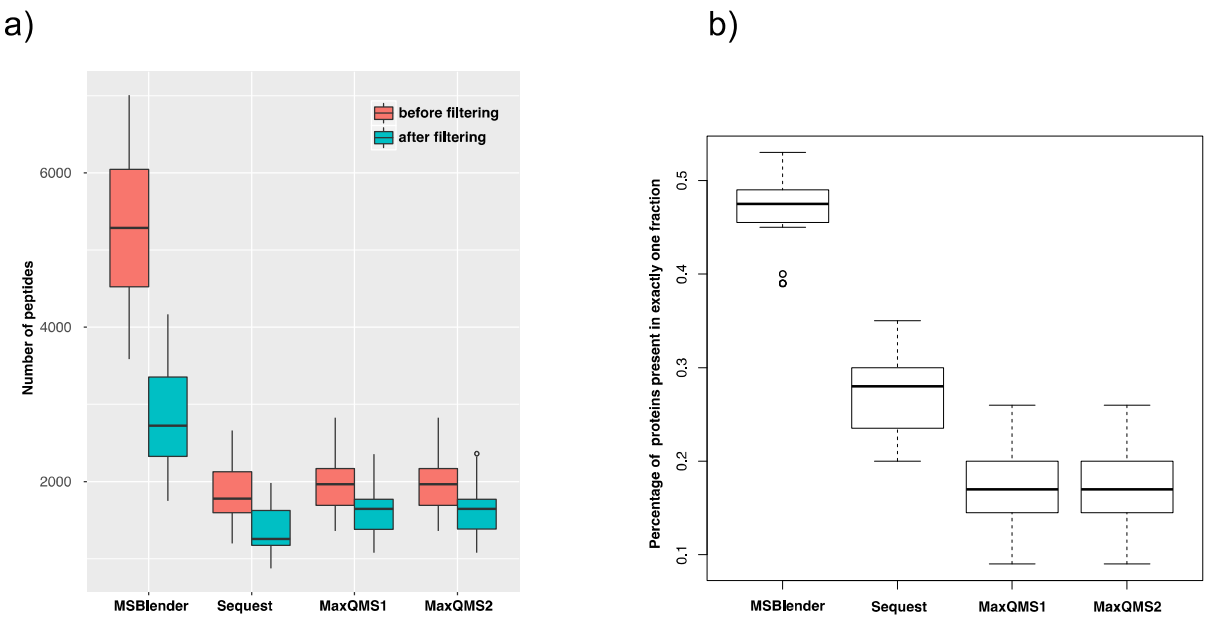

Figure 2-7: Comparison of peptides identified using different search tools. a) Number of Peptides before and after removing "one-hit-wonders" for each used searching tools identified in one co-fractionation experiment. There are 16 co-fractionation experiments (n = 16). b) Percentage of one-hit-wonders for each search engine. There are 16 co-fractionation experiments (n = 16). In each box plot, the middle line is the median, the lower and upper line of the box indicates the first and the third quartile. The upper and lower whiskers extend to the largest value less than the third quartile plus 1.5 times the interquartile range (IQR) and smallest value greater than first quartile minus 1.5 times the IQR, respectively. All data points beyond the whiskers are plotted as individual points.

2.3.1.2.2 Elution data normalization

Before calculating correlation coefficient metrics, the protein elution profile matrix is normalized column-wise to correct for slight sample injection variation. The protein elution profile matrix for each co-fractionation experiment consists of ion intensity or MS2 spectral counts for M proteins across N fractions. Thus, before calculating protein elution profile similarities, the raw data of each protein in each fraction is normalized by dividing the amount of the particular

protein (either MS1 ion intensity or MS2 spectral counts) by the total amount of proteins in corresponding fractions. So given a protein elution matrix A of the size $M \times N$, where each $A_{i,j}$ denotes the value of MS1 intensity or MS2 spectral counts of a particular protein i in fraction j, the column-wise normalized protein elution profile matrix $B_{i,j}$ is calculated as:

$$B_{i,j} = \frac{A_{i,j}}{\sum_i A_{i,j}}$$

Some similarity score metrics (i.e. Euclidean distance score) require row-wise normalization after column-wise normalization to make sure the sum of each row equal 1. So the final normalized protein elution profile matrix $C_{i,j}$ is calculated as:

$$C_{i,j} = \frac{B_{i,j}}{\sum_j B_{i,j}}$$

2.3.1.2.3 Creating candidate protein pairs

In previous work (Wan et al., 2015), we first created all possible pairs of proteins for each experiment, followed by calculating their corresponding co-elution scores and then removed all protein pairs without co-elution correlation scores equal or more than 0.5. However, this approach is computationally demanding and requires high-performance computational resources to perform all calculations in a reasonable amount of time. Thus, we decided to apply a pre-filtering step: instead of calculating all possible protein pairs for each experiment we first generate a super-set of all possible protein pairs across all experiments and remove those pairs for which the two proteins do not overlap (never occur in same fraction across all experiments). Usually, this filtering-step removes a substantial (up to 60%) of possible candidate pairs, significantly reducing computational time. In the subsequent step, we calculate co-elution scores for each candidate protein pair across each experiment and then summarize the results into matrices, and then we remove all protein pairs whose co-elution score is below 0.5 across all experiments. The rationale is explained in the next section.

2.3.1.2.4 Cut-off for correlation coefficient

We plotted the histogram of maximal correlation scores for all positive PPIs among all seven different correlation coefficients (apex score is not included, since it is either 0 or 1) across all experiments performed (**Fig. 2-8**). We noticed there is a clear cut-off at 0.5, which suggests we can retain protein pairs with a co-elution correlation score over 0.5 for machine learning prediction, as pairs without any co-elution score over 0.5 are not likely to be positive interactions.

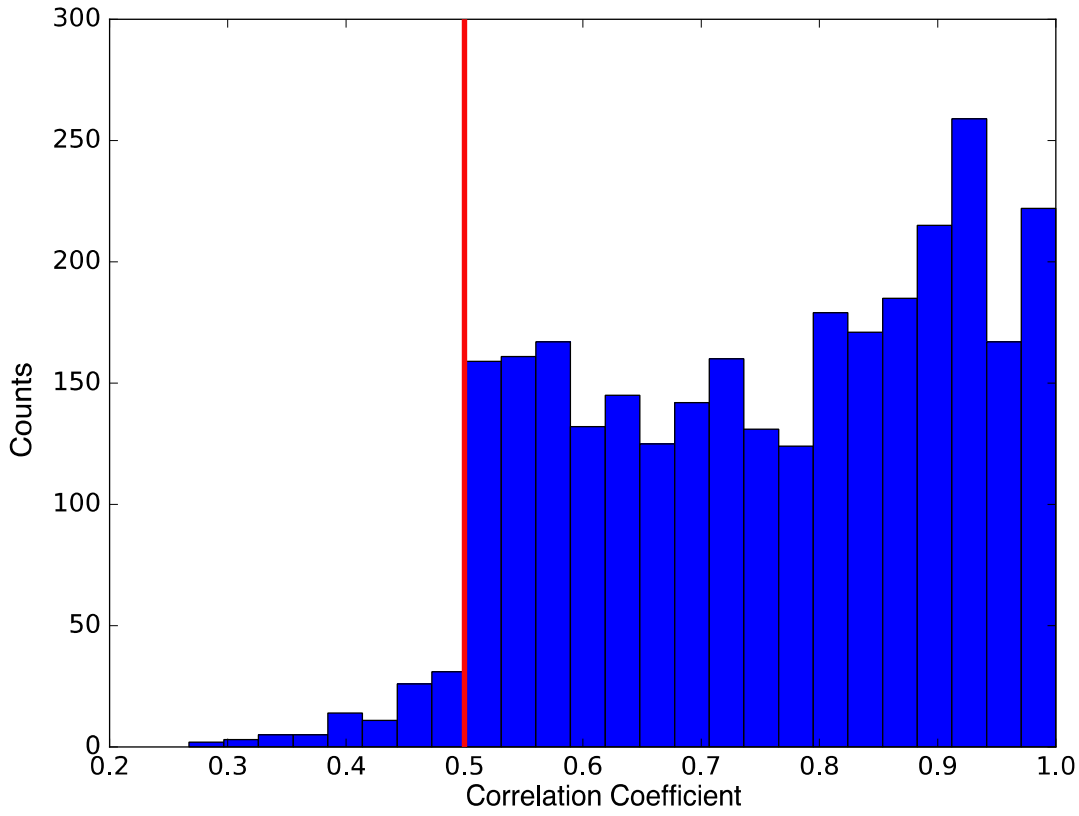

Figure 2-8: Correlation score cut-off setting. Histogram of maximal correlation scores of positive PPI pairs among all seven different correlation metrics across all 16 co-fractionation experiments. The red line indicates the cutoff chosen for EPIC.

2.3.1.2.5 Similarity metrics

Proteins that belong to the same protein complex should co-elute in the same or adjacent fractions, and thus should have similar elution profiles. In EPIC, we deploy several methods for measuring the similarity of two protein elution profiles. We treat each elution profile as a vector consisting of the observed MS2 spectral counts or MS1 ion intensities for a particular protein across the corresponding biochemical fractions, and a complete co-fractionation experiment is stored as a matrix where rows and columns represent proteins and fractions, respectively. To measure the co-elution profile similarity between two proteins, we employ various correlation metrics that range from simple scores, such as Euclidean distance, to more sophisticated metrics based on information theory. Some co-elution scores use normalized data $B_{i,j}$ while some use raw data $A_{i,j}$. In the following formulas: p_a and p_b denote protein a and protein b in the same co-fractionation experiment, N denotes the total number of proteins and M the total number of fractions.

2.3.1.2.5.1 Euclidean distance

Euclidian distance denotes the distance between two vectors (or two points) in a high-dimensional space (also known as 2-norm). The two points, for which the distance is calculated, represent a protein pair while the number of fractions is the dimension of space that the Euclidean theorem applies to. This Euclidean distance feature uses normalized counts and lies between 0 and 1, where identical elution profiles have a distance of 0 and elution profiles that differ greatly have a distance closer to 1.

2.3.1.2.5.2 Jaccard score

Jaccard score computes the ratio of how often proteins elute in the same fractions and how often proteins are detected in all fractions. Thus, the Jaccard score between two proteins is calculated

by counting the number of fractions that contain both proteins and dividing by the number of fractions that have at least one of the two proteins. The formula is as follows:

$$Jaccard(pa, pb) = \frac{|\{\#\ p\_a > 0\} \cap \{\#\ p\_b > 0\}|}{|\{\#\ p\_a > 0\} \cup \{\#\ p\_b > 0\}|}$$

2.3.1.2.5.3 **Bayes correlation**

We integrated a novel method that utilizes a Bayesian probabilistic framework for calculating correlation scores between two MS2 spectral counts based vectors. Originally, this method was proposed (Sanchez-Taltavull et al., 2016) to process RNA-seq gene expression data that is based on sequence counts for various genes under different conditions. Here, we applied the same method for peptide counts for various proteins across the biochemical fractions. The main advantage of Bayesian statistics over Pearson correlation is that it considers both measured signal magnitudes and associated uncertainties in those magnitudes. Thus, Bayesian correlation will returns high correlation values if measurement confidence is high and prevents high correlation values when the measurement confidence is low. To integrate Bayesian correlation, we integrated a public R script (http://www.perkinslab.ca/sites/perkinslab.ca/files/Bayes_Corr.R) into our python pipeline using the rpy python package that allows the import of R code into python. Bayesian correlation calculation scores support three different assumptions of how the priors are distributed: uniform, Dirichlet-marginalized and zero count-motivated. We used zero count for this work, as it performed best (**Fig. 2-9**).

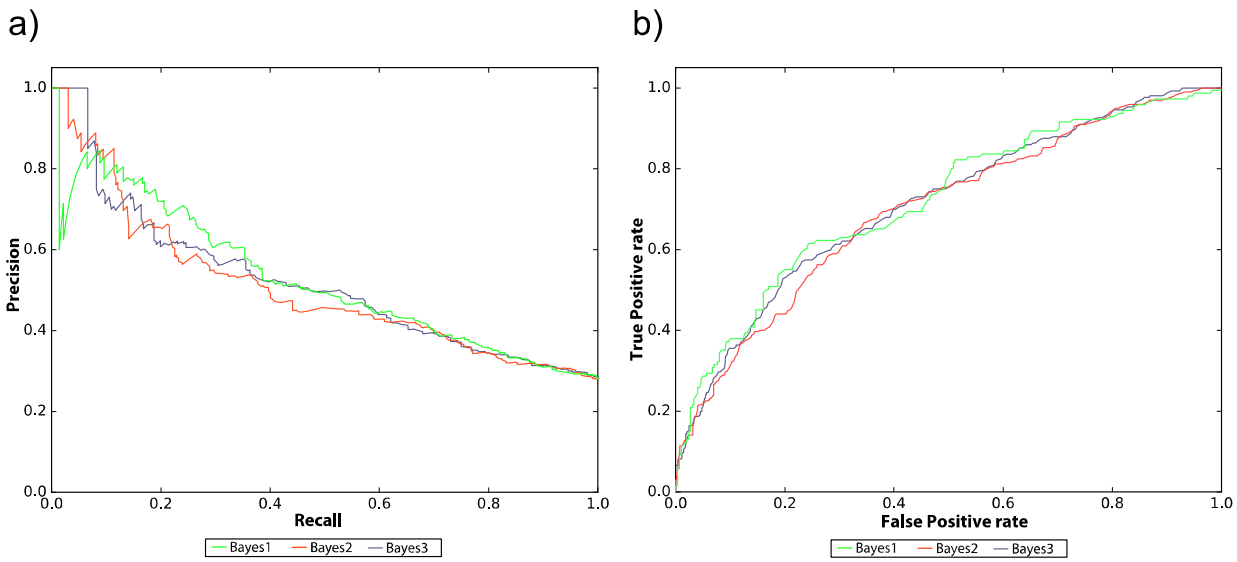

Figure 2-9: Different Bayes correlation priors comparison. Precision-recall (PR) curves (a) and Receiver-operating-characteristic (ROC) curves (b) for different Bayes correlation priors: uniform (Bayes1), Dirichlet-marginalized (Bayes2) and zero count-motivated (Bayes3).

2.3.1.2.5.4 Apex score

Most proteins tend to elute with a specific retention time, and thus the fraction that contains the largest amount of a particular protein is typically also the most critical fraction for that protein. Thus, two proteins are considered to be more likely to interact with each other if the fractions having the largest recorded amount across all fractions are the same. Based on this premise, previous co-fractionation experiments introduced the apex score (Havugimana et al., 2012), which scores protein co-elution profiles highly if their respective peak fractions are the same (apex score = 1) or else penalizes them (apex score = 0).

2.3.1.2.5.5 Pearson correlation coefficient (PCC)

Pearson correlation is used to measure the similarity of two protein co-elution profiles. In order to calculate PCC, we used the scipy package in python. PCC was calculated by using the vector of raw peptide counts or intensities obtained for each protein. From experience, PCC works well for proteins with high signal but not well for proteins with low peptide counts. Nevertheless, we decided to integrate this correlation metric into EPIC as it is a frequently used similarity metric, thus is also useful for benchmarking and evaluating other correlation metrics.

2.3.1.2.5.6 Pearson correlation coefficient plus noise (PCCN)

The Pearson correlation coefficient is relatively good at determining protein co-elution based on normalized protein elution profiles. However, proteins with low signals (low MS2 values) are more likely to co-elute by chance. To avoid this issue, the PCCN metric introduces a low level of random artificial signal on the raw co-elution data in the form of Poisson noise to each protein across all fractions, followed by co-elution matrix normalization and co-elution score calculation via Pearson correlation. This process is repeated n-times, and the resulting PCCN score is the average of those *n* runs. The same strategy has been used in creating previous co-elution networks (Havugimana et al., 2012; Wan et al., 2015), but here we systematically investigated the iteration parameter *n*.

2.3.1.2.5.7 Weighted cross correlation (WCC)

One of the issues of detecting eluting protein complexes from a liquid-chromatography based system is that the component subunits might show some residual retention time shifts. Unlike PCC, Weighted Cross Correlation (WCC) considers this small variance between otherwise similar co-elution profiles. To avoid promiscuity, stringent parameters are used to tolerate a small shift of roughly only one fraction when comparing two proteins. The WCC calculation is performed using the wccsom R package (Wehrens et al., 2005), which we integrated into our

python pipeline using the rpy2 python R interface package. WCC similarity is measured between 0 and 1.

2.3.1.2.5.8 **Mutual information (MI)**

Mutual information considers both linear and nonlinear dependencies between vectors. The initial step in calculating MI is to binarize the spectral count vector elements into 'with protein' and 'without protein', since mutual information measures statistical dependence between the two given proteins based on their relative co-elution frequency (% co-eluted fractions) and each protein's individual relative frequency (% fractions containing the respective protein). The elution matrix was binarized by temporarily changing each protein spectral count to 1 (if there were spectral counts observed in the fraction) or to 0 (if not present). Thus, $P(p_a = 1)$ denotes the individual relative frequency of p_a, which is calculated by dividing the total number of fractions with value 1 for protein p_a by the total number of fractions in the corresponding co-fractionation experiment, whereas the joint relative co-elution frequency of protein p_a and p_b named $P(p_a = 1, p_b = 1)$ is calculated by counting the total number of fractions that contain both p_a and p_b and dividing this number by the total number of fractions. MI is calculated as follows:

$$MI(p_a, p_b) = H(p_a, p_b) - H(p_a) - H(p_b)$$

In the formula above, $H(p_a)$ denotes the entropy of protein a and $H(p_a, p_b)$ the joint entropy with following formulas:

$$H(p_a) = -\sum_{i}^{\{0,1\}} P(p_a = i) * \log_2(P(p_a = i))$$

$$H(p_a, p_b) = -\sum_{j}^{\{0,1\}} \sum_{i}^{\{0,1\}} P(p_a = i, p_b = j) * \log_2(P(p_a = i, p_b = j))$$

2.3.1.3 Predict PPIs and protein complexes with the aid of machine learning

In the last step of the data pre-processing, EPIC generates a co-elution matrix, which contains rows for each protein pair and columns for each co-elution score across all co-fractionation experiments. In cases where a protein pair was not present in one of the experiments, I set all of its co-elution scores for the given experiment to zero. In the subsequent sections, I will describe how EPIC creates a co-elution PPI network and the set of protein complexes. How to evaluate the prediction results from EPIC on both PPI and complexes levels is also discussed.

2.3.1.3.1 Train machine-learning classifier and predict PPIs

The machine learning classifier is trained on the sets of positive and negative PPIs as we defined before based on CORUM, IntAct, and GO. We create the union of training set by merging the training set obtained from the above three databases, in which only the protein pairs have at least one elution profile similarity score larger than 0.5 (among all co-fractionation experiment and among all correlation metrics) are retained. We then train the classifier on this reduced set of negative and positive interactions with correlation metrics scores from different co-fractionation experiments as input features. Because the classifier is trained to distinguish true-positive co-complex membership with high co-elution score from non-interacting protein pairs including false-positive chance co-elution associations that also have high co-elution scores, we decided to additionally integrate functional evidence data (i.e. GeneMANIA, STRING and WormNet) into the machine learning method. However, to reduce circular reasoning in the machine learning step, functional evidence derived from “physical interaction”, “protein complexes” and “predicted interactions” are excluded from input features. For example, the final set of worm complexes prediction used functional evidence data from WormNet, and the lines of evidence are shown in the table below:

Data	Organism	Description
CE-CX	*C.elegans*	Inferred links by co-expression pattern of two genes (based on high-dimensional gene expression data)
CE-GN	*C.elegans*	Inferred links by gene neighbourhoods of bacterial and archaeal orthologs
CE-GT	*C.elegans*	Inferred links by genetic interactions
CE-PG	*C.elegans*	Inferred links by similar phylogenetic profiles
DM-CX	*D.melanogaster*	Inferred Links by co-expression pattern of two genes (based on high-dimensional gene expression data)
DR-CX	*D.rerio*	Inferred Links by co-expression pattern of two genes (based on high-dimensional gene expression data)
HS-CX	*H.sapiens*	Inferred Links by co-expression pattern of two genes (based on high-dimensional gene expression data)
SC-CC	*S.cerevisiae*	Inferred links by co-citation
SC-CX	*S.cerevisiae*	Inferred Links by co-expression pattern of two genes (based on high-dimensional gene expression data)
SC-GT	*S.cerevisiae*	Inferred links by genetic interactions
SC-TS	*S.cerevisiae*	Inferred Links by protein tertiary structure

Table 2-1: Lines of functional evidence taken from WormNet for worm complexes prediction.

EPIC generates a set of PPIs using the classifier trained on experimental data with an option to include functional evidence. Then a set of PPIs are predicted by the classifier trained on experimental data or optionally experimental data integrated with functional data. A protein elution profile correlation score cut-off was applied to ensure all PPIs have experimental evidence support (see above).

2.3.1.3.2 Predict protein complexes from the PPI network

In the final step, EPIC generates a set of putative complexes from the predicted protein interaction network. As with our previous work, we use the ClusterONE clustering method, and it has been shown to provide excellent performance among several different clustering algorithms for predicting protein complexes from PPI networks (Havugimana et al., 2012; Nepusz et al., 2012; Wan et al., 2015). Novel protein complexes are identified by comparing the predicted set of complexes and the curated protein complexes from the major databases (CORUM, IntAct and GO) by setting a liberal overlap score cut-off at 0.25.

2.3.1.3.3 PPI prediction metrics and evaluation

We utilize different measurements to evaluate EPIC performance based on its capabilities of predicting both PPIs and multi-protein complexes. Most of the evaluation metrics that we apply for measuring how well EPIC can predict PPIs are commonly used throughout the machine-learning field and are briefly mentioned in this section.

One first needs to define criteria of what is true for a predicted interaction. The summary of positive training protein complexes sets from reference databases is shown in the table below.

Pre-processing step	CORUM	IntAct	GO	All
Raw	1866	280	65	2211
Ortholog mapping	1342	135	65	1513
Size filtering	548	26	33	610
After merging	401	24	32	451

Table 2-2: Summary of reference protein complex datasets (CORUM, GO, IntAct). The numbers indicate the number of complexes for each dataset after each processing step.

With EPIC, evaluation of PPIs prediction is done by comparing the predicted interactions to the above mentioned generated reference data set of positive and negative protein interactions. Based on this concept, one defines precision, recall, and F-measure (also known as F1 score) as follows:

Precision:

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

Recall:

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

F-measure:

$$F\_measure = 2 * \frac{precision * recall}{precision + recall}$$

Additionally, we evaluate performance using the precision-recall (PR) and the receiver operating characteristic (ROC) curve. We use the area under the PR curve (auPR), and the area under the ROC curve (auROC) to give single value performance, which can be used to compare different methods or parameter settings.

2.3.1.3.3.1 Precision recall curve

The PR curve is created by first sorting the list of predicted protein interactions by their confidence scores and then iteratively removing the top element from that list while calculating the resulting precision and recall value for the updated list. The PR curve is the line that results by plotting those generated precision recall values. This line shows the trade-off between precision and recall, and area under precision recall curve measures the average precision of the classifier. It can be used to compare multiple models, since a better classifier will lead to a higher PR-curve and thus results in a larger auPR value.

2.3.1.3.3.2 Receiver operating characteristic curve

The ROC is generated analogously to the PR curve, but instead of plotting the resulting precision and recall values, the ROC plots true positive rate against the false positive rate. The auROC curve describes the probability of the classifier of scoring a positive interaction higher than a negative interaction, which means it shows how well the classifier can separate positive and negative PPIs. Thus, in a two groups classification problem, an auROC score of 0.5 means the classifier cannot differentiate between a positive interaction and a negative interaction, whereas a score of 1 means the classifier can perfectly predict the class labels.

2.3.1.3.4 Cluster prediction evaluation

Training a classifier on PPIs to determine whether or not a prediction is true is straightforward, as it only involves comparing the set of predicted PPIs against a set of pre-defined positive and negative protein interactions (see previous sections). However, in the case of predicting protein complexes that typically consist of three or more members, this comparison is more difficult. First, we describe a simple measurement for determining the precision of the predicted protein complexes based on the overlap of the predicted complexes to a given set of reference complexes. However, an important issue here is when one should consider two protein complexes as a match. Several protein complex prediction studies have investigated how to evaluate cluster overlap, and essentially all their measurements are based on how to evaluate the overlap between the set of proteins within complex A and the set of proteins within complex B. The overlap score between protein complexes are calculated as below (note that |A| denotes the number of proteins in complex A):

$$O(A,B) = \frac{|A \cap B|^2}{|A| * |B|}$$

It is suggested to consider two protein complexes to be matching when the overlap score between them is greater than 0.25, since two clusters of the same size would have this score if the intersection set is half of the complex size.

Additionally, we calculate prediction sensitivity, accuracy, positive predictive value, and cluster separation (Brohee and van Helden, 2006). For the following scores we consider $a_1,\ldots,a_i,\ldots,a_m$ predicted complexes which we compare to a set of $b_1,\ldots,b_j,\ldots,b_n$ reference complexes, and $T_{i,j}$ denotes the number of proteins that are found in both complex i and j.

Sensitivity (Sn): fraction of proteins in predicted complexes that are found in reference complexes.

$$Sn = \frac{\sum_{i=1}^{n} max_{j=1}^{m} t_{i,j}}{\sum_{i=1}^{n} |b_i|}$$

Positive predictive value (PPV): indicates how specific and complete the predicted complexes match the reference complexes. A score of 1 indicates that each predicted complex only overlaps which exactly one reference complex, and a low score indicates low or redundant overlap with the reference.

$$PPV = \frac{\sum_{j=1}^{m} max_{i=1}^{n} T_{i,j}}{\sum_{j=1}^{m} \sum_{i=1}^{n} T_{i,j}}$$

Accuracy (Acc): shows the trade-off between *PPV* and *Sn*.

$$Acc = \sqrt{Sn * PPV}$$

Maximum matching ratio (MMR): The MMR was developed to cope with some of the limitations of the PPV. PPV tends to be lower if there is substantial overlap in the reference data (Nepusz et al., 2012), but those overlaps are common in biological data sets such as CORUM. Our merging step only removes highly overlapping clusters, but smaller overlaps are still present. Thus, even if EPIC perfectly predicts the reference complexes it will not achieve a score of 1 for PPV and Sep (clustering-wise separation score suggested by Brohée and Van Helden (Brohee and van Helden, 2006)). MMR can cope with this problem:

$$MMR = \frac{\sum_{i=1}^{n} max_{i=1}^{m} O(n_i, m_j)}{|max_{i=1}^{n} O(n_i, m) > 0|}$$

As established by others (Nepusz et al., 2012), we summarize MMR, overlap score, and accuracy to create the composite score (the sum of MMR, overlap score and accuracy), and we consider the parameter combination with the highest composite score to be the best combination.

2.3.1.4 Optimizing EPIC performance

We extensively benchmarked EPIC and optimized parameters for each step of the EPIC pipeline on our worm data to define the final protein network. In an ideal scenario, we would evaluate the complete space of all possible parameters, however the space for searching the optimal parameter configuration grows exponentially ($2^{|parameters|}$) with the number of parameters we want to configure. Thus, to make the benchmarking of EPIC feasible, we investigated only one parameter at a time while keeping the remaining parameters fixed. First we will describe benchmarking statistics and evaluation criteria followed by the results of benchmarking.

2.3.1.4.1 Feature parameters

In this part, we evaluate the optimal parameter settings for co-elution scores (if any parameter setting is involved). From the total of eight correlation features, two of them have parameters to optimize: the prior used for the Bayes correlation and the number of noise iterations for Pearson correlation plus noise (PCCN). We evaluated those parameters based on how well they can predict PPIs (i.e. precision, recall, F1, auROC, auPR). To be consistent, all the evaluations were performed using elution data generated by the MSblender search engine, as it is the search engine that generated the largest data set with the most identified proteins. The results for number of noise iterations can be found in the figure below and we observed optimal scores obtained for five noise iterations.

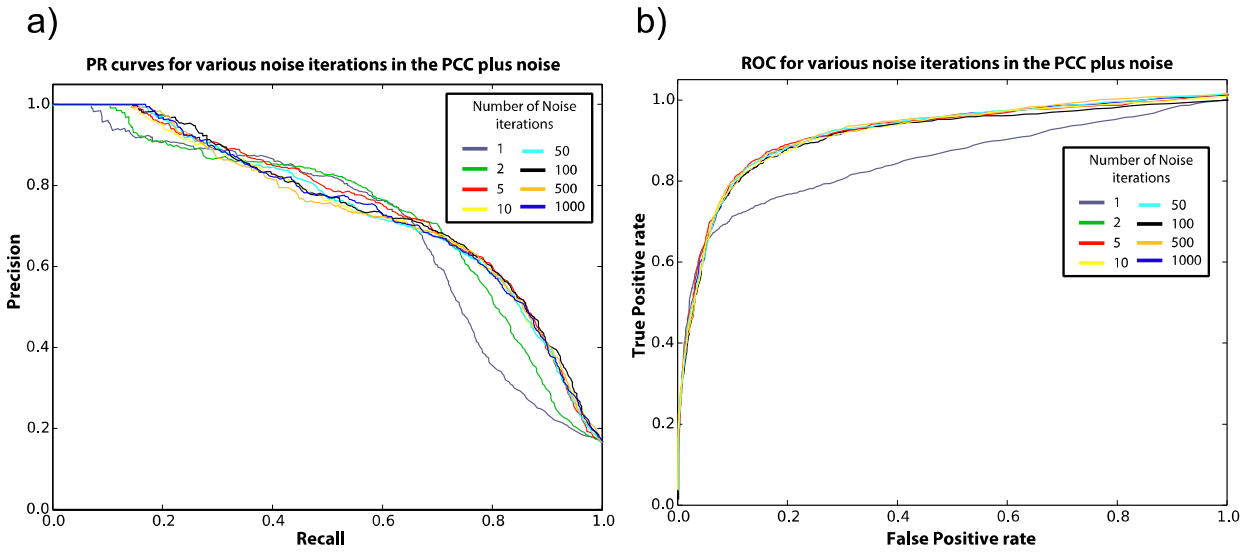

Figure 2-10: Number of Poisson noise iteration comparison. Precision-recall (PR) curves (a) and Receiver-operating-characteristic (ROC) curves (b) for different iterations of Poisson noise added in the Pearson correlation coefficient feature.

After analyzing the three possible Bayes priors, we observed no significant differences between the three different priors based on ROC and PR curves (**Fig. 2-11**).

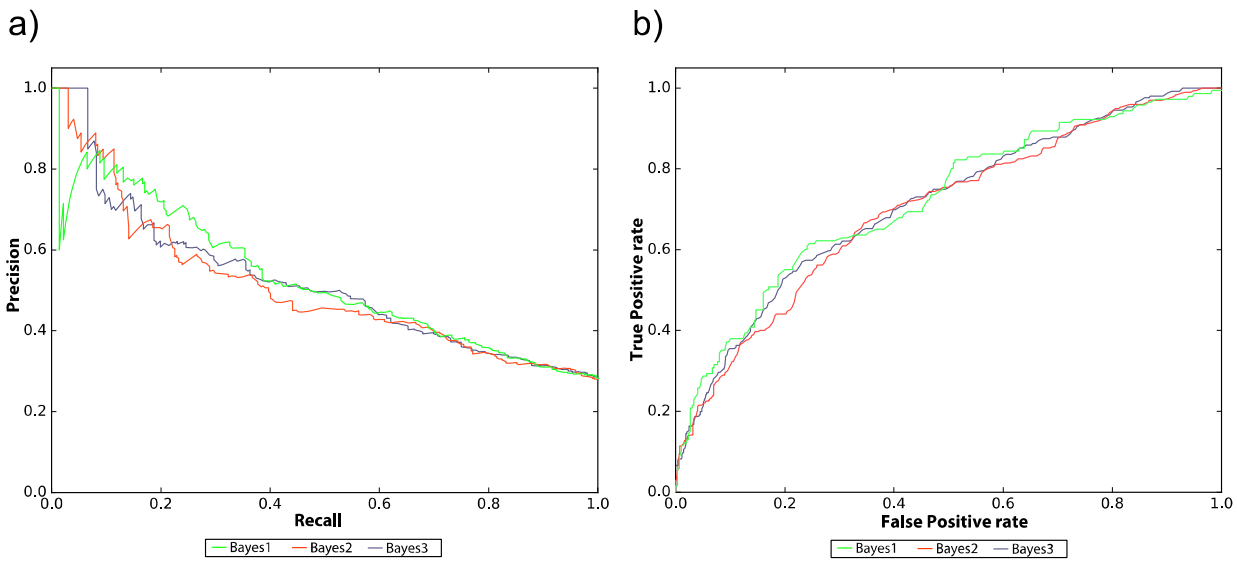

Figure 2-11: Different Bayes correlation priors comparison. Precision-recall (PR) curves (a) and Receiver-operating-characteristic (ROC) curves (b) for different Bayes correlation priors: uniform (Bayes1), Dirichlet-marginalized (Bayes2) and zero count-motivated (Bayes3).

However, if we analyze the evaluation metrics for predicted protein complexes we see the best composite score for the zero-count prior (Bayes3) (**Table 2-3**). Thus, we use the zero-count prior for EPIC.

Category	Bayes1	Bayes2	Bayes3
# Predicted PPIs	1360	2310	681
# Predicted clusters	134	204	96
Maximum matching ratio (MMR)	0.08	0.07	0.08
Overlap score	0.01	0.01	0.05
Accuracy score	0.17	0.18	0.29
Composite score	0.26	0.25	0.42

Table 2-3: Evaluation of three different available Bayes priors. The three priors are uniform (Bayes1), Dirichlet-marginalized (Bayes2), and zero count (Bayes3).

2.3.1.4.2 EPIC parameter optimization by nested cross validation

It is not possible to provide globally optimal parameters for all data sets. In EPIC, we developed a nested cross validation strategy to optimize parameters for our worm data and used the

optimized set of parameters to generate our WormMap. As described in **Fig. 2-12**, we first collected and merged all worm protein complexes from CORUM, GO and IntAct. We first used k-means clustering and an overlap score as the measurement metric to divide the whole set of reference protein complexes set into two distinct sets of complexes. We then balanced the two sets while minimizing the overlap by iteratively moving the most distinct protein complex from the set with more complexes to the set with fewer complexes. The first half is used for training (based on our co-fractionation data) while the second half is used as the 'holdout' set for evaluation (2-fold cross validation at the protein complex level). In our study and in the EPIC software, we implement two machine-learning classifiers, support for four protein searching/quantification tools and eight different correlation scores, which gives us 2,040 total parameter combinations. We trained machine-learning classifiers with our worm co-fractionation data to predict PPIs and protein complexes for each of the 2,040 different parameters combinations. The resulting 2,040 predicted protein complex sets were then benchmarked with the held out "test" half of the curated protein complexes using composite score (see above) as the evaluation metric.

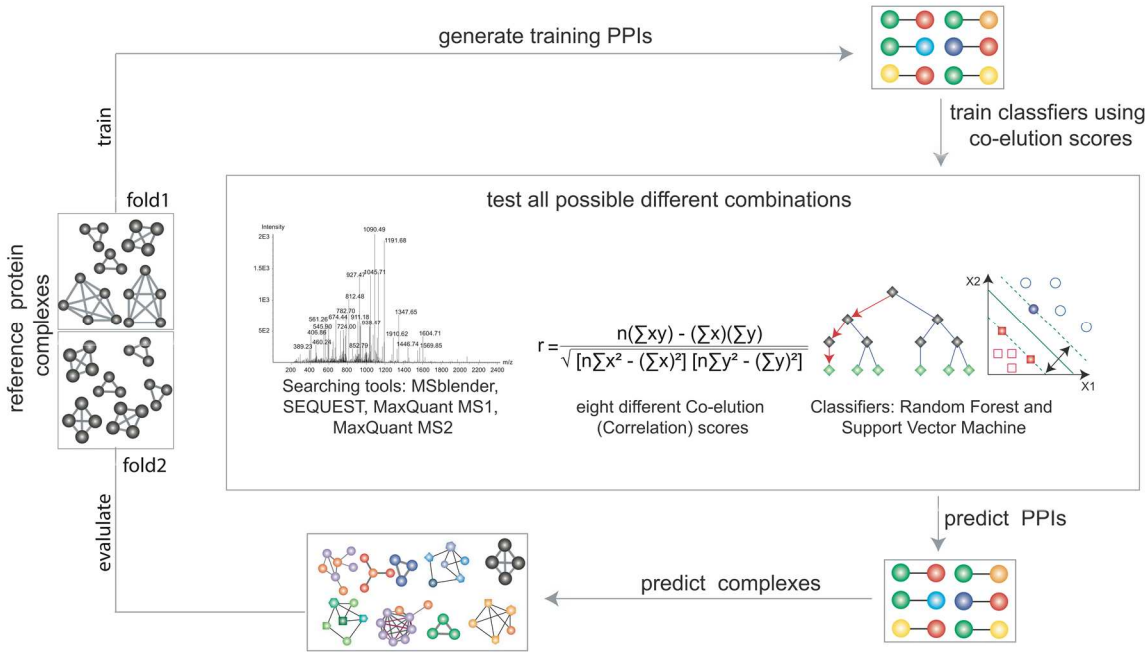

Figure 2-12: Computational procedures for protein interaction and co-complex prediction, driven by global optimization of classifier performance. The best combination of features was obtained using a nested cross-validation procedure.

Random forest in general outperformed support vector machine for predicting protein complexes. MSblender gives the best composite score compared with other protein search/quantification tools. To get a relatively good prediction, at least three different correlation scores are required. The results are shown in the figure below.

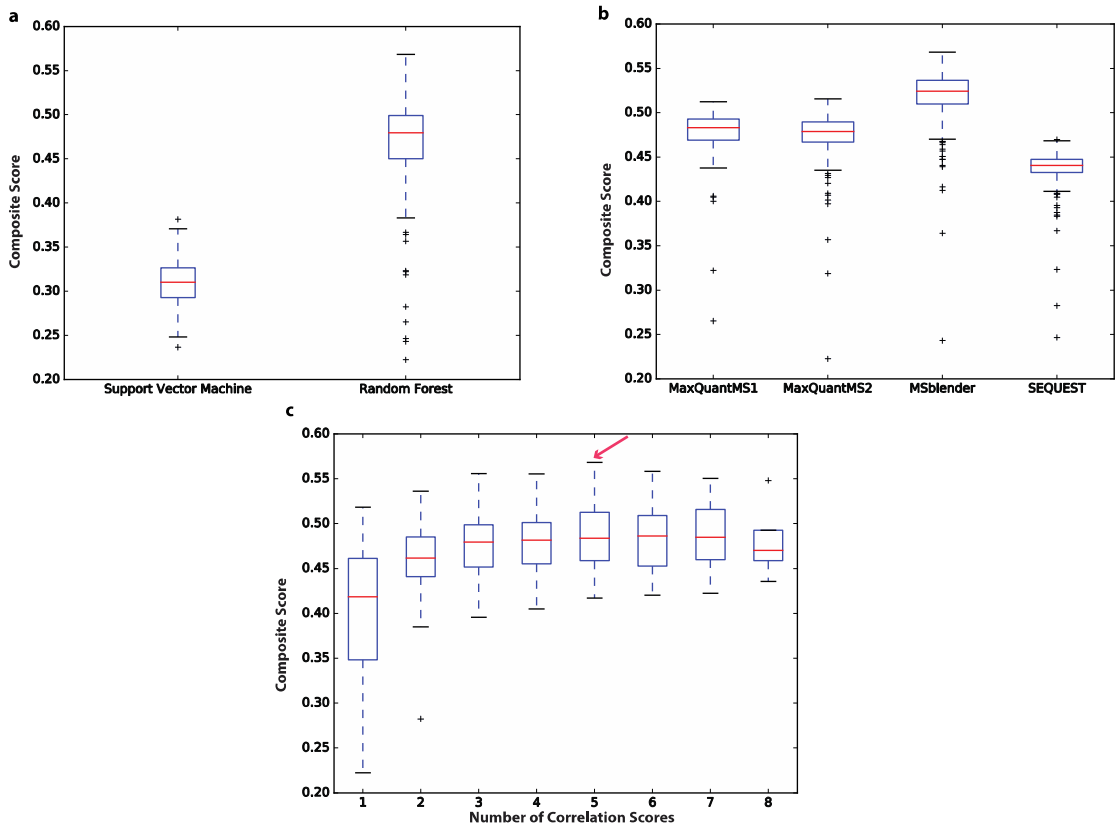

Figure 2-13: EPIC parameters global optimization by nested cross-validation. (a). Boxplot showing the complex prediction performance (composite score) from two different machine-learning classifiers (random forest n = 1014 vs. support vector machine n = 945). (b). Boxplot showing the complex prediction performance (composite score) based on the 234 results from each four different protein search/quantification tool. (c). Boxplot showing the relationship between different numbers of correlation scores and complex prediction performance (i.e. composite score). n = 28, 110, 224, 280, 224, 112, 32 and 4 are the number of composite score results with various correlation scores used (from 1 to 8). Red arrow indicates the set of (five) correlation scores producing the highest composite score. In each box plot, the red line is the median, the lower and upper line of the box indicates the first and the third quartile. The upper and lower whiskers extend to the largest value less than the third quartile plus 1.5 times the interquartile range (IQR) and smallest value greater than first quartile minus 1.5 times the IQR, respectively. All data points beyond the whiskers are plotted as individual points.

The optimized set of parameters (machine-learning classifier: random forest, protein searching/quantification tool: MSblender, correlation scores: mutual information, Bayes correlation, Euclidean distance, weighted cross correlation and apex score) for generating WormMap is the combination that gives the highest composite score. Functional evidence data was then added to the matrix formed by the optimal set of correlation scores for predicting PPIs. Since extensive computational resources are required for this optimization, we performed this analysis on the SciNet supercomputing platform (https://www.scinethpc.ca/). We provide a parameter optimization function in the EPIC software and encourage users to optimize their parameters using their own data if a super computing resource is available, but otherwise, we recommend using the default EPIC parameters, which are the ones that were found optimal for WormMap using the above procedure.

2.3.1.5 Exploring the value of additional experiments

After nested cross validation, the selected optimal correlation score combination and random forest machine learning classifier was used for evaluating if a pre-enrichment step improves protein complex prediction and what is the most economic way to perform experiments. We performed the analysis using data collected from pre-enrichment, non-pre-enrichment (IEX) and the combination of both (all experiments), individually. Similar to the step of nested cross validation, we benchmarked the predicted protein complexes using composite score, based on our 2-fold cross-validation strategy. For each specific number of experiments, we considered all combinations and reported the average of the evaluating metrics. For example, for the first point in the plot indicating use of one experiment, we analyzed each of our seven IEX experiments individually to predict complexes, evaluated the composite score and then calculated the average of number predicted complexes and composite scores over the seven experiments. We observed a positive correlation between composite score and the number of experiments (**Fig. 2-14a**). After five experiments, using IEX alone performed much better than using all experiments. Similarly, a sharp increase is observed for the last point of the "all experiments" line (red line). We then asked if the sharp increase of IEX performance is the result of sacrificing the coverage of predicted protein complexes. To balance the coverage of predicted protein complexes and

composite score, we then plotted "composite score × the number of predicted complexes" vs. "number of experiments" (**Fig. 2-14b**). In this plot, we noticed the "all experiments" line reached its stationary phase at nine experiments. We also noticed a dramatic decrease of the "IEX" line at seven experiments, which shows that the sharp increase of composite score for "IEX" is due to a decrease in the number of predicted protein complexes. Also, when using all 16 experiments, the composite score is maximized. Thus, the general guideline would be to use as many experiments as possible and that pre-enrichment will help protein complex prediction in terms of both composite score and coverage, however, if mass spectrometry time is limited, a reasonable lower bound is to run four IEX experiments.

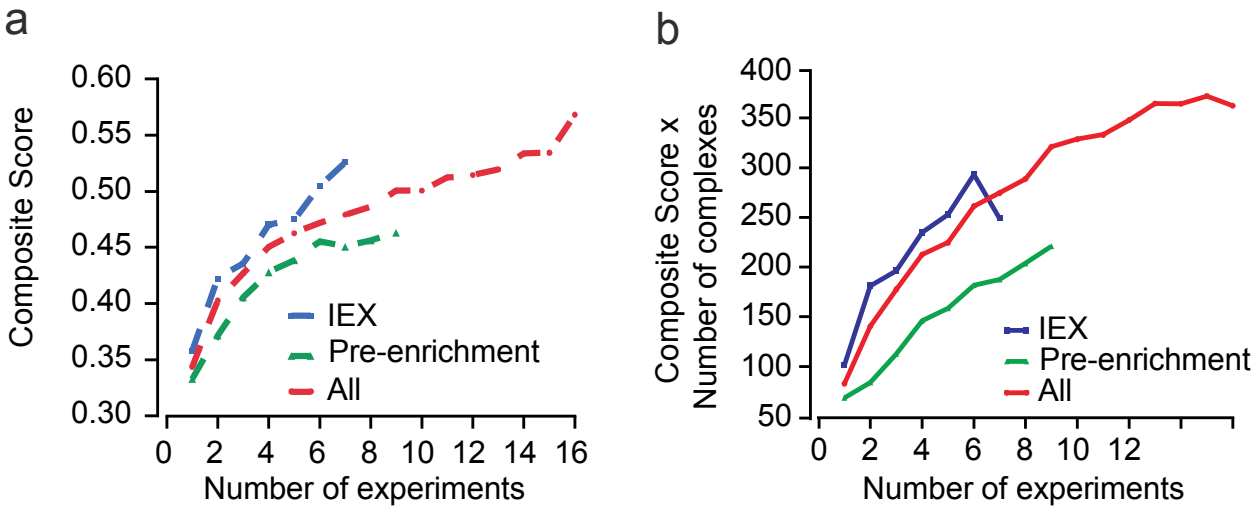

Figure 2-14: Exploring the value of additional experiments. (a). Line plot of the number of experiments and corresponding averaged composite score. (b). Line plot of the number of experiments and the corresponding averaged value of composite score times the number of predicted protein complexes.

2.3.1.6 Comparison of EPIC with PrInCE

PrInCE is a software tool that is recently introduced by Foster group to analyze co-fractionation data (Stacey et al., 2017). To objectively compare the performance of the two tools (PrInCE vs. EPIC), we downloaded the example SILAC co-fractionation data available from the PrInCE website (condition1.csv and condition2.csv) and used this as input data to predict protein complexes using both PrInCE and EPIC. We then compared the results (predicted complexes) with a benchmark set of reference assemblies (i.e. CORUM) using the multifactor composite score as the stringent evaluation metric. The resulting set of protein complexes predicted by EPIC with the SILAC data alone produced a substantially higher composite score than PrInCE achieved (**Table 2-4**) and that EPIC also predicted up to five times as many complexes (with comparable or higher reliability) than PrInCE (**Table 2-4**).

	EPIC	**PrInCE**
Composite score	1.014	0.658
Maximum matching ratio (MMR)	0.267	0.059
Overlap score	0.327	0.369
Accuracy score	0.42	0.23
# Complexes	333	65

Table 2-4: Performance comparison with an existing approach (PrInCE)

2.3.2 Results

Since stably-associated components within a complex are expected to co-fractionate together, EPIC first computes pairwise protein profile similarity using up to eight correlation metrics

(Euclidean, Jaccard, Apex, Pearson, Pearson with Poisson noise, weighted cross correlation, mutual information, and Bayes correlation that emphasize different profile features. Positive and negative reference co-complex PPIs display distinct correlation distributions as shown below.

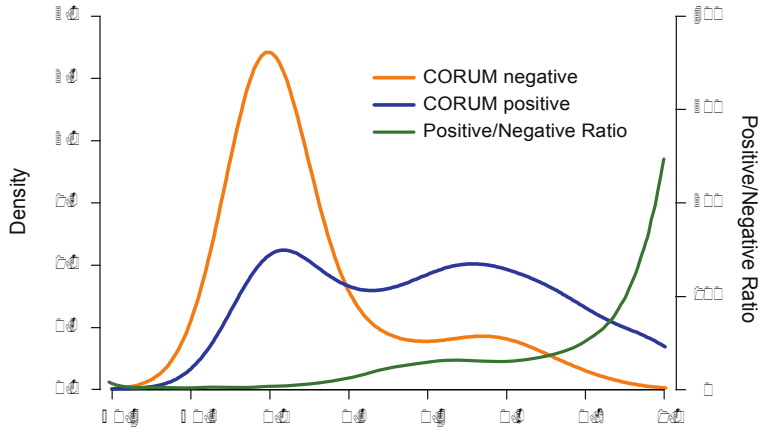

Figure 2-15: Co-elution profile similarity predicts PPIs. Plots showing the Pearson correlation coefficients (distribution density curves) obtained for a representative worm protein co-fractionation experiment; positive (CORUM derived; *blue*) and negative (randomized; *orange*) co-complex interactions, as well as the positive/negative ratio (*green*), are shown.

While it is not possible to pre-define a universally optimal combination of correlation metrics for all possible CF/MS experiments, EPIC provides default parameters tuned on comprehensive CF/MS data (described below), and can optimize settings for any given data set. To reduce computational time, proteins observed in only one fraction and protein pairs with co-fractionation correlation scores less than 0.5 are removed before generating a scored co-complex PPI vector for each input experiment. Multiple correlation vectors are then combined and input into a supervised machine-learning model that is both trained to predict new PPIs and benchmarked against reference positive (annotated) PPIs (i.e. co-complex relationships curated

in the CORUM, IntAct and GO databases) and negatives (i.e. combinations of proteins in distinct complexes).

To generate a comprehensive reference (gold standard) set for both training and benchmarking, EPIC retrieves species-specific complexes from the IntAct and GO complex databases. Since positive examples are limited for certain species, like *C. elegans*, the benchmark is supplemented by mapping annotated human protein complexes from the CORUM database based on stringent one-to-one orthology (InParanoid). To minimize redundancy and bias, complexes with the majority of subunits in common (overlap score >0.8) are merged, while large assemblies with 50+ members (e.g. ribosome) that could dominate learning are eliminated.

EPIC uses support vector machine (SVM) and random forest (RF) classifiers by default, but other algorithms can be substituted programmatically. Since CF/MS data is often incomplete (e.g. due to proteome under-sampling) or noisy (e.g. chance co-elution of unrelated proteins), EPIC can integrate additional supporting evidence (e.g. functional interactions inferred from co-expression, domain co-occurrence, and co-citation) from public sources such as GeneMANIA or STRING, thereby producing richer and more accurate interaction networks. To avoid circularity, functional interactions based on published PPIs are excluded. To ensure all complexes have CF/MS experimental support, those complexes inferred based solely on functional evidence are removed. Prediction performance is evaluated by 2-fold cross-validation (i.e. against an independent 'holdout' set of reference protein complexes).

Finally, EPIC applies network-partitioning to define complex membership. ClusterONE (Nepusz et al., 2012) is used by default, though other algorithms can be evaluated to optimize complex definition (Wiwie et al., 2015). Each cluster is compared to annotated complexes curated in CORUM, GO and IntAct, and overall performance is measured by three complementary evaluation metrics (maximum matching ratio, accuracy, and overlap score; as documented above), from which a single summary composite score is calculated to assign prediction quality (Nepusz et al., 2012)

We evaluated EPIC performance using a novel data set of 1,380 IEX-HPLC fractions generated for soluble worm protein extracts from mixed stage *C. elegans* cultures. Co-eluting proteins were acid-precipitated, alkylated and trypsin digested, and the resulting peptide mixtures analyzed by

precision Orbitrap MS. To optimize major EPIC parameters (MS search tool, set of profile correlation metric and machine learning classifier), we compared predicted complexes from each parameter setting (2,040 parameter combinations) against an independent benchmark of known complexes compiled from CORUM, IntAct and GO using composite score as the evaluation measure (as described in **Fig. 2-12**). Optimized parameters substantially improved the resulting composite score compared to previously used parameters (Havugimana et al., 2012; Wan et al., 2015) (**Fig. 2-16**).

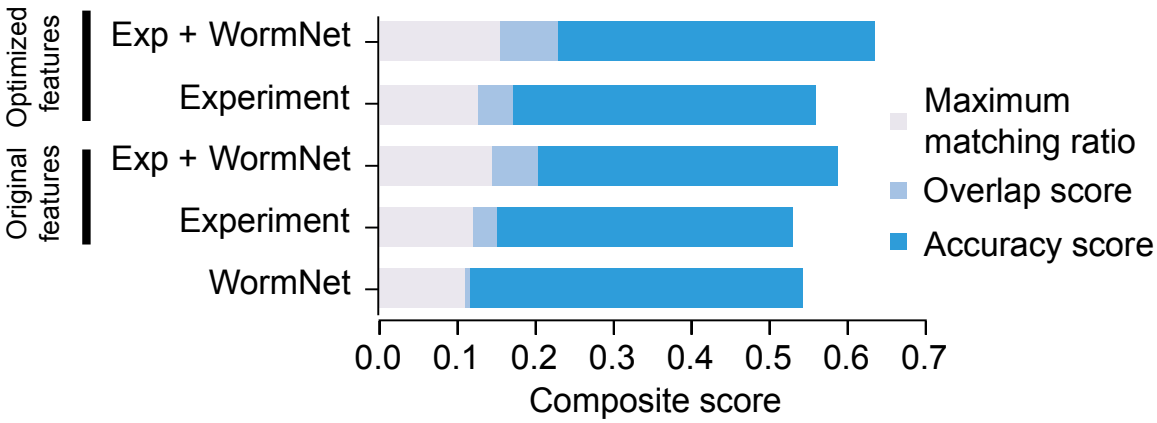

Figure 2-16: Bar chart shows predicted worm complex scores (maximal matching ratio, overlap and accuracy, the sum of which forms the composite score) using different combinations of experimental (CF/MS) data, functional evidence (WormNet) and correlation scores. "Original features" indicates results from the set of correlation metrics (parameters) used in previous publications, and "optimized features" indicates our newly optimized EPIC parameters.

We evaluated the performance benefit of integrating functional interactions with the CF/MS data, again based on composite score, and found that including GeneMANIA, STRING, or WormNet (Cho et al., 2014b) clearly improved performance (**Fig. 2-17**).

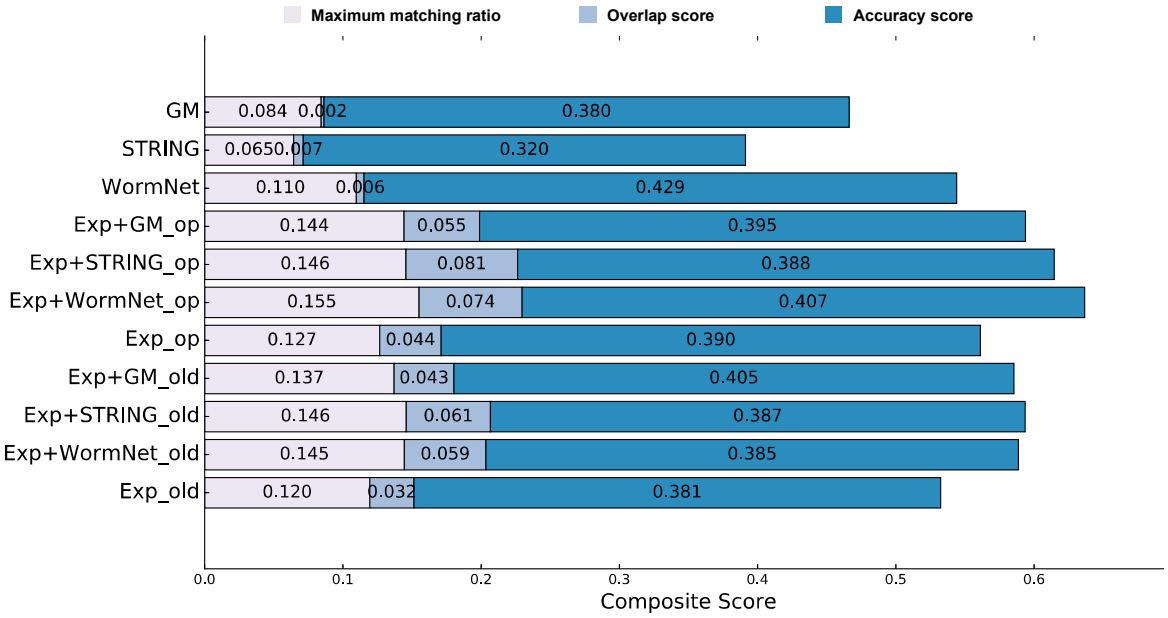

Figure 2-17: Composite score comparison for original and optimized features integrated with different sources functional evidence. Composite score analysis demonstrates that for predicting complexes, based on EPIC analysis of CF/MS data, integration of functional associations from WormNet outperforms STRING and GeneMANIA evidence. The analysis also shows an optimized set of EPIC-derived co-elution scores better predicts protein complex memberships than were reported previously.

It is also noted that functional evidence was not effective when used alone as input to predict complexes and PPIs (**Fig. 2-17** and **Fig. 2-18**).

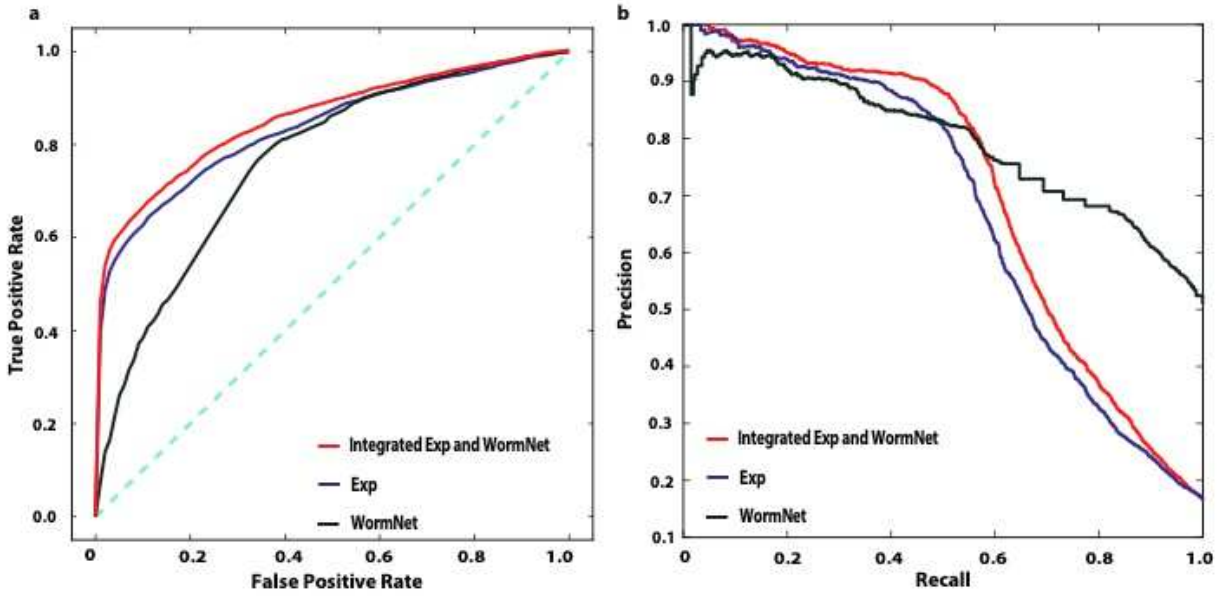

Figure 2-18: ROC curve and Precision-recall curve for co-complex PPI prediction from different input data. The plot demonstrates that the best co-complex interaction predictions were obtained after integrating experimental data with supporting functional evidence data (i.e. WormNet).

Since CF/MS studies consume considerable resources (e.g. LC/MS run time), we used EPIC to explore the 'cost/benefit' ratio of repeat biochemical fractionations by evaluating the relationship between prediction accuracy and the number of experiments performed. We calculated the average composite score by randomly sampling different numbers of co-fractionation experiments. Notably, while performance steadily improved as more data was acquired, prediction performance grew fastest over the first 2-4 separations (**Fig. 2-14**), suggesting an efficient lower bound (i.e. ~4 IEX-HPLC experiments) for study design.

2.3.3 Discussion

Current knowledge of the physical networks of cells and tissues remains limited for many species, particularly non-traditional animal models. The majority of known/curated protein

assemblies are annotated to mammals, whereas inference based on homology may not be the ideal for more distant organisms. CF/MS is an ideal experimental technology to address this, as it can be applied directly to any biological sample. However, CF/MS data is complex and challenging to process. We have developed the EPIC software to facilitate routine CF/MS analysis of native macromolecular assemblies in diverse contexts. EPIC provides optimized computational workflows, does not require expert computational skills to run, automates the entire data analysis process, and is applicable to diverse model systems. In practice, EPIC enables users to process their own data and supply their own manually curated reference protein complexes to optimize classifier training.

We have shown that EPIC predicts complexes with high accuracy, particularly if four or more biochemical separations are available. While transient or unstable macromolecules may not be efficiently detected by CF/MS, chemical cross-linking can potentially be beneficial (Liu et al., 2015), while other gentle separation techniques, such as isoelectric focusing (Pourhaghighi et al., submitted) and size-exclusion chromatography (Olinares et al., 2010), can provide complementary data. Regardless, to mitigate the false discovery rate, EPIC implements customizable data filtering procedures and can optionally integrate supporting independent functional evidence.

EPIC is both open source (https://github.com/BaderLab/EPIC) and compatible with disparate proteomic sampling techniques, including 'top-down' analysis of intact proteins (Tran et al., 2011) and sample multiplexing (isotopic labeling) (Werner et al., 2014) to map differential networks across conditions (Ideker and Krogan, 2012). To facilitate broader uptake, we provide an automatically executable Jupyter-based notebook along with a Docker container (https://hub.docker.com/r/baderlab/bio-epic/) encompassing all necessary scripts and packages, enabling easy installation, deployment and optimization on any operating system. The distributed version of EPIC has step-by-step instructions and a user-friendly interface that enables uploading of local user defined CF/MS data files and the graphical display of results.

Chapter 3

WormMap: a comprehensive map of soluble protein complexes in C. *elegans*

A paper has been published in Nature Methods (Hu et al., 2019), partially based on the content of this chapter. The work presented in this chapter was done by collaboration between Fraser and Emili labs. I received technical assistance from Dr. June Tan on C. *elegans* transgenic line generation. Mr. Eric Wolf helped on AP-MS experiments. Prof. Andrew Emili and Prof. Gary Bader co-supervised the project.

3 WormMap: a comprehensive map of soluble protein complexes in *C. elegans*

3.1 Introduction

Caenorhabditis elegans was first introduced as a model organism by Sydney Brenner for the study of developmental biology in the 1960s (Riddle et al., 1997). Decades later, it became a widely used experimental system with numerous advantages. It is easy to culture, and has a quick life cycle (~3 days), such that a single culture dish seeded with *E. coli* can generate thousands of worms (Riddle et al., 1997). Ideal for microscopic observation, *C. elegans* has transparent body, fixed cell lineages and defined tissues, and consistent cell positions, making it useful for the study of multicellular development. The complete ~97 Mb genome of *C. elegans,* comprising six chromosomes (5 autosomes and 1 sex chromosome) and around 19,000 genes, was published in 1998 (Consortium, 1998). In addition to its hermaphroditic reproductive mode, well suited to genetic screens, another special advantage is susceptibility to RNA interference (RNAi), in which *E. coli* expressing double-stranded RNAs targeting genes of interest are fed to *C. elegans* to disrupt function (Fire et al., 1998). This technique has been extensively used to study the roles of genes and pathways during development in *C. elegans*. These data are available via public databases, like WormBase (Harris et al., 2014) and modENCODE (Gerstein et al., 2010), which also collect sequence and expression data and other genomic information.

C. elegans has multiple developmental stages: embryogenesis, four postembryonic developmental larval stages (L1 to L4) and adult (Riddle et al., 1997). When food is diminished, *C. elegans* enters a quiescent dauer larval stage after the second larval molt (Riddle et al., 1997).

Gene functional association networks have been generated for *C. elegans*. For example, WormNet (Cho et al., 2014a) is a probabilistic gene function network generated for *C. elegans* based on Bayesian integration of high-throughput genomic datasets (e.g. gene co-expression, genetic interactions, etc) *(Lee et al., 2008a)*. Likewise, GeneMANIA (Montojo et al., 2014) couples linear regression and Gaussian field label propagation to integrate genomic datasets into

a composite network to predict gene functional associations for multiple model organisms (including *C. elegans*) (Mostafavi et al., 2008). Yet far fewer physical interactions have been reported in worm. Co-IP and AP-MS are generally too laborious for large-scale application to higher eukaryotes with multiple developmental stages and tissues. Surprisingly, bioinformatic predictions of physical interactions among worm proteins are also limited. To date, the largest PPIs study in *C. elegan*s, by the Vidal group in 2009, used Y2H technology to assess only around one fourth of the predicted *C. elegans* proteome, yielding 3,864 putative binary PPIs (Simonis et al., 2009). Few of these interactions have supporting biochemical evidence or define the membership of protein complexes.

In this thesis work, a more complete map of the physical organization of protein complexes in *C. elegans* is generated using data collected from CF/MS experiments with the aid of EPIC, named as WormMap. WormMap is fully supported by direct biochemical evidence. WormMap contains 612 putative complexes from a network of 16,098 high-confidence PPIs that encompassed 3,855 worm proteins, most of which have never been reported before. The resulting 'WormMap' reveals assemblies with links to disparate lineage-restricted processes, conserved animal systems and human disease. WormMap included novel subunits and assemblies unique to nematodes that we validated using orthogonal methods.

3.2 Material and methods

3.2.1 Perform co-fractionation experiments on worm lysate

The experimental details of performing CF/MS experiments on *C. elegans* have already been documented in section 2.2.1. In the following sections, only the extra validation experiments will be described. Computational functional enrichment analysis of protein complexes on WormMap will also be discussed.

3.2.2 Generating GFP tagged worm strains for AP/MS

To create GFP-tagged proteins for AP/MS experiments, *C. elegans* strains were grown and maintained at 20 °C on nematode growth media (NGM) plates seeded with *E. coli* strain OP50. Some strains (wild-type N2 and RW1596: *myo-3 (st386) stEx30 [myo-3p::GFP::myo-3 + rol-6(su1006)* (Campagnola et al., 2002)) were ordered from the CGC (https://cgc.umn.edu/). Extra-chromosomal array strains containing a C-terminal GFP translational fusion construct of F26E4.4, Y34B4A.6 and F13H8.2 were also generated in this study. For instance, the open reading frame and 617 bp promoter region of F26E4.4 (Dupuy et al., 2004) were amplified and cloned into the pPD95.75 vector (Fire Lab Vector Kit). The construct was then injected at 20 ng/µl along with pRF4 as a co-injection marker. Roller positive F2 animals were isolated and imaged to confirm the GFP expression (rol-6 was used as a co-injection marker). Mixed stage worms were harvested for AP/MS validation studies. All other GFP-tagged strains (Y34B4A.6 and F13H8.2) are generated in a similar fashion.

3.2.3 Affinity purification mass spectrometry validation

Affinity purification was performed essentially as described (Kwan et al., 2016) with minor modifications. Briefly, frozen cell pellets were re-suspended in high-salt NP-40 lysis buffer (10 mM Tris-HCl pH 8.0, 420ml NaCl, 0.1% NP-40) with protease and phosphatase inhibitors (Roche). After 3 freeze-thaw cycles, each lysate was briefly sonicated, treated with nuclease (Thermo Scientific Cat #88700), followed by centrifugation at 14,000 rpm. The resulting soluble protein extract was split for technical replicate purifications. Each lysate was incubated at 4 °C on a rotator with 1 µg of rabbit anti-GFP antibody (Thermo Scientific Cat #G10362) for 2 hrs, followed by incubation with 25 µl of Protein-G Dynabeads slurry for 1 hr. The beads were washed twice with low-salt buffer (10mM Tris-HCl pH 8.0, 100 mM NaCl) and bound proteins subsequently eluted (X 4) with 1% ammonium hydroxide pH 11. Recovered protein samples were dried, re-suspended in 50 mM ammonium bicarbonate, reduced with 5 mM DTT at 56 °C for 45 min and alkylated with 10 mM iodoacetamide at room temperature for 45 min in the dark. Trypsin digestion was performed overnight at 37 °C. Peptide samples were de-salted and re-

suspended in 1% formic acid and then analyzed by data dependent (top-15 MS2) acquisition on a Q Exactive HF mass spectrometer (Thermo Scientific) using a 90-minute gradient on the same HPLC system described above. The resulting MS spectra were searched with MSblender.

3.2.4 Disease and phenotype enrichment analysis

Since there is a lack of information available for Worm gene disease associations, we combined several human resources and mapped human gene names to worm gene names via 1:1 orthology using InParanoid. Gene disease associations were retrieved from the Online Mendelian Inheritance in Man (OMIM), UniProt, and ClinVar databases. However, OMIM only provides gene-disease associations, and thus we retrieved a mapping from gene name to UniProt identifier via the UniProt identifier mapping web service. Moreover, OMIM does not provide a classification system for their diseases and different OMIM IDs might describe the same disease (e.g. Alzheimer has multiple identifiers depending on the types). Thus, we mapped each OMIM disease identifier to their corresponding disease ontology identifier (DOID). In the final step, we combined the resulting data set with a set of DOID annotations for Worm genes from the WormBase database. For phenotype analysis, we annotated our protein complexes with phenotype information taken from WormBase. Statistical enrichment for both phenotype and disease was determined by Fisher exact test, and Benjamini-Hochberg procedure was applied for multiple testing corrections.

3.2.5 GO Enrichment

The Gene Ontology (GO) is a controlled vocabulary that describes genes by using three categories: molecular function, cellular component and biological process. We inferred enriched GO terms using the g:Profiler R package (Reimand et al., 2016). To ensure we only get significant hits we only considered GO terms with less than 500 proteins annotated to them, and the p-value was corrected by the conservative Bonferroni correction procedure.

3.3 Results

Using all 16 *C. elegans* co-fractionation experiments with optimized parameter settings and including functional interactions, EPIC predicted 16,098 high-confidence co-complex PPIs among 3,855 worm proteins (~25% of the nematode proteome), each directly supported by CF/MS data (at least one co-elution correlation score >0.5). Most (13,547) of these PPIs have not been reported before (compared to iRefWeb (Turner et al., 2010), BioGRID (Chatr-Aryamontri et al., 2017) or our previously generated Metazoan Complex Map (Wan et al., 2015)) (**Fig. 3-1;** The complete listing is available at https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-019-0461-4/MediaObjects/41592_2019_461_MOESM3_ESM.txt).

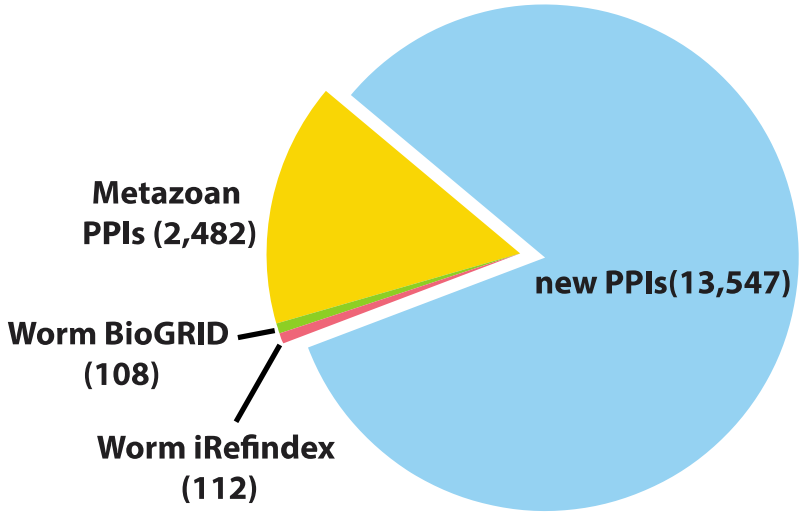

Figure 3-1: Pie chart showing overlap of predicted co-complex interactions with PPIs from BioGRID, iRefIndex and previously reported conserved metazoan complex map.

Partitioning the network using ClusterONE predicts 612 complexes (**Fig. 3-2a**) of which only 150 map to known assemblies in CORUM, GO, and IntAct. Most of the novel complexes appear to be clade-specific as only 89 are also found in the Metazoan Complex Map (The complete listing is available at https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-019-0461-4/MediaObjects/41592_2019_461_MOESM4_ESM.txt).

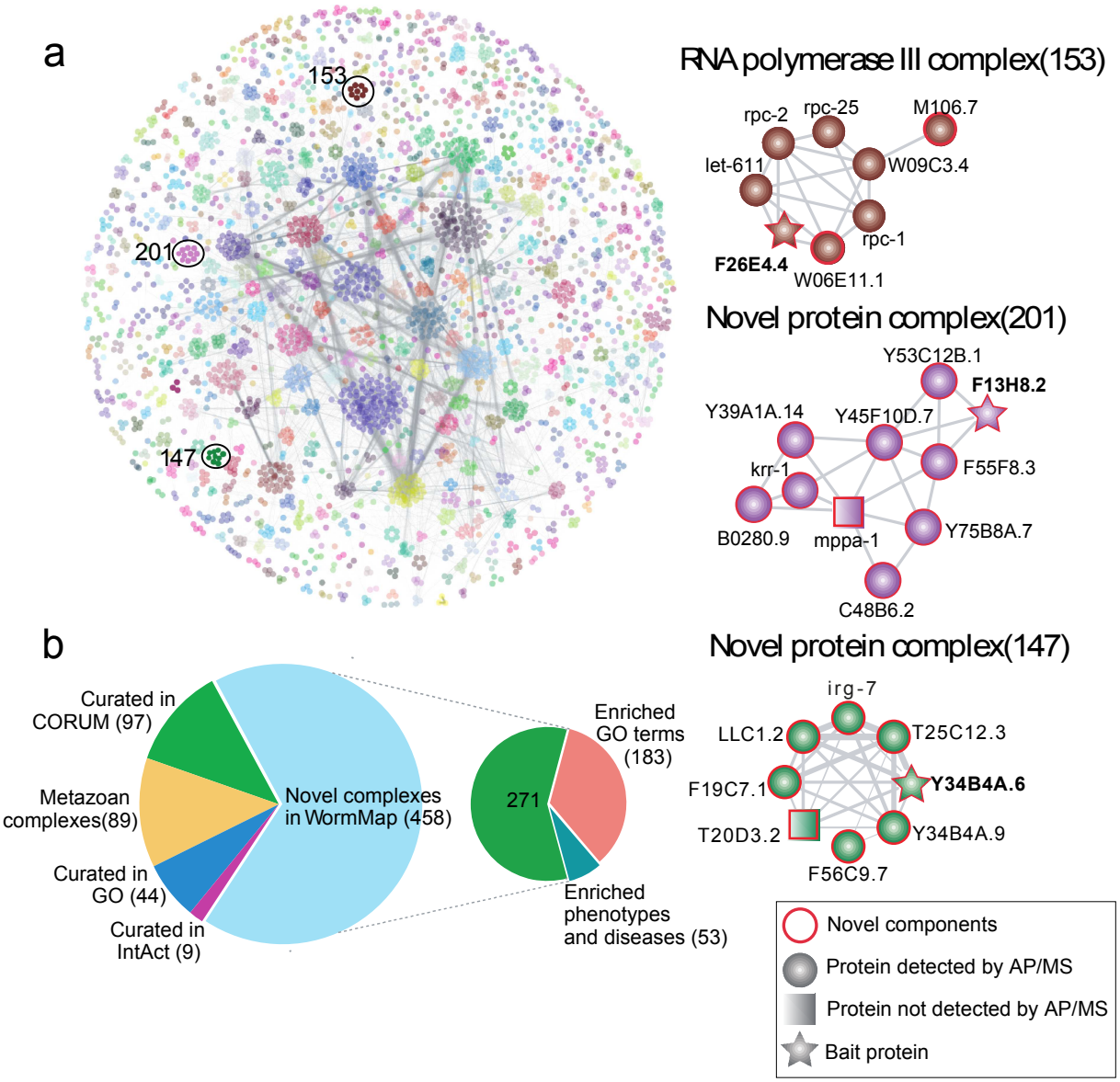

Figure 3-2: a) EPIC-derived WormMap. The left side shows the global overview of WormMap. Complexes validated using AP/MS are circled and AP/MS results are shown on the right, including novel components of the RNA polymerase III complex, as well as two novel complexes. Protein nodes are coloured according to complex assignments, with novel assemblies and components highlighted with red circles. Grey lines between proteins indicate interactions that are supported by strong co-elution evidence. Bait proteins are shown as stars, prey proteins as circles and undetected proteins as squares. Novel

components are indicated by a red node outline. AP/MS spectral counts are summarised in Supplementary Table 4. b) Pie charts showing the overlap of predicted worm complexes found by EPIC with previously known macromolecules (from CORUM, GO, IntAct, and the metazoan protein complex map (Wan et al., 2015)) and enrichment of putative novel assemblies for select biological function (GO terms), phenotype and/or disease associations.

We used multiple independent approaches to assess the accuracy of the predicted worm protein complexes. Experimentally, we used an established, orthogonal biochemical approach (AP/MS; see Online Methods) to validate both entirely novel assemblies as well as previously reported assemblies for which EPIC predicted unexpected new components (**Fig. 3-2a** and **Table 3-1**).

Complex (AP/MS validation)	APMS Bait	F13H8.2		F26E4.4.1		Y34B4A.6		
		replicate 1	*replicate 2*	*replicate 1*	*replicate 2*	*replicate 1*	*replicate 2*	protein name/description
RNA polymerase III (WormMap #153)	F26E4.4	0	0	18	17	0	0	Uncharacterized protein
	W06E11.1	0	0	11	9	0	0	Uncharacterized protein; predicted to have DNA-directed 5'-3' RNA polymerase activity
	rpc-1	0	0	36	29	0	0	DNA-directed RNA polymerase subunit (EC 2.7.7.6)
	M106.7	0	0	1	1	0	0	Uncharacterized protein; predicted to have catalytic activity and nucleotide binding activity
	rpc-25	0	0	3	3	0	0	RNA Polymerase, Class III (C)
	rpc-2	3	2	26	22	0	0	DNA-directed RNA polymerase subunit beta (EC 2.7.7.6)
	let-611	0	0	10	5	0	0	an ortholog of human POLR3C (RNA polymerase III subunit C
	W09C3.4	0	0	7	3	0	0	an ortholog of human POLR3F (RNA polymerase III subunit F)
Novel Protein Complex #201	F13H8.2	84	83	4	1	1	2	Uncharacterized protein; an ortholog of human WDR3 (WD repeat domain 3)
	B0280.9	5	4	0	0	0	0	U3 small nucleolar RNA-associated protein 18 homolog
	F55F8.3	38	33	0	0	0	0	Periodic tryptophan protein 2 homolog
	C48B6.2	4	7	0	0	0	0	Putative 40S ribosomal protein S4-like
	Y39A1A.14	10	10	0	0	0	0	Ribosomal RNA small subunit methyltransferase nep-1 (EC 2.1.1.-) (18S rRNA (pseudouridine-N1)-methyltransferase) (Ribosome biogenesis protein nep-1)
	Y45F10D.7	18	22	0	0	0	0	Uncharacterized protein; an ortholog of human WDR36 (WD repeat domain 36)
	Y75B8A.7	15	17	0	0	0	1	U3 small nucleolar ribonucleoprotein protein MPP10
	Y53C12B.1	54	55	0	0	0	0	Uncharacterized protein; an ortholog of human TBL3 (transducin beta like 3)
	krr-1	3	2	0	0	0	0	KRR1 small subunit processome component (KRR-R motif-containing protein 1)
Novel Protein Complex #147	Y34B4A.6	0	0	0	0	3	0	Uncharacterized protein
	Y34B4A.9	0	0	0	0	2	2	Uncharacterized protein
	F58C9.7	0	0	0	0	0	1	Uncharacterized protein
	F19C7.1	0	0	0	0	0	1	Uncharacterized protein
	LLC1.2	0	0	0	0	1	1	Uncharacterized protein
	irg-7	0	0	0	0	15	18	Protein irg-7 (Infection response protein 7)

	T25C12.3	0	0	0	0	22	16	Uncharacterized protein

Table 3-1: Results of AP/MS validation experiments. Table of spectral counts recorded in follow up AP/MS experiments, all performed in duplicate, for all co-purifying proteins identified with each bait protein (as indicated in header). A red protein name indicates either novel components assigned to a known complex (RNA polymerase III) or totally novel complexes (Complex 201 and 147). Bold numbers are spectral counts obtained for subunits predicted by EPIC that were also detected by AP/MS.

For example, we verified three new nematode-specific components (F26E4.4, W06E11.1 and M106.7) of the worm RNA polymerase III machinery, one of which (M106.7) has DNA and nucleotide binding activity (Mulder et al., 2003) (**Fig. 3-2a**). We also validated unc-15 as part of a large myosin complex, an association not reported in a public database or our training set, but has been observed in previous work (Kagawa et al., 1989). Likewise, we verified a predicted novel 10-member complex (**Fig. 3-2a**), for which most components have limited functional annotation in WormBase(Harris et al., 2010), suggesting an overlooked biological role. Two of the subunits (B0280.9 and krr-1) are orthologs of human small-subunit processome components involved in ribosomal biogenesis, suggesting a related function in nematodes. Intriguingly, another subunit, Y45F10D.7, is an ortholog of human WDR36, which is linked to primary open-angle glaucoma type 1G (GLC1G) (Monemi et al., 2005), potentially providing a mechanistic connection. We also confirmed another putative novel complex with eight protein components (**Fig. 3-2a**) containing mostly uncharacterized components according to UniProt (UniProt, 2015) and WormBase (Harris et al., 2010). Irg-7 is the only annotated subunit, with links to innate immunity and expression in the intestine (Yunger et al., 2017), suggesting a potential role in the host response to animal pathogens. Some interacting proteins identified by AP/MS with low counts, indicating a weak MS signal, were nonetheless consistent with co-elution evidence.

To assess the physiological significance of the putative worm assemblies, we analyzed the network of complexes for coherent biological functions (based on GO annotations), mutant phenotypes (based on information from WormBase (Harris et al., 2010)), or disease associations (based on orthology to human proteins in genetic disorder databases such as OMIM (Amberger et al., 2015) and HGMD (Stenson et al., 2014)). Strikingly, almost half of the novel complexes in WormMap were enriched for associations to essential processes, phenotypes or diseases (**Fig. 3-2b**). For example, knockdown of components of dozens of complexes either cause embryonic

lethality or sterility, and have links to cancer in humans, reinforcing the utility of EPIC for gaining fundamental mechanistic insight into large CF/MS data.

The GO term enrichment result is available at: https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-019-0461-4/MediaObjects/41592_2019_461_MOESM4_ESM.txt

The phenotypic enrichment result is available at: https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-019-0461-4/MediaObjects/41592_2019_461_MOESM7_ESM.xlsx

The disease enrichment result is available at: https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-019-0461-4/MediaObjects/41592_2019_461_MOESM7_ESM.xlsx

3.4 Discussion

The nematode *C. elegans* is a powerful model organism that has been extensively used to study different fundamental biological questions. However, knowledge of the physical interaction network supporting the development and cell biology of this animal is currently limited. The work I proposed here closes this gap by comprehensively identifying protein complexes present in *C. elegans*. In this chapter, we described how to use EPIC to map protein complexes in *C. elegans* based on CF/MS data, which has classically been studied using genetic methods, thereby revealing nematode-specific biochemical network adaptations. We integrated functional evidence with co-fractionation experimental data into EPIC to minimize co-elution and finalize the final network. It is argued that integrating functional evidence will reduce false negative PPIs, but may introduce bias towards well-studied proteins (Skinnider et al., 2018). While it is difficult to evaluate this bias, we note that many WormMap complexes, including those validated by AP/MS, contain uncharacterized proteins or proteins with diverse functional annotations, which suggests that EPIC is not strongly affected by this bias. Regardless, we believe WormMap will serve a valuable public resource for worm community.

CHAPTER 4

Applying deep learning to predict PPIs

All the work conducted in this chapter was done by myself. Prof. Andrew Emili and Prof. Gary Bader co-supervised the project.

4 Apply deep learning to predict PPIs

4.1 Introduction

The original implementation of EPIC requires calculating eight different correlation scores for each set of elution profiles between two different proteins. Additionally, optimizing the results of the final predicted protein complexes map across a range of parameters is computationally intensive. A deep neural network (or deep learning) approach may be able to overcome this problem by automatically learning patterns from elution profiles to make predictions without any feature engineering (correlation score calculations). In this chapter, I will use deep learning to perform PPI prediction using raw co-fractionation experimental data collected from our worm samples. Similar to most deep learning work, efforts are required for hyperparameter optimization before achieving satisfying results. In this work, a tree-structured Parzen Estimator Approach (TPE) (Bergstra et al., 2011) was used for optimizing hyperparameters (negative/positive ratio, PPI score cut-off, optimizer, learning rate, drop out probability and activation function). Briefly, the TPE algorithm randomly samples hyperparameters from an initial uniform distribution, and evaluates the loss function for each random set of inputs. And then the initial uniform distribution is replaced by a new distribution based on the results from regions of the sampled distribution that minimize the loss function. The advantage of using a deep neural network based approach to predict protein complexes is that we skip the feature-engineering step, which is time consuming to create and optimize. The downside is that the hyperparameter optimization step is time consuming and requires high performance computing resources with GPU hardware. However, the result generated by deep learning is not as good as EPIC judging by the composite score evaluation introduced before. Regardless, through our work, we introduce and provide a framework for using deep learning to predict protein complexes using co-fractionation data, and hopefully this work can serve as a good starting point for others to adopt this approach into their own research work.

4.2 Material and methods

4.2.1 Applying neural network to predict PPIs

The general workflow of using deep learning to predict protein complexes is very similar to the EPIC pipeline (**Fig. 4-1**). Instead of performing pair-wise correlation metric computation of protein elution profiles in individual CF/MS experiments for feature engineering, all CF/MS data are directly concatenated to form a master matrix for input into the learner. For example, if protein A has the vector reading *m* (protein elution profile) from the first CF/MS experiment and has another vector reading *n* from the second CF/MS experiment, we concatenate the two vectors as $m + n$, then the concatenated vector from all proteins are stacked row by row to form the master matrix. If multiple CF/MS experiments are performed, all the vector readings from protein A are concatenated first. In the case where protein A is not detected in one CF/MS experiment, a zero vector (the length of the zero vector is the same as the number of fractions collected in that particular CF/MS experiment) is used to represent the vector reading. The training set and evaluation sets of PPIs are generated in the same manner as the EPIC pipeline discussed in the previous chapters. The difference in the second part of this workflow is that a deep neural network is used as the machine learning classifier instead of random forest or support vector machine, and the input of the deep learning classifier is the raw CF/MS output of the two proteins from the master matrix instead of correlation scores after feature engineering. Also, functional evidence is not incorporated into the machine learning prediction, as we try to predict PPIs purely based on CF/MS experimental data using deep learning. The last step is the same as in the EPIC workflow – the predicted network is segmented using a clustering algorithm to generate protein complexes.

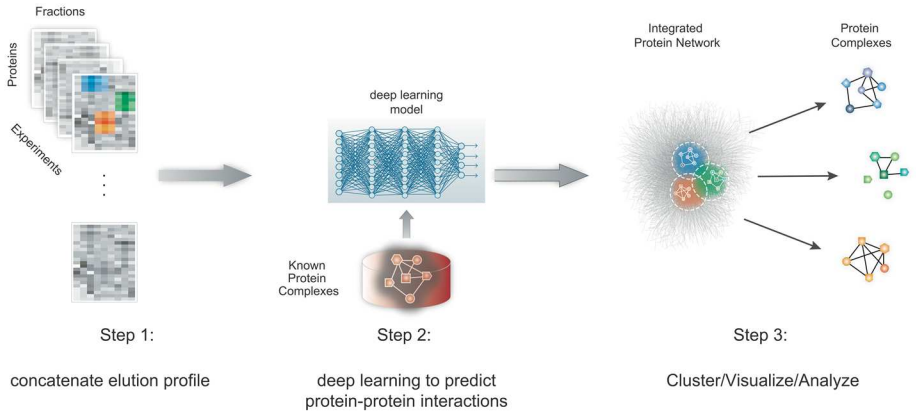

Figure 4-1: A modified computational workflow using deep-learning takes CF/MS data as input to predict protein complexes: (i) concatenate individual elution profile from individual co-fractionation experiment into a master elution profile (a master matrix); (ii) co-complex PPI scoring using deep learning model; (iii) prediction, clustering, and benchmarking of derived complexes.

Inspired by previous work of applying deep learning in computational biology to predict specific gene patterns (Alipanahi et al., 2015; Xiong et al., 2015) I started with the 1D convolutional neural network architecture ranging from three to five layers and using Max Pooling layers to reduce dimensionality and control overfitting. However, the accuracy of predictions from the convolutional neural network was low. I decided to switch to the standard multilayer perceptron (MLP) neural network to predict PPIs from CF/MS profile sequence data. In this work, the input data are elution profile vectors of two proteins across all CF/MS experiments, which are sequence data. After testing, we built our MLP model with the architecture as shown in the figure below:

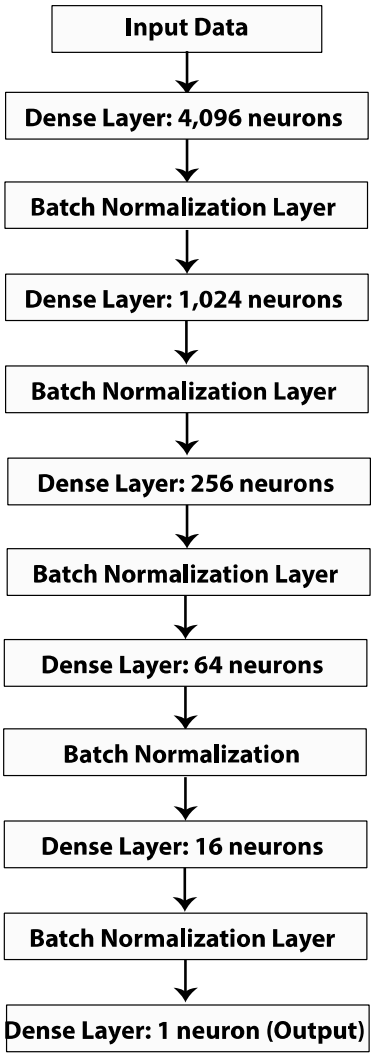
Input Data
Dense Layer: 4,096 neurons
Batch Normalization Layer
Dense Layer: 1,024 neurons
Batch Normalization Layer
Dense Layer: 256 neurons
Batch Normalization Layer
Dense Layer: 64 neurons
Batch Normalization
Dense Layer: 16 neurons
Batch Normalization Layer
Dense Layer: 1 neuron (Output)

Table 4-1: The architecture of the MLP model used in our prediction. In this architecture, the input layer is followed by six dense layers. Between each two dense layers, there is a batch normalization layer. The last dense layer uses sigmoid activation function to give a probabilistic output of prediction.

In this architecture, I also introduced a batch-normalization layer that performs normalization for each training mini-batch to minimize the internal covariate shift problem, so a higher learning rate and more flexible initialization parameters are permitted (Ioffe and Szegedy, 2015). In this thesis work, we have collected 1,380 IEX HPLC fractions in total for the WormMap dataset. To input this data to the MLP, we concatenate the protein elution profiles from two proteins formed by two 1,380 long numeric vectors to give a combined vector with the length of 2,760. The training set of data was generated as in the EPIC work (Chapter 2), with the positive data labeled as "1", and the negative data labeled as "0".

4.2.2 Hyper-parameter optimization by tree-structured Parzen Estimator Approach (TPE)

To achieve good results for deep learning, hyper-parameter tuning is a necessary step. In this work, a tree-structured Parzen estimator (TPE) is used to optimize the hyper-parameters for the deep neural network architecture described above (Bergstra et al., 2011). This Bayesian based optimization algorithm has been shown to perform better than other popular optimization algorithms in efficiency and accuracy, especially for deep neural networks (Bergstra et al., 2013). Like in any optimization problem, an objective loss function L is defined as shown below:

$$L(\theta) = -average\left[\sum_{n=1}^{5}\left(AUPRC_{\theta,n}\right)\right]$$

In the above formula, θ is the set of hyper-parameters used in the deep neural network model. *AUPRC* is the area under the PR curve for the *n*th fold under a five-fold cross validation scheme. The average of the *AUPRC*s for the five-fold cross validation is the value we try to maximize,

and the negative of the average of the *AUPRC*s is taken as the loss function. We pick *AUPRC* instead of area under *ROC* curve, because we used all available labeled training data (much more positive ones), as deep learning model training requires lots of data. The way I coded cross validation makes sure that the portion of positive and negative data in each fold is the same as the one in the original training data set. In general, a Bayesian optimization algorithm tries to follow the probability model for a surrogate function and pick the most likely set of parameters evaluated by the true loss function. The TPE algorithm first randomly initializes the set of hyper-parameters based on the pre-defined distribution of each parameter. Using a sequential model-based global optimization (SMBO) scheme, TPE replaces the initial distribution by a newly defined one using the formula defined below:

$$p(x|y) = \begin{cases} l(x); \ if \ y < y^* \\ g(x); \ otherwise \end{cases}$$

In the formula above, x is the set of parameters and y is the value of the objective function that is L as defined above. In the TPE algorithm, y^* is set as some quintile γ of the observed values of y, thus $\gamma = p(y < y^*)$. The TPE algorithm optimizes Expected Improvement (EI) that is defined as:

$$EI_{y^*}(x) = \int_{-\infty}^{y^*} (y^* - y) p(y|x)\, dy = \int_{-\infty}^{y^*} (y^* - y) \frac{p(x|y)}{p(x)} dy$$

since $\gamma = p(y < y^*)$, and we know:

$$p(x) = \int p(x|y)\, p(y) dy = \gamma l(x) + (1 - \gamma) g(x)$$

The formula of EI could be transformed as:

$$EI_{y^*}(x) = \frac{l(x)}{p(x)} \int_{-\infty}^{y^*} (y^* - y) p(y) dy = \frac{\gamma y^* l(x) - l(x) \int_{-\infty}^{y^*} y p(y) dy}{\gamma l(x) + (1 - \gamma) g(x)} \propto \left(\gamma + \frac{g(x)}{l(x)} (1 - \gamma) \right)^{-1}$$

To maximize the value of *EI*, the ideal x should be sampled with high probability under *l(x)* and low probability under *g(x)*. The TPE algorithm uses an Adaptive Parzen Estimator to yield models of *l(x)* and *g(x)* by placing density based on previously sampled K observation over

iterations. Each continuous hyper-parameter is pre-defined by a prior distribution specified by the user. In this work, we choose to optimize the following parameters:

Parameters	Prior
optimizer for training	Adam, Nadam, RMSprop
learning rate	Loguniform (-0.5, 0.5)
dropout rate	Uniform (0.35, 0.65)
activation functions	relu, elu, tanh, sigmoid
# non-zero fractions	Range (2, 10)
random seeds	Range (0, 5)

Table 4-2: Hyper-parameters optimized by TPE in this work. Note that this is a selection among many possible hyper-parameters that can be optimized.

4.3 Results

We used the TPE algorithm to optimize hyper-parameters as described above for 100 iterations. The set of parameters that achieved the highest average of the *AUPRC*s after five-fold cross validation is shown in Table 4-3.

Parameters	Prior
optimizer for training	Nadam
learning rate	0.828
dropout rate	0.45
activation functions	elu
# non-zero fractions	9
random seeds	0

Table 4-3: The optimized set of hyper-parameters after 100 iterations of TPE optimization.

With the optimized set of hyper-parameters, the ROC and PR curves based on the five-fold cross validation are shown in the figure below:

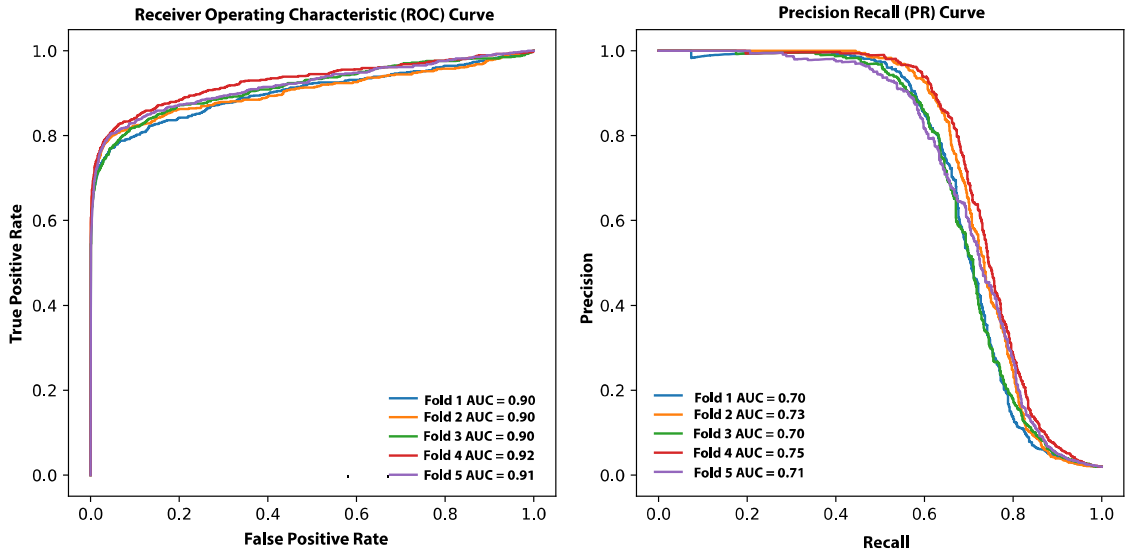

Figure 4-2: ROC and PR curves of the optimized set of hyper-parameters based on five-fold cross validation.

Based on the figure above, it is found that average *AUPRC* is about 0.70. The values of area under *PR* curves are consistent across all folds based on the cross validation results that suggest the deep learning model is robust The results are also consistent for area under *ROC* curves, although we didn't use it as the metric for hyper-parameter optimization. However, I noticed that the performance of the deep learning model is highly sensitive to hyper-parameter settings during parameter tuning. We then used the optimized set of hyper-parameters and all available training data to train the model. The trained deep learning model was applied to predict PPIs using all the available CF/MS data. If we set recall cutoff at 0.36, in total we obtained 56,189 PPIs and 752 protein complexes. The composite score for the resulting protein complexes is 0.4766 that is lower than the one obtained by EPIC (0.636). The results are briefly summarized in the table below:

Number of PPIs	56,189
Number of complexes	752
Maximum matching ratio	0.0806
Overlap score	0.0173
Accuracy score	0.3787
Composite score	0.4766

Table 4-4: The clustering results of overlapped PPIs between deep learning results and WormMap.

However, after checking the newly predicted list of PPIs, I found many well-known PPIs are only detected by the deep learning approach. A good example is the interaction between two subunits (cra-1 and natb-1) of N-terminal AceTyltransferase B (NatB) complex. Although the overlap between deep learning and EPIC is small (~2,200 PPIs), the preliminary result here already suggests deep learning and EPIC seem to predict different subsets of PPIs. A natural way to think of this problem is to treat the PPI prediction problem as an ensemble approach to combine the most confident PPIs from EPIC and deep learning and throw away the less confident ones. Ideally, in the future, more different deep learning architectures could be explored to generate a more comprehensive set of PPIs that covers both EPIC and deep learning.

CHAPTER 5

Thesis summary and future directions

5 Thesis summary and future directions

5.1 Thesis summary

Mapping PPIs or protein complexes is important to help us understand the cellular processes in a biological system. Through the guilt-by-association principle, the knowledge of protein complexes also helps annotate the functions of uncharacterized genes. Until now, using a systematic AP-MS approach, protein complexes have been mapped in many different biological systems including unicellular prokaryotic/eukaryotic model organisms, multi-cellular model organisms, human cell lines, diseased samples and host-pathogen infectious systems, which answered many important biological questions. However, the AP-MS approach is laborious and can only be applied to genetically tractable biological systems. CF/MS based protein complexes detection method has recently been developed, which requires no tagging on individual protein. The tagless CF/MS approach is fast, which makes charting an interactome map of a novel organism in a few months possible. However, bioinformatics analysis of CF/MS data to infer protein complexes while eliminating false positive PPIs is the key of this type of tagless screening method. The computational analysis part of this method is not trivial and requires large efforts consisting of feature engineering, training data preparation, machine learning inference and complex prediction. In this thesis work, we united all the computational steps and developed a new software tool named EPIC to automatically score, predict and evaluate protein complexes using CF/MS data. We also used EPIC to analyze CF/MS data collected from *C. elegans* to generate the first biochemically supported nematode protein complex map, termed WormMap. To demonstrate the reliability of this prediction, we used an orthogonal approach, AP-MS, to validate novel protein complexes from this map. We also showed that by using a deep neural network approach, we could infer PPIs without performing feature engineering, which provides a new direction for analyzing CF/MS data.

Chapter 1 starts with a brief introduction of how a mass spectrometry machine can be used in proteomics research. The major historical breakthrough is the development of liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) that makes peptide separation and protein identification much easier. Protein quantification methods in proteomics including both isotopic labeling and label-free approaches are also discussed in this chapter. I

also introduced the two most widely used large-scale PPI screening assays, Y2H and AP-MS, and reviewed how the AP-MS approach can be applied to study PPIs and complexes in different types of samples to answer different biology questions. In the second part of Chapter 1, I introduced the development of computational analysis methods in proteomics. There are two directions of computational research in proteomics. The first direction focuses on data acquisition: how to develop better algorithms and software tools to more efficiently detect and quantify proteins from the mass spectrometry machine signal. The second direction is how to use "smart" computational methods, including machine learning, to extract useful information from collected proteomics data to answer biological questions. In the end of Chapter 1, I documented mathematical details of the machine learning algorithms used in this thesis work. Chapter 2 can be divided into two parts: the CF/MS experimental procedures and the EPIC computational workflow. A CF/MS experiment consists of three main steps: fraction collection, LC-MS/MS analysis and co-fractionation profile generation. Using *C. elegans* as an example, I described the details of the cell lysate preparations, the setup of HPLC fractionation and the parameters of the LC-MS/MS detection system. I explored a pre-enrichment approach to capture a sub-proteome using affinity beads and showed it both improves proteome coverage and benefits the detection of low abundance proteins. The computational part of Chapter 2 (the software development of EPIC) is the major focus of this thesis, which includes elution profile correlation metric calculation, PPI inference by machine learning and protein complex prediction. EPIC uses many public databases to download supporting information automatically. For example, it automatically collects gold standard protein complexes from CORUM, IntAct and GO by mapping orthologous proteins to a target species using InParanoid with a stringent cutoff. Heavily overlapping protein complexes are removed or merged to minimize redundancy. Feature engineering is used for machine learning-based co-complex PPI prediction. Before calculating protein elution profile co-elution scores, several filtering steps were used to remove suspicious and weakly associated protein pairs to reduce computational cost. EPIC contains eight different similarity metrics to compute protein elution profile similarity to capture different aspects of co-elution patterns. It can also combine functional evidence data with co-fractionation experimental data to help increase prediction accuracy, where the functional evidence data can either be supplied by users or automatically downloaded from two popular databases (GeneMANIA and STRING) by supplying the target species taxonomy identifier. EPIC evaluates its prediction

performance at PPI and protein complex levels. In Chapter 2, I discussed how different evaluation metrics can be used to assess PPI and protein complex prediction using reference data. There are many parameters used in our prediction task. For example, in the WormMap project, we used multiple proteomics search and quantification methods, eight different correlation metrics to calculate protein elution profile similarity and two different machine-learning classifiers. We used a nested-cross validation approach to select the optimal combination of parameters to derive the final worm protein complex set. We also did some extra analysis to see how many co-fractionation experiments are needed to derive a protein complex map with an acceptable quality. Unsurprisingly, we found that using more co-fractionation experiments leads to better protein complex prediction, but four IEX experiments are a reasonable lower bound if resources (i.e. mass spectrometry time) are limited. EPIC is also compared with similar software (PrInCE) and we found that EPIC predicted more complexes at a higher quality judged by the composite score. Chapter 3 is about WormMap, the first biochemically supported protein complexes map derived by CF/MS experiments and EPIC. We mapped 612 putative protein complexes and 16,098 PPIs in this work. We benchmarked our WormMap with previously published or curated worm PPIs and protein complexes. We found that the majority of co-complex PPIs and complexes in WormMap are novel. To demonstrate the reliability of WormMap, we tagged three genes (F26E4.4, Y34B4A.6 and F13H8.2) in *C. elegans* with GFP and performed AP-MS on these three GFP-tagged strains. Through AP-MS experiments, we successfully validated two novel components in the well-known RNA polymerase III complex. We also validated two novel protein complexes with components mostly uncharacterized or annotated with diverse functions. In the end of Chapter 3, we performed enrichment analysis on the novel complexes set from WormMap and found many of them are related to phenotypes and diseases, which demonstrates the richness of biology captured by WormMap. Chapter 4 is about applying deep learning to predict protein complexes using CF/MS data without performing extensive feature engineering. I built a MLP based deep learning architecture and applied the TPE algorithm to optimize its hyper-parameters. The set of optimized hyper-parameters are used to finalize the deep learning model to predict PPIs using all available CF/MS data. The predicted PPIs network does not have a comparable quality to the one predicted by EPIC, but the overlapped set of PPIs between EPIC and deep learning gives very

high composite score after clustering, which suggests this set of PPIs are highly confident co-complex PPIs.

5.2 Future directions

5.2.1 Experimental advances on CF/MS pipeline

One of the goals of a co-fractionation MS experiment is to maximize the coverage of proteome detection to maximize the protein complexes identification. Meanwhile, the integrity of protein complexes should be well preserved during chromatographic separations. To achieve these goals, several additional experimental techniques can be explored in future research. For example, non-ionic detergents could be added to solubilize hydrophobic complexes (Babu et al., 2012), chemical cross-linkers can be used to stabilize labile assemblies (Liu et al., 2015), and organelle compartments can be enriched prior to HPLC co-fractionation. The chemical cross-linking strategy is one of the most promising directions for CF/MS pipeline. It is known some of the unstable complexes or transient interactions will not be preserved during chromatographic separation. Chemical cross-linking could solve this problem by stabilizing native protein complexes. However, most chemical cross-linking approaches nowadays suffer from a low labeling efficiency, which hinders its application in large-scale PPIs or protein complexes mapping (Liu et al., 2015; Walzthoeni et al., 2015). In other words, if more efficient cross-linking chemistry is developed, the combination of CF/MS and cross-linking could be a very powerful tool for mapping protein complexes. The current EPIC pipeline uses label-free technique (MS1 ion intensity or MS2 spectral counts) to quantify proteins from hundreds of fractions, which could be expensive on mass spectrometry machine time. As discussed in Chapter 1, TMT based quantification approach can provide attractive multiplexing advantage. With advanced mass spectrometry machine, TMT multiplexing can allow quantification up to 10 or 11 different conditions in one mass spectrometry run (Werner et al., 2014). In this way, fractions generated by the EPIC pipeline could be individually labeled but analyzed in one mass spectrometry run, which will dramatically reduce mass spectrometry machine time.

5.2.2 Protein complex map for other biological systems.

One of the most attractive properties of CF/MS and EPIC platform is it provides a standardized and unified pipeline to map protein complexes across different biology samples, especially for those genetically intractable systems. We expect biologists from other laboratories can adopt this pipeline and software into their own research to map protein complexes in diverse animal species. Meanwhile, tissue specificity plays a central role in human disease and physiology, thus mapping tissue specific protein network can help us to understand how genes change functions across different tissues and their relationship among different diseases (Greene et al., 2015). Coupled with gentle tissue separation technique, CF/MS and EPIC can be utilized to map tissue specific protein complexes in an efficient manner. Indeed, our laboratory has already started mapping brain specific protein complexes map in mouse that may be the first study of this kind. The preliminary results have already shed light on its importance in neuron disease related biology. Protein complexes closely related to Autism have been identified and functionally validated.

5.2.3 Protein network dynamic by CF/MS and statistical inference

Most available protein interaction networks capture the overall static landscape of biological systems. However, the structure and architecture might be dramatically re-wired under different biological conditions (Ideker and Krogan, 2012). To derive a differential protein network accurately and efficiently is a proteomics “dream”. But this work is very difficult and only very few works have been published, with many limitations (Collins et al., 2013; Lambert et al., 2013). Our CF/MS pipeline provides an attractive approach to tackle the differential network problem, however, how to automatically and rigorously detect how the interactome (or complexes) changes under different conditions is challenging. Furthermore, since mass spectrometry machine time is expensive, the straightforward hypothesis test based approach that fits a known distribution based on central limit theorem and large number theorem and normalizes using the pooled variance to detect sample difference (Listgarten et al., 2007) might be resource consuming (experimentally, it needs multiple co-fractionation runs) and not very applicable. Thus, it is necessary to develop new high-dimensional interaction statistical models

to track how protein complexes change under different conditions using CF/MS data without multiple runs while preserving the statistical power.

5.2.4 Using Natural Language Processing (NLP) techniques to annotate protein complexes

The combination of CF/MS and EPIC software described in this thesis work can be used to map protein complexes across many different biological systems. Although the knowledge of protein complexes is important, linking physical associations of proteins (complexes) to related phenotypes is not easy. Previous papers have attempted to use a Bayesian predictor to assign diseases to human protein complexes (Lage et al., 2007). However, this work only focuses on human protein complexes, and the generated dataset is static. Expanding this work to multiple model organisms to relate the richness of mapped protein complexes and their associated phenotypes would be very useful for the biology community. Inspired by the many techniques of NLP and previous applications in system biology (Jurafsky and Martin, 2009; Kveler et al., 2018), a statistical NLP model/software could be made to extract useful information from PubMed to statistically assign phenotypes to complexes. Also, since most annotated protein complexes are human complexes nowadays, EPIC uses a stringent orthologous mapping strategy to map human protein complexes to protein complexes of target species. This works well for most species but can be problematic in some cases. For example, if the organism under study is evolutionally distant from human, the direct orthologous mapping might generate inaccurate complexes. An automated NLP pipeline could also help solve this problem by collecting protein complex information for target species from literature. Our group has published a deep learning based NLP pipeline that could extract useful biochemical information from large amount of literature can be potentially useful for protein complexes generation in EPIC (Giorgi and Bader, 2019).

5.2.5 Large scale directed signaling protein-protein interaction network

So far all the networks discussed are undirected, which means the proteins are known to interact with each other either physically or functionally, but the direction or the causality is unclear or missing. Understanding directions in protein network is informative: many protein interactions represent components in signaling pathways and adding direction or causality information to a protein network can help us classify activators and inhibitors in a pathway and better understand the signaling pathway architecture. Building a protein network is not a trivial task, and causality inference in a network is an even more difficult task. In 2005, a small causal signaling network derived from single cell measurements was published (Sachs et al., 2005). In this paper, the levels of 11 different phospho-proteins and phospho-lipids were recorded using flow-cytometry under different conditions. All the data were collected and used to train a Bayesian network (a type of graphical probabilistic network). A Bayesian network is a powerful machine learning technique and can combine protein-signaling information as a joint probability distribution over proteins. The interaction direction is encoded as conditional probability in the joint distribution. Many of the established directed protein interactions were confirmed by previous publications or validated by new experiments performed by the authors. This work is promising in that it demonstrates an approach to modeling directed PPI networks using a probabilistic graphic model. An interesting extension to this work would be to scale up the phosho-signalling protein measurement using quantitative mass spectrometry based proteomics experiments, by which we could measure thousands of signaling proteins simultaneously across different conditions to build a much more comprehensive causal protein network.

References:

Aebersold, R., Agar, J.N., Amster, I.J., Baker, M.S., Bertozzi, C.R., Boja, E.S., Costello, C.E., Cravatt, B.F., Fenselau, C., Garcia, B.A.*, et al.* (2018). How many human proteoforms are there? Nature chemical biology *14*, 206-214.

Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell *92*, 291-294.

Alex, K., Sutskever, I., and Hinton, G.E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. 1097--1105.

Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature biotechnology *33*, 831-838.

Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic acids research *43*, D789-798.

Arabidopsis Interactome Mapping, C. (2011). Evidence for network evolution in an Arabidopsis interactome map. Science *333*, 601-607.

Babu, M., Bundalovic-Torma, C., Calmettes, C., Phanse, S., Zhang, Q., Jiang, Y., Minic, Z., Kim, S., Mehla, J., Gagarinova, A.*, et al.* (2018). Global landscape of cell envelope protein complexes in Escherichia coli. Nature biotechnology *36*, 103-112.

Babu, M., Krogan, N.J., Awrey, D.E., Emili, A., and Greenblatt, J.F. (2009). Systematic characterization of the protein interaction network and protein complexes in Saccharomyces cerevisiae using tandem affinity purification and mass spectrometry. Methods Mol Biol *548*, 187-207.

Babu, M., Vlasblom, J., Pu, S., Guo, X., Graham, C., Bean, B.D., Burston, H.E., Vizeacoumar, F.J., Snider, J., Phanse, S.*, et al.* (2012). Interaction landscape of membrane-protein complexes in Saccharomyces cerevisiae. Nature *489*, 585-589.

Bennett, E.J., Rush, J., Gygi, S.P., and Harper, J.W. (2010). Dynamics of cullin-RING ubiquitin ligase network revealed by systematic quantitative proteomics. Cell *143*, 951-965.

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. Paper presented at: 25th Annual Conference on Neural Information

Processing Systems (NIPS 2011) (Granada, Spain: Neural Information Processing Systems Foundation).

Bergstra, J., Yamins, D., and Cox, D.D. (2013). Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28 (Atlanta, GA, USA: JMLR.org), pp. I-115-I-123.

Bishop, C.M. (2006). Pattern recognition and machine learning (New York: Springer).

Boekema, E.J., Hifney, A., Yakushevska, A.E., Piotrowski, M., Keegstra, W., Berry, S., Michel, K.P., Pistorius, E.K., and Kruip, J. (2001). A giant chlorophyll-protein complex induced by iron deficiency in cyanobacteria. Nature *412*, 745-748.

Brohee, S., and van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. Bmc Bioinformatics *7*, 488.

Butland, G., Peregrin-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N.*, et al.* (2005). Interaction network containing conserved and essential protein complexes in Escherichia coli. Nature *433*, 531-537.

Campagnola, P.J., Millard, A.C., Terasaki, M., Hoppe, P.E., Malone, C.J., and Mohler, W.A. (2002). Three-dimensional high-resolution second-harmonic generation imaging of endogenous structural proteins in biological tissues. Biophys J *82*, 493-508.

Celis, J.E., Gromov, P., Ostergaard, M., Madsen, P., Honore, B., Dejgaard, K., Olsen, E., Vorum, H., Kristensen, D.B., Gromova, I.*, et al.* (1996). Human 2-D PAGE databases for proteome analysis in health and disease: http://biobase.dk/cgi-bin/celis. FEBS Lett *398*, 129-134.

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N.K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A.*, et al.* (2017). The BioGRID interaction database: 2017 update. Nucleic acids research *45*, D369-D379.

Cho, A., Shin, J., Hwang, S., Kim, C., Shim, H., Kim, H., Kim, H., and Lee, I. (2014a). WormNet v3: a network-assisted hypothesis-generating server for Caenorhabditis elegans. Nucleic acids research *42*, W76-W82.

Cho, A., Shin, J., Hwang, S., Kim, C., Shim, H., Kim, H., Kim, H., and Lee, I. (2014b). WormNet v3: a network-assisted hypothesis-generating server for Caenorhabditis elegans. Nucleic acids research *42*, W76-82.

Collins, B.C., Gillet, L.C., Rosenberger, G., Rost, H.L., Vichalkovski, A., Gstaiger, M., and Aebersold, R. (2013). Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. Nature methods *10*, 1246-1253.

Consortium, C.e.S. (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. Science *282*, 2012-2018.

Cox, J., Hein, M.Y., Luber, C.A., Paron, I., Nagaraj, N., and Mann, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Molecular & cellular proteomics : MCP *13*, 2513-2526.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nature biotechnology *26*, 1367-1372.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V., and Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. Journal of proteome research *10*, 1794-1805.

Diament, B.J., and Noble, W.S. (2011). Faster SEQUEST searching for peptide identification from tandem mass spectra. Journal of proteome research *10*, 3871-3879.

Drew, K., Muller, C.L., Bonneau, R., and Marcotte, E.M. (2017). Identifying direct contacts between protein complex subunits from their conditional dependence in proteomics datasets. PLoS Comput Biol *13*, e1005625.

Dupuy, D., Li, Q.R., Deplancke, B., Boxem, M., Hao, T., Lamesch, P., Sequerra, R., Bosak, S., Doucette-Stamm, L., Hope, I.A.*, et al.* (2004). A first version of the Caenorhabditis elegans Promoterome. Genome Res *14*, 2169-2175.

Eng, J.K., Fischer, B., Grossmann, J., and Maccoss, M.J. (2008). A fast SEQUEST cross correlation algorithm. Journal of proteome research *7*, 4598-4602.

Eng, J.K., McCormack, A.L., and Yates, J.R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom *5*, 976-989.

Fields, S., and Song, O. (1989). A novel genetic system to detect protein-protein interactions. Nature *340*, 245-246.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806-811.

Florens, L., Washburn, M.P., Raine, J.D., Anthony, R.M., Grainger, M., Haynes, J.D., Moch, J.K., Muster, N., Sacci, J.B., Tabb, D.L.*, et al.* (2002). A proteomic view of the Plasmodium falciparum life cycle. Nature *419*, 520-526.

Gatlin, C.L., Eng, J.K., Cross, S.T., Detter, J.C., and Yates, J.R., 3rd (2000). Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. Analytical chemistry *72*, 757-763.

Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B.*, et al.* (2006). Proteome survey reveals modularity of the yeast cell machinery. Nature *440*, 631-636.

Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M.*, et al.* (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature *415*, 141-147.

Gerstein, M.B., Lu, Z.J., Van Nostrand, E.L., Cheng, C., Arshinoff, B.I., Liu, T., Yip, K.Y., Robilotto, R., Rechtsteiner, A., Ikegami, K.*, et al.* (2010). Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science *330*, 1775-1787.

Gessulat, S., Schmidt, T., Zolg, D.P., Samaras, P., Schnatbaum, K., Zerweck, J., Knaute, T., Rechenberger, J., Delanghe, B., Huhmer, A.*, et al.* (2019). Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nature methods *16*, 509-518.

Giorgi, J.M., and Bader, G.D. (2019). Towards reliable named entity recognition in the biomedical domain. Bioinformatics.

Gnatt, A.L., Cramer, P., Fu, J., Bushnell, D.A., and Kornberg, R.D. (2001). Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. Science *292*, 1876-1882.

Goll, J., and Uetz, P. (2006). The elusive yeast interactome. Genome Biol *7*, 223.

Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E., Sealfon, S.C.*, et al.* (2015). Understanding multicellular function and disease with human tissue-specific networks. Nature genetics *47*, 569-576.

Guruharsha, K.G., Rual, J.F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., Beekman, C., Wong, C., Rhee, D.Y., Cenaj, O.*, et al.* (2011). A protein complex network of Drosophila melanogaster. Cell *147*, 690-703.

Harris, T.W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., Chen, W.J., De La Cruz, N., Davis, P., Duesbury, M., Fang, R.*, et al.* (2010). WormBase: a comprehensive resource for nematode research. Nucleic acids research *38*, D463-467.

Harris, T.W., Baran, J., Bieri, T., Cabunoc, A., Chan, J., Chen, W.J., Davis, P., Done, J., Grove, C., Howe, K.*, et al.* (2014). WormBase 2014: new views of curated biology. Nucleic acids research *42*, D789-793.

Hart, G.T., Lee, I., and Marcotte, E.R. (2007). A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. Bmc Bioinformatics *8*, 236.

Havugimana, P.C., Hart, G.T., Nepusz, T., Yang, H., Turinsky, A.L., Li, Z., Wang, P.I., Boutz, D.R., Fong, V., Phanse, S.*, et al.* (2012). A census of human soluble protein complexes. Cell *150*, 1068-1081.

Hebert, A.S., Richards, A.L., Bailey, D.J., Ulbrich, A., Coughlin, E.E., Westphall, M.S., and Coon, J.J. (2014). The one hour yeast proteome. Molecular & cellular proteomics : MCP *13*, 339-347.

Hein, M.Y., Hubner, N.C., Poser, I., Cox, J., Nagaraj, N., Toyoda, Y., Gak, I.A., Weisswange, I., Mansfeld, J., Buchholz, F.*, et al.* (2015). A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell *163*, 712-723.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B.*, et al.* (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. IEEE Signal Processing Magazine *29*, 82-97.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K.*, et al.* (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature *415*, 180-183.

Hu, L.Z., Goebels, F., Tan, J.H., Wolf, E., Kuzmanov, U., Wan, C., Phanse, S., Xu, C., Schertzberg, M., Fraser, A.G.*, et al.* (2019). EPIC: software toolkit for elution profile-based inference of protein complexes. Nature methods.

Hu, P., Janga, S.C., Babu, M., Diaz-Mejia, J.J., Butland, G., Yang, W., Pogoutse, O., Guo, X., Phanse, S., Wong, P.*, et al.* (2009). Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. PLoS biology *7*, e96.

Huttlin, E.L., Bruckner, R.J., Paulo, J.A., Cannon, J.R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M.P., Parzen, H.*, et al.* (2017). Architecture of the human interactome defines protein communities and disease networks. Nature *545*, 505-509.

Huttlin, E.L., Ting, L., Bruckner, R.J., Gebreab, F., Gygi, M.P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K.*, et al.* (2015a). The BioPlex Network: A Systematic Exploration of the Human Interactome. Cell *162*, 425-440.

Huttlin, E.L., Ting, L., Bruckner, R.J., Gebreab, F., Gygi, M.P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K.*, et al.* (2015b). The BioPlex Network: A Systematic Exploration of the Human Interactome. Cell *162*, 425-440.

Ideker, T., and Krogan, N.J. (2012). Differential network biology. Molecular systems biology *8*, 565.

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37 (Lille, France: JMLR.org), pp. 448-456.

Jager, S., Cimermancic, P., Gulbahce, N., Johnson, J.R., McGovern, K.E., Clarke, S.C., Shales, M., Mercenne, G., Pache, L., Li, K.*, et al.* (2011). Global landscape of HIV-human protein complexes. Nature *481*, 365-370.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science *302*, 449-453.

Jurafsky, D., and Martin, J.H. (2009). Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition, 2nd edn (Upper Saddle River, N.J.: Pearson Prentice Hall).

Kagawa, H., Gengyo, K., McLachlan, A.D., Brenner, S., and Karn, J. (1989). Paramyosin gene (unc-15) of Caenorhabditis elegans. Molecular cloning, nucleotide sequence and models for thick filament structure. J Mol Biol *207*, 311-333.

Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. In arXiv e-prints.

Kislinger, T., Cox, B., Kannan, A., Chung, C., Hu, P., Ignatchenko, A., Scott, M.S., Gramolini, A.O., Morris, Q., Hallett, M.T., *et al.* (2006). Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. Cell *125*, 173-186.

Kislinger, T., Rahman, K., Radulovic, D., Cox, B., Rossant, J., and Emili, A. (2003). PRISM, a generic large scale proteomic investigation strategy for mammals. Molecular & cellular proteomics : MCP *2*, 96-106.

Kristensen, A.R., Gsponer, J., and Foster, L.J. (2012). A high-throughput approach for measuring temporal changes in the interactome. Nature methods *9*, 907-909.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., *et al.* (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature *440*, 637-643.

Kveler, K., Starosvetsky, E., Ziv-Kenet, A., Kalugny, Y., Gorelik, Y., Shalev-Malul, G., Aizenbud-Reshef, N., Dubovik, T., Briller, M., Campbell, J., *et al.* (2018). Immune-centric network of cytokines and cells in disease context identified by computational mining of PubMed. Nature biotechnology *36*, 651-659.

Kwan, J., Sczaniecka, A., Heidary Arash, E., Nguyen, L., Chen, C.C., Ratkovic, S., Klezovitch, O., Attisano, L., McNeill, H., Emili, A., *et al.* (2016). DLG5 connects cell polarity and Hippo signaling protein networks by linking PAR-1 with MST1/2. Genes Dev *30*, 2696-2709.

Kwon, T., Choi, H., Vogel, C., Nesvizhskii, A.I., and Marcotte, E.M. (2011). MSblender: A probabilistic approach for integrating peptide identifications from multiple database search engines. Journal of proteome research *10*, 2949-2958.

Lage, K., Karlberg, E.O., Storling, Z.M., Olason, P.I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tumer, Z., Pociot, F., Tommerup, N., *et al.* (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. Nature biotechnology *25*, 309-316.

Lambert, J.P., Ivosev, G., Couzens, A.L., Larsen, B., Taipale, M., Lin, Z.Y., Zhong, Q., Lindquist, S., Vidal, M., Aebersold, R., *et al.* (2013). Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition. Nature methods *10*, 1239-1245.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436-444.

Lee, I., Ambaru, B., Thakkar, P., Marcotte, E.M., and Rhee, S.Y. (2010a). Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. Nature biotechnology *28*, 149-156.

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

Lee, I., Date, S.V., Adai, A.T., and Marcotte, E.M. (2004). A probabilistic functional network of yeast genes. Science *306*, 1555-1558.

Lee, I., Lehner, B., Crombie, C., Wong, W., Fraser, A.G., and Marcotte, E.M. (2008a). A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. Nature genetics *40*, 181-188.

Lee, I., Lehner, B., Crombie, C., Wong, W., Fraser, A.G., and Marcotte, E.M. (2008b). A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. Nat Genet *40*, 181-188.

Lee, I., Lehner, B., Vavouri, T., Shin, J., Fraser, A.G., and Marcotte, E.M. (2010b). Predicting genetic modifier loci using functional gene networks. Genome Res *20*, 1143-1153.

Leuenberger, P., Ganscha, S., Kahraman, A., Cappelletti, V., Boersema, P.J., von Mering, C., Claassen, M., and Picotti, P. (2017). Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. Science *355*.

Listgarten, J., Neal, R.M., Roweis, S.T., Wong, P., and Emili, A. (2007). Difference detection in LC-MS data for protein biomarker discovery. Bioinformatics *23*, e198-204.

Liu, F., Rijkers, D.T., Post, H., and Heck, A.J. (2015). Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. Nature methods *12*, 1179-1184.

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. Science *285*, 751-753.

Markham, K., Bai, Y., and Schmitt-Ulms, G. (2007). Co-immunoprecipitations revisited: an update on experimental concepts and their implementation for sensitive interactome

investigations of endogenous proteins. Analytical and Bioanalytical Chemistry *389*, 461-473.

Michalski, A., Damoc, E., Hauschild, J.P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M., and Horning, S. (2011). Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. Molecular & cellular proteomics : MCP *10*, M111 011015.

Miyagi, M., and Rao, K.C. (2007). Proteolytic 18O-labeling strategies for quantitative proteomics. Mass Spectrom Rev *26*, 121-136.

Monemi, S., Spaeth, G., DaSilva, A., Popinchalk, S., Ilitchev, E., Liebmann, J., Ritch, R., Heon, E., Crick, R.P., Child, A.*, et al.* (2005). Identification of a novel adult-onset primary open-angle glaucoma (POAG) gene on 5q22.1. Hum Mol Genet *14*, 725-733.

Montojo, J., Zuberi, K., Rodriguez, H., Bader, G.D., and Morris, Q. (2014). GeneMANIA: Fast gene network construction and function prediction for Cytoscape. F1000Research *3*, 153.

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., and Morris, Q. (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol *9 Suppl 1*, S4.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P.*, et al.* (2003). The InterPro Database, 2003 brings increased coverage and new features. Nucleic acids research *31*, 315-318.

Murphy, K.P. (2012). Machine learning a probabilistic perspective. In Adaptive computation and machine learning series (Cambridge, MA: MIT Press), pp. xxix, 1067 p.

Nahnsen, S., Bielow, C., Reinert, K., and Kohlbacher, O. (2013). Tools for label-free peptide quantification. Molecular & cellular proteomics : MCP *12*, 549-556.

Navarro, P., Kuharev, J., Gillet, L.C., Bernhardt, O.M., MacLean, B., Rost, H.L., Tate, S.A., Tsou, C.C., Reiter, L., Distler, U.*, et al.* (2016). A multicenter study benchmarks software tools for label-free proteome quantification. Nature biotechnology *34*, 1130-1136.

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nature methods *9*, 471-472.

Old, W.M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K.G., Mendoza, A., Sevinsky, J.R., Resing, K.A., and Ahn, N.G. (2005). Comparison of label-free methods for quantifying human proteins by shotgun proteomics. Molecular & cellular proteomics : MCP *4*, 1487-1502.

Olinares, P.D., Ponnala, L., and van Wijk, K.J. (2010). Megadalton complexes in the chloroplast stroma of Arabidopsis thaliana characterized by size exclusion chromatography, mass spectrometry, and hierarchical clustering. Molecular & cellular proteomics : MCP *9*, 1594-1615.

Ong, S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Molecular & cellular proteomics : MCP *1*, 376-386.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N.*, et al.* (2014). The MIntAct project-IntAct as a common curation platform for 11 molecular interaction databases. Nucleic acids research *42*, D358-D363.

Orphanides, G., Lagrange, T., and Reinberg, D. (1996). The general transcription factors of RNA polymerase II. Genes Dev *10*, 2657-2683.

Pankow, S., Bamberger, C., Calzolari, D., Martinez-Bartolome, S., Lavallee-Adam, M., Balch, W.E., and Yates, J.R., 3rd (2015). F508 CFTR interactome remodelling promotes rescue of cystic fibrosis. Nature *528*, 510-516.

Park, S.K., Venable, J.D., Xu, T., and Yates, J.R., 3rd (2008). A quantitative analysis software tool for mass spectrometry-based proteomics. Nature methods *5*, 319-322.

Ratushny, V., and Golemis, E. (2008). Resolving the network of cell signaling pathways using the evolving yeast two-hybrid system. Biotechniques *44*, 655-662.

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., and Vilo, J. (2016). g:Profiler-a web server for functional interpretation of gene lists (2016 update). Nucleic acids research *44*, W83-89.

Riddle, D.L., Blumenthal, T., Meyer, B.J., and Priess, J.R. (1997). Introduction to C. elegans. In C elegans II, D.L. Riddle, T. Blumenthal, B.J. Meyer, and J.R. Priess, eds. (Cold Spring Harbor (NY)).

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Seraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. Nature biotechnology *17*, 1030-1032.

Rolland, T., Tasan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R.*, et al.* (2014). A proteome-scale map of the human interactome network. Cell *159*, 1212-1226.

Rost, H.L., Liu, Y., D'Agostino, G., Zanella, M., Navarro, P., Rosenberger, G., Collins, B.C., Gillet, L., Testa, G., Malmstrom, L.*, et al.* (2016). TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. Nature methods *13*, 777-783.

Rout, M.P., Aitchison, J.D., Suprapto, A., Hjertaas, K., Zhao, Y., and Chait, B.T. (2000). The yeast nuclear pore complex: composition, architecture, and transport mechanism. J Cell Biol *148*, 635-651.

Rozenblatt-Rosen, O., Deo, R.C., Padi, M., Adelmant, G., Calderwood, M.A., Rolland, T., Grace, M., Dricot, A., Askenazi, M., Tavares, M.*, et al.* (2012). Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins. Nature *487*, 491-495.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.W. (2010). CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic acids research *38*, D497-501.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning internal representations by error propagation. In Parallel distributed processing: explorations in the microstructure of cognition, vol 1, E.R. David, L.M. James, and C.P.R. Group, eds. (MIT Press), pp. 318-362.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., and Nolan, G.P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. Science *308*, 523-529.

Sanchez-Taltavull, D., Ramachandran, P., Lau, N., and Perkins, T.J. (2016). Bayesian Correlation Analysis for Sequence Count Data. Plos One *11*, e0163595.

Shah, P.S., Link, N., Jang, G.M., Sharp, P.P., Zhu, T., Swaney, D.L., Johnson, J.R., Von Dollen, J., Ramage, H.R., Satkamp, L.*, et al.* (2018). Comparative Flavivirus-Host Protein Interaction Mapping Reveals Mechanisms of Dengue and Zika Virus Pathogenesis. Cell *175*, 1931-1945 e1918.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.

Simonis, N., Rual, J.F., Carvunis, A.R., Tasan, M., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Sahalie, J.M., Venkatesan, K., Gebreab, F.*, et al.* (2009). Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. Nature methods *6*, 47-54.

Sinha, A., Huang, V., Livingstone, J., Wang, J., Fox, N.S., Kurganovs, N., Ignatchenko, V., Fritsch, K., Donmez, N., Heisler, L.E.*, et al.* (2019). The Proteogenomic Landscape of Curable Prostate Cancer. Cancer Cell *35*, 414-427 e416.

Skinnider, M.A., Stacey, R.G., and Foster, L.J. (2018). Genomic data integration systematically biases interactome mapping. PLoS Comput Biol *14*, e1006474.

Sonnhammer, E.L., and Ostlund, G. (2015). InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic acids research *43*, D234-239.

Sowa, M.E., Bennett, E.J., Gygi, S.P., and Harper, J.W. (2009). Defining the human deubiquitinating enzyme interaction landscape. Cell *138*, 389-403.

Stacey, R.G., Skinnider, M.A., Scott, N.E., and Foster, L.J. (2017). A rapid and accurate approach for prediction of interactomes from co-elution data (PrInCE). Bmc Bioinformatics *18*, 457.

Stefely, J.A., Kwiecien, N.W., Freiberger, E.C., Richards, A.L., Jochem, A., Rush, M.J.P., Ulbrich, A., Robinson, K.P., Hutchins, P.D., Veling, M.T.*, et al.* (2016). Mitochondrial protein functions elucidated by multi-omic mass spectrometry profiling. Nature biotechnology *34*, 1191-1197.

Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A., and Cooper, D.N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet *133*, 1-9.

Stiernagle, T. Maintenance of C. elegans. In WormBook : the online review of C elegans biology, T.C.e.R. Community, ed. (WormBook).

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P.*, et al.* (2017). The STRING database in 2017: quality-

controlled protein-protein association networks, made broadly accessible. Nucleic acids research *45*, D362-D368.

The Gene Ontology, C. (2017). Expansion of the Gene Ontology knowledgebase and resources. Nucleic acids research *45*, D331-D338.

Thul, P.J., Akesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Bjork, L., Breckels, L.M.*, et al.* (2017). A subcellular map of the human proteome. Science *356*.

Ting, Y.S., Egertson, J.D., Bollinger, J.G., Searle, B.C., Payne, S.H., Noble, W.S., and MacCoss, M.J. (2017). PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. Nature methods *14*, 903-908.

Tiwary, S., Levy, R., Gutenbrunner, P., Salinas Soto, F., Palaniappan, K.K., Deming, L., Berndl, M., Brant, A., Cimermancic, P., and Cox, J. (2019). High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. Nature methods *16*, 519-525.

Tran, J.C., Zamdborg, L., Ahlf, D.R., Lee, J.E., Catherman, A.D., Durbin, K.R., Tipton, J.D., Vellaichamy, A., Kellie, J.F., Li, M.*, et al.* (2011). Mapping intact protein isoforms in discovery mode using top-down proteomics. Nature *480*, 254-258.

Tsou, C.C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A.C., and Nesvizhskii, A.I. (2015). DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. Nature methods *12*, 258-264, 257 p following 264.

Turner, B., Razick, S., Turinsky, A.L., Vlasblom, J., Crowdy, E.K., Cho, E., Morrison, K., Donaldson, I.M., and Wodak, S.J. (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. Database (Oxford) *2010*, baq023.

Tyanova, S., Temu, T., and Cox, J. (2016). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc *11*, 2301-2319.

UniProt, C. (2015). UniProt: a hub for protein information. Nucleic acids research *43*, D204-212.

von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., and Bork, P. (2005). STRING: known and predicted protein-protein

associations, integrated and transferred across organisms. Nucleic acids research *33*, D433-437.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. Nature *417*, 399-403.

Walzthoeni, T., Joachimiak, L.A., Rosenberger, G., Rost, H.L., Malmstrom, L., Leitner, A., Frydman, J., and Aebersold, R. (2015). xTract: software for characterizing conformational changes of protein complexes by quantitative cross-linking mass spectrometry. Nature methods *12*, 1185-1190.

Wan, C., Borgeson, B., Phanse, S., Tu, F., Drew, K., Clark, G., Xiong, X., Kagan, O., Kwan, J., Bezginov, A.*, et al.* (2015). Panorama of ancient metazoan macromolecular complexes. Nature *525*, 339-344.

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

Wehrens, R., Melssen, W., Buydens, L., and de Gelder, R. (2005). Representing structural databases in a self-organizing map. Acta Crystallogr B *61*, 548-557.

Werner, T., Sweetman, G., Savitski, M.F., Mathieson, T., Bantscheff, M., and Savitski, M.M. (2014). Ion coalescence of neutron encoded TMT 10-plex reporter ions. Analytical chemistry *86*, 3594-3601.

White, C.A., Oey, N., and Emili, A. (2009). Global quantitative proteomic profiling through 18O-labeling in combination with MS/MS spectra analysis. Journal of proteome research *8*, 3653-3665.

Williams, E.G., Wu, Y., Jha, P., Dubuis, S., Blattmann, P., Argmann, C.A., Houten, S.M., Amariuta, T., Wolski, W., Zamboni, N.*, et al.* (2016). Systems proteomics of liver mitochondria function. Science *352*, aad0189.

Wiwie, C., Baumbach, J., and Rottger, R. (2015). Comparing the performance of biomedical clustering methods. Nature methods *12*, 1033-1038.

Wu, Y., Williams, E.G., Dubuis, S., Mottis, A., Jovaisaite, V., Houten, S.M., Argmann, C.A., Faridi, P., Wolski, W., Kutalik, Z.*, et al.* (2014). Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population. Cell *158*, 1415-1430.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R.*, et al.* (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806.

Yates, J.R., 3rd, Eng, J.K., McCormack, A.L., and Schieltz, D. (1995). Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Analytical chemistry *67*, 1426-1436.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N.*, et al.* (2008). High-quality binary protein interaction map of the yeast interactome network. Science *322*, 104-110.

Yunger, E., Safra, M., Levi-Ferber, M., Haviv-Chesner, A., and Henis-Korenblit, S. (2017). Innate immunity mediated longevity and longevity induced by germ cell removal converge on the C-type lectin domain protein IRG-7. PLoS genetics *13*, e1006577.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S.*, et al.* (2014). Proteogenomic characterization of human colon and rectal cancer. Nature *513*, 382-387.

Zhang, Y., Fonslow, B.R., Shan, B., Baek, M.C., and Yates, J.R., 3rd (2013). Protein analysis by shotgun/bottom-up proteomics. Chem Rev *113*, 2343-2394.

Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nature methods *12*, 931-934.

Zuberi, K., Franz, M., Rodriguez, H., Montojo, J., Lopes, C.T., Bader, G.D., and Morris, Q. (2013). GeneMANIA prediction server 2013 update. Nucleic acids research *41*, W115-122.