A machine learning approach to decipher protein-protein interactions in human plasma to facilitate the characterization of metabolic pathways

Emily Hashimoto-Roth

Thesis submitted to the University of Ottawa

In partial fulfillment of the requirements for the

MSc degree in Biochemistry specializing in Bioinformatics

Department of Biochemistry, Microbiology and Immunology

Faculty of Medicine

University of Ottawa

To my grandparents,

John Junichi Hashimoto and Lois Yuriko Nakashima.

Acknowledgements

Having joined Dr. Lavallée-Adam's lab as an undergraduate student with zero research experience, I can only describe my time under his supervision as a rich learning experience – academically and professionally. In the lab, he taught me to be an early career bioinformatician in the proteomics and mass spectrometry fields and, professionally, he has opened the door to numerous community involvement opportunities. Because of these opportunities, I can now say that I have colleagues all around the world – USA, UK, Australia, Japan, China, Sweden, India, and a few others! My experience as a Master's student was only further enriched when Dr. Bennett agreed to co-supervise my project. As a member of her lab, I was able to learn about lipidomics and metabolomics from the perspective of researchers that run the workbench experiments (and I got to run a couple samples on her QTRAP 5500 mass spectrometer – *Charron*)! Both Dr. Lavallée-Adam and Dr. Bennett have been unconditionally supportive, providing me with the guidance I needed to become a more confident young scientist. I owe them my deepest appreciation and gratitude.

My appreciation also extends to our collaborators at the *Institut de recherches cliniques de Montréal* in Dr. Benoit Coulombe's lab, for providing me with the data I used to develop our computational methodologies and the NSERC-CREATE MATRIX graduate trainee program.

Much of my gratitude also goes out to my parents, Grace Hashimoto and Danny Roth, who supported me endlessly and tirelessly, especially when my studies became an entirely work-from-home endeavour. Last, but certainly not least, my close friend, Nina Hadžimustafić, continuously supported me while also pursuing her MSc and, now, her MD. Our friendship has always been a source of love and encouragement that I will cherish forever.

Abstract

Immunoprecipitation coupled to mass spectrometry (IP-MS) methods are often used to identify protein-protein interactions (PPIs) in biological samples. While these approaches are prone to false-positive identifications through contamination and antibody non-specific binding, their results can be filtered by combining the use of negative controls and computational modelling. However, such filtering does not effectively detect false-positive interactions when IP-MS is performed on human plasma samples, given a higher propensity for non-specific interactions. Therein, proteins cannot be overexpressed or inhibited, and existing modelling algorithms are not adapted for execution without such controls. Hence, we introduce MAGPIE, a novel machine learning-based approach for identifying PPIs in human plasma using IP-MS, which leverages negative controls that include antibodies targeting proteins not known to be present in human plasma. Unsupervised learning algorithms are first applied to label-free MS quantification data to identify a set of high-quality negative controls that can be used for false-positive interaction modelling. MAGPIE then uses a logistic regression classifier to assess the reliability of PPIs detected in IP-MS experiments using antibodies targeting known plasma proteins. When applied to five IP-MS experiments, our algorithm identified 68 PPIs with an FDR of 20%. MAGPIE significantly outperformed a state-of-the-art PPI discovery tool, detecting three times more interactions at half the FDR. PPIs identified by MAGPIE are further supported by known or predicted interactions in the STRING PPI repository. Finally, our approach provides an unprecedented ability to detect human plasma PPIs, enabling a better understanding of biological processes in plasma.

Table of Contents

List of abbreviations

PPI	Protein-protein interaction
IP	Immunoprecipitation
LC	Liquid chromatography
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
DI	Direct infusion
DIA	Data-independent acquisition
SPA	Shotgun proteome analysis
LFQ	Label-free quantification
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
mRNA	Messenger RNA
Y2H	Yeast-two-hybrid
GAL4	Galactose-responsive transcription factor
PfCA	Protein fragment complementation assay
TAP	Tandem affinity purification
CBP	Calmodulin binding protein
TEV	Tobacco etch virus
IgG	Immunoglobulin G
MAGPIE	Machine learning assessment with logistic regression of protein-protein interactions
SAINT	Significance Analysis of INTeractome
CRAPome	Contaminant repository for affinity purification-mass spectrometry data
FDR	False discovery rate
PCA	Principal component analysis
t-SNE	*t*-distributed stochastic neighbour embedding
UMAP	Uniform manifold approximation and projection
CE	Cross-entropy
LOCO	Leave-one-control-out
GO	Gene ontology
MR	Mutual rank
CNDP1	Beta-Ala-His dipeptidase
KLK6	Kallikrein-6
SNCA	Alpha-synuclein
PCSK9	Proprotein convertase subtilisin/kexin type 9
HA	Hemagglutinin
LC3B	Microtubule-associated proteins 1A/1B light chain 3B
METTL23	Methyltransferase-like protein 23
RPAP2	RNA polymerase II-associated protein 2

List of tables

List of figures

1 Introduction

1.1 Proteomics and protein-protein interactions

Proteins are biomolecules that largely govern cellular function, but seldomly do they work alone. Their interactions, be they stable or transient, determine the physiological outcome of many biological mechanisms. Therefore, the ability to identify protein-protein interactions within cells, tissues, and organisms is essential to characterize and understand innate biological function. Such research falls under the field of study known as proteomics, defined by the large-scale study of proteins[1,2]. Over the last three decades, the proteomics field has reached an intersection of disciplines: biochemistry and computer science. This intersection, representing the advent of bioinformatics for proteomics, introduces the sophisticated computational analyses of large biological datasets compiled from high-throughput laboratory experiments into mass spectrometry analyses[3]. These include, but are not limited to, algorithms for identification, quantification, and functional annotation of proteins and their interactions. It is these comprehensive analyses that give researchers the means to identify protein complexes and machineries, discern their interacting components, and characterize their functions and effects on given biological processes[4]. This further gives rise to the ability of studying protein-protein interactions associated with disease, wherein endogenous proteins behave abnormally, or protein-protein interactions become perturbed, and alter a healthy biological state[5–7]. This discipline is not without challenges, such as complexity in experimental design or overwhelming outputs of dynamic data. Therefore, as more proteomic experiments are performed, and as public data repositories grow larger, it is imperative that computational tools are developed to continue pushing research efforts beyond their current limitations.

1.2 Protein-protein interactions mapping efforts

There have been numerous notable contributions for mapping protein-protein interactions in model organisms and some human cell lines to help unveil their underlying biology. To no small feat, groups have attempted to map the entire protein-protein interaction network in yeast (*sp. Saccharomyces cerevisiae*) on multiple occassions[8–10]. In budding yeast, both a protein kinase and phosphatase interactome have also been mapped[11]. Similarly, the protein-protein interaction network in nematodes (*sp. Caenorhabditis elegans*) has also been largely mapped[12]. In humans, a network of protein interactions has been characterized for RNA polymerase II, a key enzyme in eukaryotic transcription of mRNA[13], and a proximity-dependent interaction network, Cell Map, has also been made available[14]. Quantitative proteomics has been used to define the interactome network topology of HeLa cell lines, proposing the importance of interaction stoichiometry for creating complete networks[15,16]. Finally, the BioPlex protein-protein interaction network is revered as one of the largest networks for interactions characterized in humans, composed of more than 23,000 interactions[17]. This non-exhaustive list exists because of the many experimental techniques that have been developed to discover protein-protein interactions.

1.2.1 Early experimental techniques for isolating proteins and their interactors

Experimental protocols for discovering protein-protein interactions have increased in number over the last 40 years. In the early days, the development of the binary yeast-two-hybrid (Y2H) system was a novel achievement, leveraging advances in genetic engineering, for identifying interacting proteins[18]. The system works by utilizing the galactose-responsive transcription factor (*GAL4*), an essential enzyme in *S. cerevisiae* for activating transcription of

galactose-consuming genes. *GAL4* can be fragmented into two functional domains: an N-terminal domain that binds to the upstream activating sequence for a given reporter gene and a C-terminal domain to activate the complex and induce transcription. The N-terminal domain, also referred to as the DNA binding domain, is often expressed by a known bait protein, while the C-terminal domain, or activation domain, is expressed by potential protein prey interactors. If the bait-prey pairs constitute a bona fide interaction, they will come into physical contact. This allows for the binding and activation domain fragments to come within proximity and induce transcription of the reporter gene. The Y2H system quickly became the archetypical assay for discovering interacting proteins, giving rise to systems such as the protein-fragment complementation assay (PfCA)[19,20], which functions using the same logic of bringing two functional domains together to induce a visual response. However, the PfCA methods instead requires the bait-prey pairs to be covalently linked to fragments of a reporter protein that will, for example, fluoresce if the fragments are brought together. These systems have facilitated many successful protein-protein interactions mappings. For instance, researchers have recently used the Y2H system to screen and characterize small-molecule inhibitors for the known interaction between tumor protein P53 and E3 ubiquitin-protein ligase Mdm2, to enhance the tumour-suppressing ability of the P53 protein[21]. Even more recently, the Y2H system was used to identify interacting proteins in Western diamondback rattlesnake *(sp. Crotalus atrox*) venom, a complex cocktail of proteins, that may serve as a basis for future development of new snake bite treatments[22].

Despite these successes, there are still several limitations associated to the Y2H system and other split protein assays[23,24]. Notably, false-positive identifications are prevalent when screening resulting reporter gene colonies, many of which are due to the inability to model native

environments when implementing these assays. In other words, when screening interactors using a library, biologically irrelevant interactions are often detected between proteins that are not normally located in the same cellular compartment. Additionally, because they are binary in nature, the Y2H and PfCA systems can only detect a small number of putative interactions at a time and putative interactors must be chosen prior to executing the assay. These limitations are largely resolved with mass spectrometry-based screening of protein-protein interactions, as they encompass various high-throughput methodologies accompanied by computational tools for filtering false-positive identifications in established cell lines[25,26].

1.2.2 Mass spectrometry-based screening of protein-protein interactions

Mass spectrometry is the gold-standard technology for proteomics research and the screening of protein-protein interactions is no exception. Often, immunoprecipitation coupled to mass spectrometry (IP-MS) is used for such high-throughput screening[27]. These versatile strategies require a protein of interest to artificially express a molecular tag, through DNA recombination. This tag can then be targeted with high specificity by an antibody, allowing for the protein of interest to be isolated from solution in complex with its interactors, enzymatically digested into peptides, and identified, as well as quantified by mass spectrometry. Moreover, these methodologies are often supplemented by liquid chromatography (LC) that separates peptides before entering the mass spectrometer and favours their detection[28]. LC columns are composed of a stationary and mobile phase, for which the peptides flowing through will have different affinities. Such affinities cause peptides to be retained for varying amounts of time within the column, thus staggering their elution into the mass spectrometer and improving their detectability. In fact, understanding and predicting the retention time of peptides for unique LC

methodologies is, in its own right, a major research area. For example, Pelletier et al. include peptide retention time prediction as a key element for their algorithm, MealTime-MS, which implements a machine learning model for the real-time identification of proteins in mass spectrometry experiments[29].

Typically, the antibodies used in an IP-MS experiment have been precisely engineered to recognize a given molecular tag to optimize its specificity and the overall experiment. The tandem affinity purification (TAP) tag is widely used for isolating a protein of interest with its interactors because its construct allows for a two-step sequential purification process prior to quantification[30,31]. From its N-terminal, the TAP tag is composed of a calmodulin binding peptide (CBP), followed by a tobacco etch virus (TEV) cleavage site, and IgG binding domain. This dual-affinity system is more sensitive than the previously described Y2H system and allows for the isolation and purification of multi-component, complex interactions (including indirect interactions) and fewer non-specific protein bindings[32]. In addition to the TAP tag, the FLAG tag is a commonly used molecular tag for IP-MS experiments[33]. Being only eight amino acids in length (Asp-Tyr-Lys-Asp-Asp-Asp-Asp-Lys), this short sequence is likely to minimize structural alteration to the protein of interest, which may otherwise have a negative impact on the IP experiment's capacity to isolate bona fide interactors. Furthermore, with a single purification step, the FLAG tag can typically achieve a greater degree of protein-protein interaction detection sensitivity than the TAP tag, at the expense of the detection of more false-positive interactions. While the use of these molecular tags is considered a standard approach for isolating interacting proteins, an important drawback to note is their limited ability to detect transient interactions. This is especially true for the TAP tag because of the two-step purification process required. Nonetheless, these systems – and variations thereof – have enabled numerous mappings of

protein-protein interaction networks, such as that for the human transcription machinery[34] as well as the human Oct4 protein, a key protein in the regulation of embryonic stem cell pluripotency[35,36]. Further progress in the development of IP-MS systems dawned proximity labelling-based mass spectrometry, for the identification of proximal proteins. One such system is BioID[37]. This system requires a protein of interest to express a BirA tag, which biotinylates proximal proteins. The biotinylated protein of interest and its biotinylated putative interactors are then purified using standard biotin-affinity capture. Unlike the TAP and FLAG systems, BioID has a greater capacity for detecting weak, transient, and indirect interactions. Furthermore, because biotinylation is a rare naturally occurring protein modification, the BioID system is considered to have high selectivity. However, a proximal protein detected with BioID may not be physically interacting with a given protein of interest. Hence, careful interpretation is necessary when exploring BioID results. Another useful approach for detecting protein-protein interactions is the cross-linking strategy[38,39]. Here, interacting proteins are covalently bonded prior to enzymatic digestion. The cross-linked peptides can then be enriched, and the resulting quantification data is information about the direct physical contact of the interacting proteins, providing insight into the structural biology of these interaction pairs[40]. Cross-linking strategies are especially effective at differentiating direct from indirect interactions that were previously reported. Ultimately, both the BioID and cross-linking approaches for detecting protein-protein interactions help meet the needs where classic IP-MS systems fall short, experimentally[41].

It is clear that the choice of detection and purification system massively influences experimental design, and these methods can be even further optimized to target historically difficult proteins to quantify, such as transmembrane proteins[42]. With each of these techniques, the purified protein and its interactors are identified and quantified by mass spectrometry.

Typically, tandem mass spectrometry (MS/MS)[27] is performed exploiting two mass analyzers are used instead of one. MS/MS requires the enzymatically digested proteins to be ionized prior to entering the first mass analyzer by, for example, electrospray ionization[43]. Peptides are subsequently fragmented after passing through the first mass analyzer, producing MS1 spectra. The resulting fragments are then measured as they pass through the second mass analyzer, producing MS2 spectra, which are used to identify the peptides and their corresponding parent proteins. Briefly, these identifications are commonly made by using database search algorithms, which match the experimental MS2 spectra to in silico MS2 spectra generated from a protein sequence database[44]. The SEQUEST algorithm is a standard approach for such database searching[45].

It is common to use label-free quantification (LFQ) IP-MS/MS for investigating protein-protein interactions[46]. This approach allows for the measurement of interacting protein relative abundance across samples, omitting the use of stable isotope labelling during the sample preparation steps, and producing quantification measurements of either precursor ion intensities or spectral counts. The former refers to the intensity of the parent peptide in the MS1 spectra. These peptide intensities can then be combined to yield the relative protein quantification. Spectral counts refer to the total number of MS2 spectra that contributed to the identification of a given protein. Spectral counts have been shown to be heavily correlated with protein abundance[47]. From choosing the appropriate detection system to the intricacies of the mass spectrometry analysis, these methodologies succumb to many false-positive identifications. This issue becomes more prevalent as the biological matrix used for the IP experiments become more complex[48].

1.2.3 Filtering false-positive identifications from data acquired from IP-MS/MS experiments

When performing IP-MS/MS experiments, both the implemented immunoprecipitation strategy and mass spectrometry portions of the methodology will generate false-positive protein-protein interaction identifications. Bioinformatic strategies are required to address this "noise" in measurement. With IP experiments, contamination by exogenous proteins artifactually introduced to samples is a well-known phenomenon[49]. These include various keratin proteins, bovine serum albumin, Protein A, which includes the IgG binding domain of the TAP tag, or even tryptic enzymes used for digestion during MS sample preparation. However, these contaminating proteins are well-characterized in public repositories such as the contaminant repository for affinity purification-mass spectrometry data (CRAPome)[50], that can be used to query and exclude these known contaminants from IP-MS/MS acquired data. A more confounding source of false-positive identifications stems from antibody non-specific binding, wherein proteins in a sample bind to the antibody of the IP system or the molecular tag expressed by the protein of interest. This issue becomes more persistent as the biological matrix of the IP experiment becomes more complex. While correct matches may be made for these non-specifically bindings, the proteins involved are unlikely to be valid interactors of the protein of interest. These false-positive identifications have been shown to be effectively filtered out by combining the use of experimental negative controls and computational modelling[51–53]. Examples of negative controls for an IP experiment, such as one implementing the FLAG tag, would be to perform the antibody purification in a system not expressing the FLAG molecular tag, or one wherein a protein foreign to the organism is fused with the FLAG tag and then purified. It is assumed that proteins purified in these negative controls are examples of antibody non-specific binding. Therefore, any interactions identified and quantified by mass spectrometry

can then be used to model their background abundance in empirical experiments. The general workflow of IP-MS/MS experiments and example of a common negative control are depicted in Figure 1.

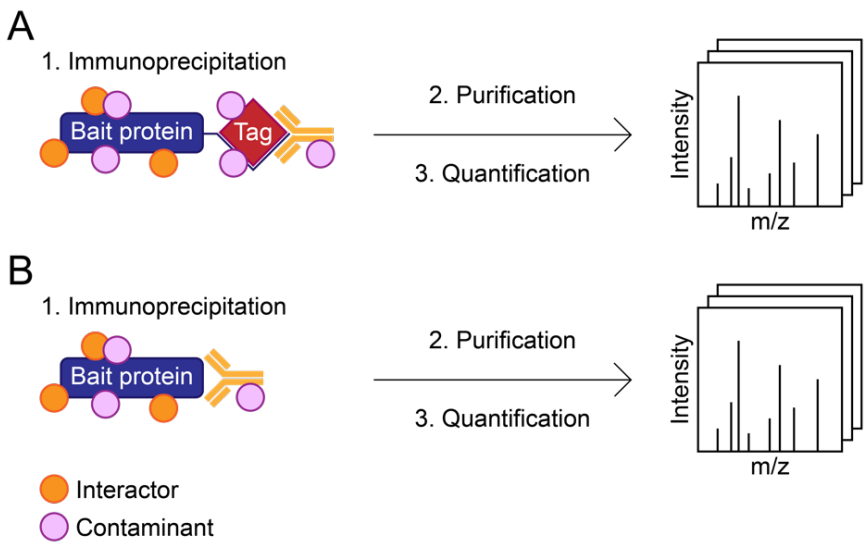

Figure 1. Simplified workflow for IP-MS/MS experiments. IP-MS/MS workflow for empirical experiments, wherein the protein of interest is expressing a molecular tag (A). Complimenting IP-MS/MS workflow for negative control experiments, wherein the protein of interest is not expressing a molecular tag (B).

1.2.4 Computational approaches for filtering false-positive protein-protein interaction identifications from IP-MS/MS acquired data

Generally, computational approaches function by assessing the confidence of a putative interactor for a given protein of interest in the IP dataset, such that it is deemed to be a bona fide interactor if it is present at a significantly higher abundance than in the negative controls. This

logic emphasizes the importance of choosing negative controls that are highly unlikely to result in the notable purification of a bona fide interactor, that they are only composed of the non-specific bindings. Many of these approaches use label-free quantification MS/MS data to assign a confidence score to a successfully purified protein-protein interaction[54], either using precursor signal intensity or spectral counts as quantification measures. Among the prominent algorithms to identify bona fide protein-protein interactions from MS/MS data, the Significance Analysis of INTeractome (SAINT) algorithm is considered a state of the art approach that can be used with either precursor intensity or spectral count quantification measures[52]. SAINT is a mixture model that derives distributions of the probabilities for a set of interactors given that they are true, $P(x_{ij}|True)$, and the probabilities for a set of interactors given that they are false, $P(x_{ij}|False)$, where x_{ij} is an interaction between the prey interactor i and the bait protein j. These two probability distributions are then used to calculate a posterior probability that a putative interactor is true, $P(True|X_{ij})$. SAINT uses these probabilities to estimate a false discovery rate (FDR) for the unique identifications. It is incumbent that the user decides the FDR threshold with which they are comfortable, a decision that may very well be influenced by the chosen experimental design and difficulty thereof and perform further experiments to manually validate the most confidently identified protein-protein interactions. The computation of posterior probabilities for putative bait-prey pairs and the implementation of Bayesian inference to identify true interactions has been used in a number of algorithms[55]. Another such algorithm of note is Decontaminator, which, instead of either precursor intensity or spectral count measures to assess to confidence of putative bait-prey pairs, uses Mascot scores[53]. Briefly, Mascot is a software package in its own right that determines the most likely protein identity for mass spectral data, outputting a score reporting the quality of these matches[56]. Typically, the greater the Mascot

score, the higher the abundance of the proteins, as proteins with high abundance are more easily identified. Similarly to SAINT, Decontaminator computes a true and false probability distribution, though for Mascot scores of the purified prey proteins, and uses these distributions to compute *p*-values for the putative bait-prey pairs. These *p*-values are then used to estimate the corresponding FDRs for the pairs. Another notable algorithm for the confidence assessment of protein-protein interactions is CompPASS[51]. The CompPASS algorithm computes Z-scores from spectral count data, measuring the number of standard deviations σ away a data point a lies from a group mean μ,

$$Z = \frac{a - \mu}{\sigma}$$

and D-scores, a unique score derived by the authors to assess the uniqueness of a putative interactor for a given bait and its reproducibility across biological replicates. The authors note that their algorithm favours interactors that are unique to the IP experiments of a given bait and across its replicates and those with high spectral count values. Therefore, a notable limitation of the CompPASS algorithm is that it may struggle to identify bona fide protein-protein interactions that are present at low abundances. Another genre of computational approaches for evaluating protein-protein interactions is through the use of network topology[57]. These approaches stipulate that a given interaction pair can be supported by other biologically relevant interactions in a given organism. Finally, new computational methods are being designed that integrate existing algorithms into their workflow. For example, research has been done whereby the SAINT algorithm was used to initialize a computational analysis and was followed by a data refinement phase in order to identify direct protein-protein interactions[58].

The biochemical and computational technique for filtering out false-positive protein-protein interactions from IP-MS/MS data have been extensively used in the context of cell

cultures. However, these techniques do not translate for use in human plasma. The diverse levels of abundance of circulating plasma proteins are the root cause for the difficulty in confidently identifying and quantifying protein-protein interactions. An added layer of difficulty lies with the methodology that can be used for the isolation of these protein-protein interactions. Unlike conventional cell lines, systems requiring molecular tags cannot feasibly be implemented with human plasma. Instead, modified immunoprecipitation assays must be used, wherein proteins are directly targeted by an antibody[59]. Because of these challenges, the ability to fully characterize the resulting dynamic protein profile datasets from experiments carried out in human plasma remains a largely unmet need in the research and medical fields. The development of an approach to confidently identify and quantify protein-protein interactions occurring in human plasma would address a critical gap in knowledge and accelerate proteomic bioinformatics.

1.3 Investigating protein-protein interactions in human plasma

While an incredibly challenging field of research, mining information from the human plasma proteome is likely to be highly rewarding[60]. Many metabolic and molecular pathways are black-boxed in human plasma due to a lack of the necessary experimental approaches and computational tools that would be required for confident analyses. For instance, proprotein convertase subtilisin/kexin type 9 serine protease (PCSK9) present at varying levels in human plasma has been linked to different cardiovascular disease phenotypes[59,61]. The apolipoprotein E4 allele (APOE) and its isoforms, associated to Alzheimer's disease, are found to be present at different levels in plasma than in the central nervous system[62]. Furthermore, numerous neurological studies conclude that the investigation regarding how plasma APOE predisposes APOE εE carriers (the strongest known allelic risk factor for Alzheimer's disease[63,64]) to the

disease is required[62,64,65]. Thoroughly investigating PCSK9 and APOE in human plasma requires the characterization of their interactions to determine what role they play in their respective pathologies and how they may be regulated. This statement is the reality for all metabolic and molecular pathways with proteins that are secreted into circulation. However, these proteins make up only a minority of human plasma protein composition (5%), while the majority of human plasma is composed of albumin (55%) and globulins (40%)[66,67]. Because of this stark difference in abundance, human plasma samples generally require pre-processing, typically albumin and/or globulin depletion, to decrease sample complexity[68,69]. The resulting depleted plasma samples less resemble their natural state in the human body. Moreover, it is well understood that albumin plays key roles in plasma, such as acting as a transporter protein for fatty acids[70], steroids[71], thyroxine[72], and interacts with a number of proteins in serum and cells[73]. These potential interactions are lost with depleted samples.

1.3.1 Recent attempts to confidently identify protein-protein interactions in human plasma and blood

While difficult, attempts have been made in the last decade to confidently identify protein-protein interactions in human plasma, serum, and whole blood. However, these attempts are not free of notable limitations. A study in 2012 used LC-MS/MS quantitative proteomics to map the human platelet proteome in human serum samples from healthy donors[74]. The authors proposed that ~85% of the platelet proteome has little variation between healthy donors and, thus, that perturbations in these protein-protein interaction networks could be a starting point for disease research. However, these platelet samples were obtained by the fractionation of fresh blood samples to reduce the contamination of leukocytes, erythrocytes, and plasma to negligible

amounts. Such fractionation disrupts the actual interactions that may be taking place. Therefore, while the human serum platelet network produced in this study may serve as a useful resource with which to compare future studies, it is likely to be missing some biologically relevant information regarding platelet interactions in the native human body environment. A more recent 2019 study attempted to assess the selectivity of antibodies for their target, without the use of molecular tags, in IP-MS/MS experiments[75]. To do so, the authors targeted proteins directly with their respective antibodies in human plasma and systematically evaluated each antibodies' enrichment for its target by computing Z-scores of the LFQ intensities for each purification. The authors considered an antibody to be enriched for its target if the quantified protein obtained a Z-score ≥ 3. In addition, if an antibody successfully purified its target together with a second plasma protein associated with a Z-score ≥ 3, the authors labelled the second protein as a co-target. While not explicitly stated by the authors, the co-target enrichments could be biologically relevant interactors with the target protein, requiring further empirical validation. Intuitively, higher Z-scores may suggest that the enrichment has biological relevance, particularly for the co-target protein purifications. However, Z-scores are limited in that they assume the data being evaluated is normally distributed and they require a larger sample sizes to get a good estimate of the standard deviation to be reflective of any statistical significance. It is, however, quite possible and likely that the abundance of protein-protein interactions in the human plasma proteome is not normally distributed.

In summary, there are a number of limitations to be overcome to confidently identify protein-protein interactions in human plasma. Experimentally, while reducing the complexity of plasma samples by, for example, depleting albumin, could improve mass spectrometry protein identification sensitivity, it does not allow for the complete characterization of the human plasma

proteome. Computationally, certain statistical assumptions made may not be true and current tools for filtering out false-positive protein-protein interactions from IP-MS/MS data are only optimized for *in-vitro* experiments (i.e., in cell lines), which produce much less dynamic and less challenging datasets than does the plasma proteome. In this thesis, I present MAGPIE, a machine learning assessment with logistic regression of protein-protein interactions. MAGPIE is a novel two-phase computational approach for discriminating between putative protein-protein interactions and antibody non-specific binding. The first development phase aimed to confidently and unbiasedly identify a set of experimental negative controls to use for modelling contamination and antibody non-specific binding. These controls will be identified from a set of antibody purifications targeting proteins known with high confidence to not to be present in human plasma, wherein the quantitative measurement used for these purifications was spectral counts. It was assumed that interactions identified with these foreign antibodies are non-specific, lending themselves to modelling the abundance of non-specific binding that occurs in human plasma. The second development phase implemented a supervised machine learning algorithm to predict whether a putative protein-protein interaction detected in an antibody purification targeting a protein expected to be present in human plasma is biologically relevant or the result of non-specific binding.

1.4 Unsupervised machine learning

Unsupervised machine learning is typically used in an exploratory manner when beginning a new machine learning-based project. This is because the dataset inputted into these algorithms is usually a large collection of unlabelled training examples. Each unlabelled training example, n, is described by its features, often in the form of a vector with p dimensions.

Therefore, given an unlabelled dataset, the problem at hand for the algorithm is to discern patterns, or relationships in the data based on the features provided, and create a representative model. Two of the most common tasks for such machine learning are clustering and dimensionality reduction algorithms.

1.4.1 Agglomerative hierarchical clustering

As its name suggests, agglomerative hierarchical clustering is an iterative bottom-up approach, wherein individual input data are evaluated and clustered together until the algorithm is left with one large group – from many to one[76]. This approach for clustering data constitutes a family of unsupervised machine learning algorithms, each recognized by their precise method used to group the input data together: the distance measure and linkage criterion. The distance measure is used to compute how close two groups of data are to each other. Often, Euclidean distance is implemented, wherein the distance between elements (*a* and *b*) associated with *n* features (*a, b*) is computed as follows:

$$||a - b||_2 = \sqrt{\sum_{i=1}^{N}(a_1 - b_1)^2}$$

Another common metric used for computing distance is the Pearson's correlation. While the Pearson's correlation is not directly a distance measure, this coefficient still reports on the similarity between two groups. The Pearson's correlation, *r*, is computed as follows:

$$r = \frac{\sum_{i=1}^{N}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{N}(a_i - \bar{a})^2 \sum_{i=1}^{N}(b_i - \bar{b})^2}}$$

The linkage criterion is a function for determining the pairwise distance from a set of observations to be used when choosing which groups to cluster in each iteration. For example, complete-linkage selects points *a* and *b* that will maximize the distance, *d(a, b)*, between clusters *A* and *B*[77].

$$max\{d(a,b) : a \in A, b \in B\}$$

Conversely, single-linkage selects points *a* and *b* that will minimize the distance, *d(a, b)*, between clusters *A* and *B*.

$$min\{d(a,b) : a \in A, b \in B\}$$

Finally, another common linkage that a user may choose is average-linkage, which computes the distance between clusters *A* and *B*, *D(A, B)*, as the average distance between all points within both clusters.

$$D(A,B) = \frac{1}{|A| \bullet |B|} \sum_{i=1}^{|A|} \sum_{j=1}^{|B|} d(a_i, b_j)$$

The choice of linkage greatly depends on the user's preference and the dataset itself being clustered. Returning to the idea of unsupervised machine learning being used in exploratory fashions, it is common that a user attempts multiple combinations of distance measures and linkage criteria until an optimal combination is established based on the user's goals, and defined clusters are formed. This trial-and-error approach is merely a first layer of investigation into the relationships present within the dataset. A dendrogram is then used to report the successful results of hierarchical clustering, wherein the "leaves" represent the data points and the "branches" represent the relationship and distance between the data points (Figure 2). These dendrograms are typically interpreted such that data points clustered together (or "leaves from

the same group of branches") are more similar to each other than they are to other clusters elsewhere in the dendrogram.

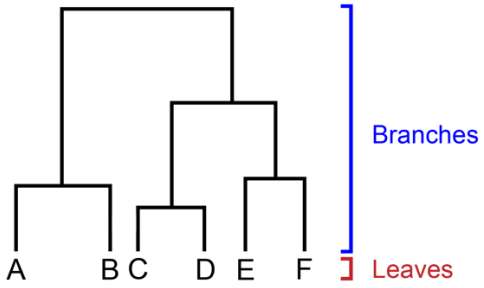

Figure 2. Toy example of dendrogram representing hierarchical clustering results.

1.4.2 Principal component analysis

A principal component analysis (PCA) serves two purposes. Primarily, it is known as a dimensionality reduction algorithm – an essential component of machine learning[78,79]. Often, the datasets used for building machine learning models have high dimensionality, meaning there are many features describing each input example, or that the number of features is far greater than the number of input examples, $p >> n$. This is particularly true with biological datasets produced from high-throughput experiments. For instance, in an IP-MS experiment, one protein purification will lead to the quantification of numerous protein interactors, which represent such features. Such high-dimensional datasets are likely to contain non-informative and redundant features and using them to build a model would result in poor performance and a model that is unlikely to be generalizable to new datasets. PCA is one such dimensionality reduction algorithm that then facilitates clustering of the reduced data, allowing the user to visualize the distinct

groups within their dataset. It helps to reveal relationships in the dataset, specifically by reporting on the variance within the dataset.

The PCA algorithm begins by standardizing the input variables, often by Z-score standardization, without which differences in the scale of measures between features would produce skewed results downstream. Once all variables are within the same scale, the algorithm creates a symmetrical *p* x *p* covariance matrix, where *p* is the number of features, which summarizes the correlations between all pairs of variables in the dataset. For example, a 3-dimensional dataset with three input examples, *a*, *b*, and *c*, would have the following covariance matrix,

$$\begin{bmatrix} Cov(a,a) & Cov(a,b) & Cov(a,c) \\ Cov(b,a) & Cov(b,b) & Cov(b,c) \\ Cov(c,a) & Cov(c,b) & Cov(c,c) \end{bmatrix}$$

From this matrix, eigenvector and eigenvalue pairs are computed, wherein the number of pairs computed equals the number of dimensions in the dataset. The eigenvectors are the new variables – the principal components. Using a simpler example, a two-dimensional dataset may produce the following two pairs of eigenvectors and eigenvalues.

$$v_1 = \begin{bmatrix} 0.827412 \\ 0.973845 \end{bmatrix} \quad \lambda_1 = 1.3578$$
$$v_2 = \begin{bmatrix} -0.432796 \\ 0.742435 \end{bmatrix} \quad \lambda_2 = 0.7643$$

When the components are ordered based on the descending value of their corresponding eigenvalues, the result is the principal components become ordered based on their significance. Using the above toy example, $\lambda_1 > \lambda_2$, thus v_1 would correspond to the first principal component and v_2 to the second component. They are ordered such that the first principal component explains the most amount of variance within the dataset, the second principal

component explains the most amount of variance of what remains, and so on. Above, v_1 represents 64% of the dataset's variance and v_2 represents 36%. Finally, the user will select how many principal components to plot, in order to visualize the variance within the dataset. Often, the first two principal components are plotted and, occasionally, the first three principal components. Two-dimensional and three-dimensional PCA scatter plots are feasible and generally interpretable (Figure 3). If the input dataset contains distinctly different classes of samples, clusters for each class will typically be observed in these plots. In the case where the user would like information regarding the significance of the fourth principal component and beyond, a discrete plot of the percentage of explained variance against principal component number can also be created.

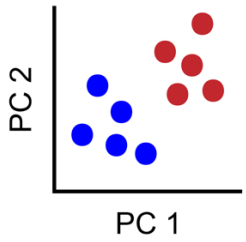

Figure 3. Toy example of a two-dimensional principal component analysis with two groups of data points that are colour-coded.

1.5 Supervised machine learning

Supervised machine learning algorithms are a family of algorithms that focus on the task of learning from data how to classify or perform regression on a new set of data. In contrast to the unsupervised algorithms, supervised machine learning requires training with labelled

examples, such that the model can, for example, predict the discrete label for an unseen testing example. This form of supervised machine learning is referred to as classification. These models, called classifiers, can be trained to perform binary classification (i.e., between two classes) or multi-class classification (i.e., between more than two classes). Because of the plethora of models available, supervised machine learning classifiers have rapidly emerged in bioscience and biomedical research. Conversely, regression models are used to estimate the relationships between dependent and independent variables. These models are particularly adept at determining which independent variable (i.e., predictors or features) impact the dependent variable, making them favourable models for prediction problems.

1.5.1 Logistic regression classifiers

Among the most popular supervised machine learning algorithms is the logistic regression model. While logistic regression models are often used for classification problems, they are not classifiers themselves. They are statistical models that, when trained, output the probability that a given input sample belongs to a given class. However, they are frequently used as classifiers by selecting a threshold to discriminate the outputted probabilities between the classes of the dataset (Figure 4).

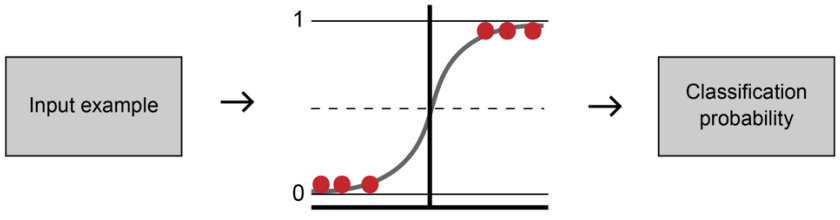

Figure 4. Toy example of a logistic regression model. Input examples are classified by the logistic regression model and the probability of that input belonging to a given class is outputted.

The probability threshold for determining which class the input belongs to is represented by the dotted line. Red data points represent training examples.

At their core, logistic regression models function by optimizing a linear regression model and transforming its output using a sigmoid function. For example, a model with three features would begin with the generalized formula,

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

Here, x_{1-3} are the features, θ_0 is the y-intercept, θ_{1-3} are the feature coefficients, and z is the linear regression output, which is mapped to a probability through the sigmoid function below,

$$probability = \frac{1}{1 + e^{-z}}$$

The feature coefficients are determined through the use of learning procedures, such that these coefficients can be optimized to build a generalizable model (*see Introduction 1.5.2 – 1.5.4*).

For example, a 2020 study constructed a logistic regression classification model for discriminating early-stage gastric cancer (EGC) patients from healthy controls based on the quantification of eleven plasma proteins found to be differentially expressed in EGC patients[80]. Differentially expressed proteins were identified by, first, depleting patient plasma samples of highly abundant proteins, then comparing their abundances in ECG patients and healthy controls by tandem mass tagging-LC-MS/MS. Their final model was trained to output the probability that a given test sample belonged to an EGC patient or healthy control, wherein the features were the quantification measurements of the eleven differentially expressed proteins. An earlier 2009 study implemented a logistic regression model to predict whether a given putative protein-

protein interaction was true while building a network of interactions for the RNA polymerase II and other human transcription machinery proteins[57].

The simplicity and transparency of logistic regression models are largely what makes them an alluring option for machine learning problems. Therein, users are able to retrieve the feature coefficients, which facilitates the interpretation of how much weight each feature has on the trained model.

1.5.2 Cross-entropy loss function

When optimizing a machine learning model, loss functions (sometimes called cost functions) are used to compute the deviations of predicted values from actual labels[81]. The computed loss therefore indicates to the optimization function being used whether the model's performance is improving, or if performance has plateaued. Intuitively, there is no one loss function best for all machine learning problems. Some are robust to outliers, while others heavily penalize outliers in predictions. One such example of the latter, that is typically used with logistic regression models, is the cross-entropy (CE) loss function. This function is defined by the following formula,

$$CE = -\frac{1}{N}\sum_{i=1}^{N}(y_i log(p_i) + (1 - y_i) log(1 - p_i))$$

Here, y_i is the binary indicator (1 or 0), p_i is the predicted probability of class 1, $(1 - p_i)$ is the predicted probability of class 0, and N is the total number of examples. In cases where the given predicted probability belongs to a sample in class 1, the second term of the formula is reduced to zero. Conversely, when a predicted probability belongs to a sample in class 0, the first term of the formula is reduced to zero. This logic is why the cross-entropy loss function is considered to

harshly penalize incorrect predictions. Moreover, the cross-entropy loss function for logistic regression is a convex problem. This means that, during the optimization process, a global minimum can be computed, which is the most desired outcome when training a machine learning model.

1.5.3 ℓ_2-norm for model regularization

Generalizability is a key aspect of a machine learning model, especially if that model is intended to be used as a tool. Therein, if a model is overfit for its training dataset, then it will perform well on the training data, but perform poorly when applied to new datasets. A common strategy for reducing the risk of overfitting is to apply regularization when computing the model's error, by adding a penalty term to the loss function. Often, with logistic regression, the ℓ_2-norm is applied as a penalty, as a means to constrain the feature coefficients of the model to reduce the model's complexity.

$$\ell_2 = \alpha \sum_{i=1}^{N} \theta_i^2$$

Here, α is a hyperparameter, θ_i^2 is the squared magnitude of the feature coefficients, and N is the total number of examples. It is common to assess multiple values of α to find its optimal value, as a value too small can result in model overfitting (i.e., the penalty is too weak) and a value too large can result in model underfitting (i.e., the penalty is too strong).

1.5.4 Newton conjugate gradient optimization

The feature coefficients for a model are determined through the implementation of an optimization algorithm. These iterative algorithms function by making small changes to the

coefficient values until a set is found that minimizes the model's loss function. A common optimization algorithm is the Newton method, but it requires the computation of a Hessian matrix (a square matrix of second-order partial derivatives), making it a computationally expensive algorithm with a long run-time. Therefore, the Newton method is not ideal for large and/or high-dimensional datasets. The Newton conjugate gradient algorithm is a variation of the original algorithm that only requires the computation of Hessian vector products, thus reducing run-time and computational resources. It is by far the most common strategy to optimize the coefficients of a logistic regression classifier.

1.5.5 Cross-validation

Supervised learning models must be evaluated for their ability to be generalized to different datasets, other than the one it was trained on. A model that can be generalized will perform roughly as well at making predictions both the data with which it has been trained and with new datasets it has not previously trained on. Conversely, a model that is not generalizable would perform much better at predicting on its training data than on other datasets. This undesired property is called overfitting. To assess if a model is overfitting, cross-validation strategies are often implemented[82]. These strategies require stratifying the input data into different subsets named training and testing sets. A common cross-validation strategy is called *k*-fold cross-validation, which splits the data into *k* subsets. Each subset is then iteratively left out (i.e. not used in training the model) and, after training with the remaining subsets, is used to test the model. Performances across the *k* iterations are used to inform the user whether their model generalizes well. Poor performance during cross-validation is an indication of being overfit (i.e., bias) to the training data. Often, 5- and 10-fold cross-validation are used and, if not

computationally prohibitive, *n*-fold cross-validation[83]. This strategy is also referred to as leave-one-out cross-validation, wherein *n* is equal to the number of data instances within the dataset. This strategy has the advantage of considering every single data point as a test set and therefore in the performance evaluation of the model.

1.6 Hypothesis and objectives

I hypothesize that a machine learning strategy can be used to assess the confidence of putative protein-protein interactions in human plasma, detected by mass spectrometry. This hypothesis is addressed by the two following objectives:

Objective 1: Build an unsupervised machine learning strategy to identify a set of experimental negative controls from antibodies targeting proteins known not to be present in human plasma to assess non-specific protein binding.

Objective 2: Design a supervised machine learning classifier to discriminate between bona fide protein-protein interactions and antibody non-specific binding in human plasma, using data acquired from immunoprecipitation experiments and liquid chromatography coupled to tandem mass spectrometry, including the negative controls identified in Objective 1.

2 Methods

The methods used and algorithm developed in this thesis are summarized in Figure 5. Mass spectral count datasets were produced from IP-LC-MS/MS experiments and used to detect protein-protein interactions in human plasma samples (Figure 5A). The empirical dataset was composed of a collection of IP experiments using antibodies to directly target proteins known to be present in human plasma, while the negative control dataset was composed of a collection of IP experiments using antibodies targeting proteins expected not to be present in human plasma. Hierarchical clustering and a principal component analysis were used to confidently identify a set of negative control experiments. This was followed by the construction of MAGPIE (Machine learning assessment with logistic regression of protein-protein interactions), a supervised machine learning-based assessing the confidence of putative protein-protein interactions in human plasma (Figure 5B). MAGPIE was then benchmarked against a leading algorithm for assessing protein-protein interactions, SAINT. External repositories of protein-protein interactions, functional annotations, and gene co-expression were then used to further investigate MAGPIE's high-confidence identifications.

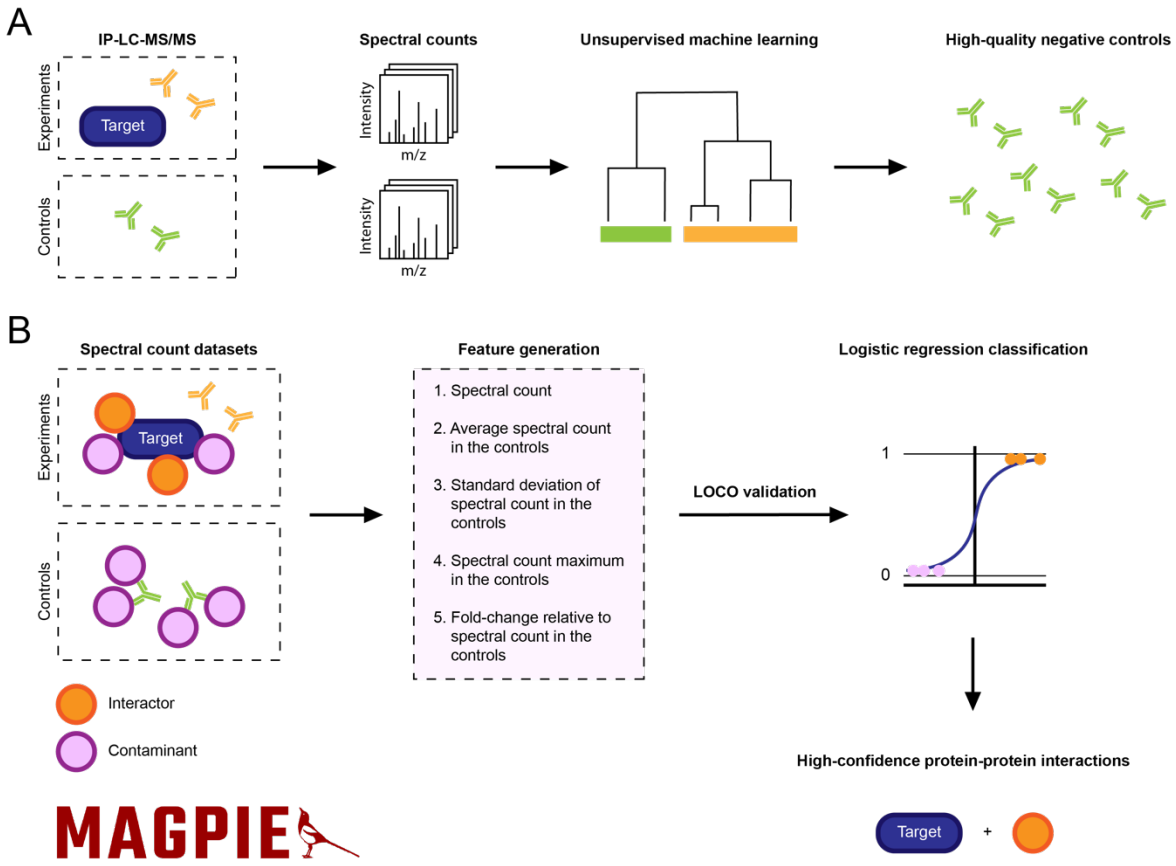

Figure 5. Workflow schematic. The mass spectral count datasets of empirical and negative control IP-LC-MS/MS experiments performed in human plasma samples were analyzed by unsupervised machine learning algorithms (hierarchical clustering and principal component analysis) to identify a set of high-quality negative controls (A). Features for supervised machine learning were engineered and assigned to all training and testing examples. A logistic regression classifier model was constructed and trained to output the probability that a given putative protein-protein interaction was true, from which an FDR is derived (B).

2.1 Protein affinity capture coupled to quantitative mass spectrometry assay

The following laboratory methodology (*Methods 2.1.1 – 2.1.4*) was designed and executed by Dr. Benoit Coulombe's research group at the *Institut de recherches cliniques de Montréal*.

2.1.1 Chemicals and reagents

Affinity pipettes fitted with porous microcolumns coupled to streptavidin (Thermo Scientific, 991STR11) were coupled to biotinylated antibodies (Antibody IDs: Anti-FLAG-IgG, Anti-HA-IgG, Anti-LC3B-IgG, Anti-METTL23-IgG, Anti-RPAP2-IgG, Anti-SRB7-IgG Ab 1, Anti-SRB7-IgG Ab 2, Anti-CNDP1-IgG Ab 1, Anti-KLK6-IgG, Anti-SNCA-IgG, Anti-PCSK9-IgG, Anti-CNDP1-IgG Ab 2) using a Versette automatic liquid handler (Thermo Fisher Scientific) as previously described[84,85]. Iodoacetamide, DTT, and glucagon were from Sigma, sequencing grade modified trypsin from Promega, and HBS-EP buffer from GE Healthcare. HPLC grade water, trifluoroacetic acid, and acetonitrile were purchased from Fisher.

2.1.2 Protein affinity capture

Plasma samples (250 µl) were diluted with 175 µl of HBS-EP buffer, dispensed in 96-well robotic PCR plates (Abgene). Protein affinity capture was automated using a Versette automatic liquid handler (Thermo Fisher Scientific). The affinity pipettes coupled to antibodies were mounted to the Versette's head and were first equilibrated by 15 aspiration and dispensing cycles of 100 µl of HBS-EP buffer. For affinity capture, 250 aspiration and dispensing cycles of 250 µl of diluted samples or standards were performed. This was followed by washes consisting of 10 aspiration and dispensing cycles of 150 µl of HBS-EP and 2 x 10 aspiration and dispensing

cycles of 150 µl of water from their respective 96-well plates. Finally, enriched proteins were eluted by 250 aspiration and dispensing cycles of 30 µl of an elution buffer consisting of 33% acetonitrile and 0.4% trifluoroacetic acid. Eluates were evaporated using a speed vacuum centrifuge (Eppendorf) and stored at -20ºC until digestion.

2.1.3 Sample preparation for LC-MS/MS

Samples were reconstituted in 45 µl of 4M urea, 100 mM ammonium bicarbonate, 2.5% N-propanol and 10 mM dithiothreitol on a Mixmate (Eppendorf) at 1200 rpm for 10 min. Reduction of disulfide bonds was performed at 37ºC for 30 min on a ThermoMixer (Thermo Fisher Scientific) set at 350 rpm. After cooling for 5 min at room temperature, 10 µl of 250 mM iodoacetamide was added. Alkylation was allowed to proceed for 30 min at room temperature in the dark. 69 µl of digestion buffer (100 mM ammonium bicarbonate, 2 mM $CaCl_2$ pH 8.0) and 1 µg of trypsin were added to each well. Trypsin activity was tested using Nα-benzoyl-L-arginine ethyl ester. The plate was sealed, mixed on a Mixmate at 350 rpm for 5 min and spun down. The digestion reaction was allowed to proceed for 20h on a ThermoMixer set at 350 rpm and 37ºC, after which the plate was cooled on ice for 5 min and spun down. The reaction was then quenched by the addition of 3 µl of 100% formic acid and 2.4 µg of glucagon as a peptide carrier.

2.1.4 LC-MS/MS conditions

Peptides samples were loaded into a 75 µm i.d. × 150 mm Self-Pack C18 column installed in the Easy-nLC 1200 system (Proxeon Biosystems). The buffers used for chromatography were 0.2% formic acid (buffer A) and 90% acetonitrile/0.2% formic acid (buffer

B). Peptides were eluted with a two-slope gradient at a flow rate of 250 nL/min. Solvent B first increased from 1 to 40% in 100 min and then from 40 to 85% B in 10 min. The HPLC system was coupled to an Orbitrap Fusion mass spectrometer (Thermo Scientific) through a Nanospray Flex Ion Source. Nanospray and S-lens voltages were set to 1.3-1.8 kV and 60 V, respectively. Capillary temperature was set to 250°C. Full scan MS survey spectra (*m/z* 360-1560) in profile mode were acquired in the Orbitrap with a resolution of 120,000 with a target value at 3e5. The 25 most intense ions were fragmented in the HCD collision cell and analyzed in the linear ion trap with a target value at 2e4 and normalized collision energy at 28. Target ions selected for fragmentation were dynamically excluded for 20 sec. The peak list files were generated with Proteome Discoverer (version 2.1, Thermo Fisher Scientific) using the following parameters: minimum mass set to 500 Da, maximum mass set to 6000 Da, no grouping of MS/MS spectra, precursor charge set to auto, and minimum number of fragment ions set to 5. Protein database searching was performed with Mascot 2.6 (Matrix Science)[86]. The mass tolerances for precursor and fragment ions were set to 10 ppm and 0.6 Da, respectively. Tryptic peptides allowing for up to 2 missed cleavages were searched by the algorithm. Cysteine carbamidomethylation was specified as a fixed modification and methionine oxidation as variable modifications. Data interpretation was performed using Scaffold (version 4.8)[87], Mascot and Qual browser (Xcalibur, Thermo Fischer Scientific) and protein identifications were performed by applying 1% FDR filtering.

2.2.1 IP-MS/MS datasets

Acquired by our collaborators, our mass spectrometry datasets consisted of the spectral count label-free mass-spectrometry-based quantification of immunoprecipitation experiments

performed on human plasma samples. Such spectral counts represent a proxy of the abundance of the proteins. The immunoprecipitation experiments were performed using twelve antibodies, targeting ten different proteins, and co-purifying a total of 226 circulating plasma proteins. Of these, five immunoprecipitation experiments were performed using antibodies targeting proteins known to be present in human plasma and, thus, containing putative protein-protein interactions. Conversely, the remaining seven immunoprecipitation experiments were performed using antibodies targeting proteins expected to not be present in human plasma. These tentative experimental negative controls were initially assessed to determine if they confidently produced useable examples of antibody non-specific binding. The resulting data was then used for the construction of a machine learning classifier. Table 1 summarizes the datasets.

Table 1. Composition of mass spectrometry datasets.

Dataset	Number of antibodies	Dataset description
Empirical experiments	5	This dataset contains the spectral count quantification of proteins purified by antibodies targeting proteins known to be present in human plasma.
Experimental negative controls	7	This dataset contains the spectral counts of proteins purified by antibodies targeting protein expected to not be present in human plasma.

2.2.2 Identifying a set of experimental negative controls using an unsupervised machine learning approach

Performing immunoprecipitation experiments using an antibodies that directly target proteins for both experimental and negative control datasets is unprecedented. Our mass spectrometry datasets were therefore analyzed to evaluate whether a subset of experimental

negative controls reproducibly captured a similar population of antibody non-specific binding. The rationale being that non-specific bindings detected are more likely to occur with abundant proteins and that such bindings should be reproduced for different antibodies. On the other hand, while proteins could be uniquely purified by a single antibody and remain non-specific binders, such an antibody would not be useful to model contamination events that can occur in empirical experiments using different antibodies. Hence, controls showing a fairly well-reproduced set of purified proteins could, therefore, be used to build a model of non-specific binding.

2.2.3 Data refinement for unsupervised machine learning

Prior to the implementation of the unsupervised machine learning algorithms, the mass spectrometry spectral count data was pre-processed to ensure confident analyses were conducted. As it is standard procedure when preparing data for unsupervised machine learning analysis, the spectral count data was normalized. Specifically, the spectral count of an interacting protein, *p*, was normalized against the total spectral count for all plasma proteins purified in a given IP experiment (bait), *b*.

$$x_{b,p\ normalized} = \frac{x_{b,p}}{x_b}$$

where, $x_{b,p}$ = spectral count of the interacting protein, *p*, when purified in IP experiment *b*
x_b = total spectral count of all purified proteins in IP experiment *b*

2.2.4 Evaluating experimental negative controls using an unsupervised machine learning approach

This methodology consisted of two unsupervised machine learning algorithms, both implemented in an exploratory manner to assess the behaviour of the targeting antibodies in human plasma. First, a complete-linkage hierarchical clustering analysis, using Euclidean

distance, was used to investigate the similarity between plasma proteins purified by the different antibodies. This was an initial assessment of how reproducibly antibody non-specific binding was captured across the negative controls and was further supplemented by applying statistical bootstrapping of the clustering. Therein, the robustness of the resulting clusters of antibody purifications was assessed to provide more confidence. This bootstrapped hierarchical clustering analysis was implemented in R (version 3.6.1), using the *pvclust* package (version 2.2.0)[88]. To confirm the observations of the previous analysis, a two-dimensional principal component analysis (PCA) was used to visualize the variance in antibody behaviour, which in turn linearly reduced the dimensionality of the normalized spectral count datasets. The PCA was implemented in Python (version 3.7), using the *Scikit-learn* package (version 0.24.0)[89]. Clustering results from both approaches were combined to determine the set of IP-MS/MS experiments being used as negative controls (*see Results 3.1*).

2.3 Classifying bona fide protein-protein interactions from antibody non-specific binding using MAGPIE

I developed MAGPIE, an algorithm implemented in Python, which is trained to output the probability that a given putative protein-protein interaction is a bona fide interaction. Upon identifying a subset of negative controls, MAGPIE was implemented as a sophisticated supervised machine learning approach for assessing confidence in putative protein-protein interactions using machine learning.

2.3.1 Data refinement by addition of spectral pseudocount values to experimental negative controls

As was made evident while identifying a subset of negative controls, antibody non-specific bindings cannot be modelled without a reasonable amount of negative control data. Our dataset contained numerous instances of a plasma protein being quantified in an empirical experiment but failing to be detected in one or more negative controls. Such proteins may be of great interest, particularly if it had a reasonably high spectral count in an empirical experiment. Nevertheless, it is also possible that such a protein was present in some of the negative controls but failed to be detected by mass spectrometry. To provide a conservative assessment of protein-protein interaction reliability, we, therefore, added spectral pseudocount values in negative controls to plasma proteins that failed to be detected in a negative control if it was quantified in at least one empirical experiment. To prevent over-representing these values in the controls and skewing our results, pseudocounts were assigned by randomly sampling a spectral count value from the bottom 10% of non-zero spectral count values belonging to that negative control. This addition enabled MAGPIE's ability to calculate various scores for all putative protein-protein interactions.

2.3.2 Establishing criteria for likely high-confidence protein-protein interactions

Prior to constructing a machine learning classifier, the biological relevance of protein-protein interactions detected in antibody purifications targeting proteins known to be present in human plasma was assessed. The purpose of this assessment was two-fold. It established a benchmark comparison with our machine learning strategy but was also crucial for identifying likely high-confidence protein-protein interactions that can be used to train the classifier.

MAGPIE began by establishing the background abundance of a given plasma protein, *p*, in the negative controls, using the average, μ_p, and standard deviation, σ_p, of its spectral count. This represented an estimation of the abundance of *p* in antibody non-specific binding. Thereafter, to determine if the same plasma protein detected by an empirical experiment was a putative bona fide interaction, MAGPIE calculated both a Z-score and fold-change value based on the abundance of that plasma protein in the negative controls and empirical experiments. A Z-score is a measure of the number of standard deviations a given piece of data lies from a respective group mean. This approach resembles the method used by Fredolini et al.[75] in their own investigation of antibody specificity in plasma. The Z-score, $z_{b,p}$, was calculated as follows,

$$z_{b,p} = \frac{x_{b,p} - \mu_p}{\sigma_p}$$

where, $x_{b,p}$ = normalized spectral count of the interacting protein, *p*, purified in IP *b*
μ_p = mean normalized spectral count of *p* in all controls
σ_p = standard deviation of the normalized spectral count of *p* in all controls

Fold-change measures the relative change of a plasma protein's spectral count between an empirical experiment and the negative controls. In this case, fold-change, $fc_{b,p}$, of a putative interacting protein, *p*, was calculated as follows,

$$fc_{b,p} = \frac{x_{b,p}}{\mu_p}$$

where, $x_{b,p}$ = normalized spectral count of the interacting protein, *p*, purified in IP *b*
μ_p = mean normalized spectral count of *p* in all controls

Notably, when computing Z-scores and fold-change values for detections in the negative controls, the spectral count average and standard deviation were recomputed, such that the normalized spectral count value belonging to a plasma protein purified by a given negative

control antibody was omitted. This negated the influence of that purification, which would otherwise skew MAGPIE's calculations to generate lower Z-scores and fold-changes.

Both these assessments were treated as confidence scores, wherein a high Z-score or fold-change value signified that the abundance of the interacting protein, *p*, in the evaluated experiment was higher than what was seen in the controls. Therefore, this was likely to be a bona fide interaction with the protein targeted by the antibody used in the immunoprecipitation experiment given that the protein targeted is present in plasma. Conversely, an average or low Z-score or fold-change value suggested the interacting protein was detected at similar or lower levels in the experiments than in the controls, showing that the plasma protein was likely binding non-specifically.

Once confidence scores were calculated for all putative protein-protein interactions and protein identified in the negative controls, MAGPIE treated these scores as thresholds, at which false discovery rates (FDRs) were estimated. Notably, false discovery rate estimation was done separately for Z-scores and fold-change values. Such estimation was performed by computing the ratio of the number of plasma proteins in the negative controls that obtained a confidence score greater than or equal to the threshold (i.e., false-positive interactions) and the ratio of the plasma proteins in the empirical experiments that obtained a confidence score greater than or equal to the threshold (i.e., true interactions). These counts were normalized against the total number of plasma proteins purified in the negative controls and empirical experiments, respectively, and used to compute the FDR at each threshold, *t*, as follows,

$$fdr(t) = \frac{\frac{\sum_{b \in B_C} \sum_{p \in P_C} 1_{s^C_{b,p} \geq t}}{|B_C \times P_C|}}{\frac{\sum_{b \in B_E} \sum_{p \in P_E} 1_{s^E_{b,p} \geq t}}{|B_E \times P_E|}}$$

where, b = IP experiment from a collection of empirical or negative control experiments
B = set of empirical or negative control experiments
p = successfully purified interaction protein
P = set of successfully purified proteins
E = empirical experiment
C = negative control experiment
s = confidence score
t = confidence score threshold (Z-score or fold-change)
1_a = 1 if a is true and 0 otherwise

Finally, a monotonic transformation was applied to the estimated FDRs when predicting the number of putative protein-protein interactions at a given FDR, calculated as follows,

$$Count(FDR_t) = min(Count(FDR_t), Count(FDR_{t+inc}))$$

where t is a confidence score threshold and $t+inc$ is the following confidence score threshold. This procedure eliminates the variation that can be observed at very stringent values of t.

The results of this analysis provided insights to choosing a criterion for defining likely high-confidence protein-protein interactions and will be used for benchmark comparison against the machine learning classifier.

2.3.3 Training and testing sets assembly

To perform a more sophisticated analysis for identifying putative protein-protein interactions in the empirical experiments using the negative controls, a supervised machine learning classifier was constructed for the binary classification of the putative protein-protein interactions. To train this classifier, a criterion defined likely high-confidence protein-protein interactions, and thus our positive training examples, as putative interactions whose Z-score was greater than or equal to 3. Negative examples were represented by protein-protein interactions detected in the negative controls. However, because there were many more negative examples than positive, a subset of negative examples was randomly sampled to create a 1:1 class-

balanced training set. The testing dataset was composed of all putative protein-protein interactions in the empirical experiments and all proteins purified in the negative controls. Normalized spectral count data were then mined to generate classifying features, summarized in Table 2.

Table 2. Classifying features for training and testing the machine learning model.

Feature	Feature description
Spectral count	Spectral count normalized against the total spectral count for all detections of a given experiment or control.
Average spectral count in the controls	Normalized average spectral count across all controls for a given detected plasma protein.
Standard deviation of spectral count in the controls	Normalized sample standard deviation across all controls for a given detected plasma protein.
Spectral count maximum in the controls	Normalized spectral count maximum across all controls for a given detected plasma protein.
Spectral count fold-change relative to the controls	Normalized fold-change relative to average spectral count across all controls for a given detected plasma protein.

2.3.4 Constructing a supervised machine learning classifier to detect bona fide protein-protein interactions

MAGPIE implements a logistic regression model for classification, which was trained to output the probability that a given putative protein-protein interaction constitutes a bona fide interaction. The logistic regression model was implemented using the Scikit-learn package (version 0.24.0) and its hyperparameters were manually optimized to minimize our model's false discovery rate as much as possible, described in *Methods 2.3.2*. The "multi_class" hyperparameter was set to "multinomial", such that the cross-entropy loss function was used.

The “penalty” hyperparameter was set to “l2” to apply an L2 regularization penalty. Finally, the “solver” hyperparameter was set to “newton-cg”.

2.3.5 Evaluating MAGPIE’s performance for detecting bona fide protein-protein interactions

Due to the lack of existing ground truth knowledge about the protein interactions of the proteins purified in our study, traditional metrics for evaluating classifier performance cannot be calculated, such as accuracy, or sensitivity. Instead, the outputted probabilities were treated as thresholds, similarly to the confidence score thresholds, at which FDRs were estimated. However, to derive an FDR without skewing results, MAGPIE implements a leave-one-control-out (LOCO) strategy. This strategy functions by executing MAGPIE *k* times, wherein *k* is the number of negative control immunoprecipitation experiments, and re-engineering the classification features for each iteration by omitting the spectral count data of a given negative control. The spectral count average, standard deviation, maximum, fold-change, and Z-scores were re-computed for each LOCO iteration. The purified plasma protein detections belonging to the omitted negative control were used as the testing dataset for predicting the number of false-positive identifications. FDRs were computed for each LOCO iteration as follows,

$$fdr(t) = \frac{\frac{\sum_{b \in B_C} \sum_{p \in P_C} 1_{prob^C_{b,p} \geq t}}{|B_C \times P_C|}}{\frac{\sum_{b \in B_E} \sum_{p \in P_E} 1_{prob^E_{b,p} \geq t}}{|B_E \times P_E|}}$$

where, b = IP experiment from a collection of empirical or negative control experiments
B = set of empirical or negative control experiments
p = successfully purified interaction protein
P = set of successfully purified proteins
E = empirical experiment
C = negative control experiment
$prob$ = logistic regression probability
t = logistic regression probability threshold
1_a = 1 if a is true and 0 otherwise

The same monotonic transformation used in *Methods 2.3.2* was implemented. Once the results of the LOCO iterations had been acquired, MAGPIE derived an overall FDR at each probability threshold by taking the ratio of the normalized summed count of predictions greater than or equal to a given probability threshold across all LOCO iterations.

Finally, because there are two instances of random sampling in each of its runs (spectral pseudocounts and training set assembly), the robustness of MAGPIE's performance was evaluated to ascertain more confidence in its results. MAGPIE was run 1000 times. The results of these randomized runs were evaluated for notable deviations in the results.

2.3.6 Benchmarking MAGPIE against SAINT

MAGPIE was benchmarked against SAINT, a leading algorithm for assessing confidence in putative protein-protein interactions. To run SAINT, Swiss-Prot and TrEMBL (both accessed May 20, 2020) were locally downloaded to retrieve protein sequence lengths, which are necessary in SAINT's assessment of the reliability of an interaction. To execute SAINT, the *saint-spc-ctrl* program was executed with recommended parameters *nburn* = 2000 (number of burn-in period for the Gibbs sampling procedure training the algorithm), *niter* = 10,000 (number of iterations in Gibbs sampling), *lowMode* = 1 (minimizing the impact on the confidence assessment of high spectral count interactions), *minFold* = 1 (forcing separation of positive and negative distributions), *normalize* = 0 (without spectral count normalization). To perform a fair comparison, SAINT was run while implementing the LOCO strategy described in *Methods 2.3.5*. In other words, SAINT was run *k* times, wherein *k* is the number of negative control immunoprecipitation experiments. The spectral count data of a negative control was omitted and used for predicting false-positive identifications. As SAINT also outputs the probability that a

given putative protein-protein interaction is a bona fide interaction, the same FDR ratio computation and monotonic transformation described in *Methods 2.3.2* were applied, though using the SAINT probability as the threshold, *t*, instead of the logistic regression probability. Therein, the results of MAGPIE and SAINT could be appropriately compared.

2.4 Supplementing MAGPIE's results with external data from public repositories

To further ascertain confidence, MAGPIE's classification results were supplemented by external data in publicly available databases. While these were not a means of validating MAGPIE's classification results, these analyses provide insight to previously characterized information about the purified plasma proteins in our spectral count datasets in the context of established cell lines, which in turn better informed our classification results.

2.4.1 Identifying protein-protein interaction subnetworks present within our datasets

Our experimental spectral count dataset was analyzed for known and predicted protein-protein interactions between the proteins targeted directly by purifying antibodies and the plasma proteins purified in the immunoprecipitation experiments. The UniProt protein IDs of all baits and preys were queried in the STRING database (version 11.0)[90]. To refine the STRING query results, we allowed for known and predicted protein-protein interactions with a STRING-derived medium confidence interaction score (score $\geq$ 0.4), STRING's default query setting. Subnetworks of the known and predicted protein-protein interactions present within our experimental dataset were created using Cytoscape (version 3.8.0)[91] and annotated to denote if MAGPIE classified these interactions with high confidence and, further, denote the FDR these classifications corresponded to.

2.4.2 Evaluating Gene Ontology semantic similarity of interacting protein pairs

Proteins that are interacting with each other are more likely to share functional annotations than proteins that do not. Gene Ontology terms are among such functional annotations. Sets of Gene Ontology (GO)[92,93] terms associated to the bait and prey proteins of our putative protein-protein interactions were therefore analyzed. The GO annotation database was downloaded locally (accessed July 14, 2020) and filtered to omit ambiguous GO terms, such that a GO term was discarded if it annotated more than 1000 proteins within the human proteome. Leaving these GO terms in the analysis would likely bias the results for those terms, masking patterns of the potentially more relevant and specific GO terms related to protein-protein interactions. After associating all bait and prey proteins to their filtered GO terms, the *GOGO* package command-line tool was used, which measures the similarity of two GO terms based on their respective ontologies[94]. A distribution of the cumulative similarity score frequencies between sets of GO terms of high-confidence protein-protein interactions (logistic regression probability ≥ 0.95) was computed, as well as for the remaining putative protein-protein interactions (logistic regression probability < 0.95), separately for the three GO namespaces: cellular component, biological process, and molecular function.

2.4.3 Evaluating gene co-expression of interacting protein pairs

Similarly to GO annotations, proteins that are interacting with each other are more likely to be co-expressed at the RNA level than proteins that do not. Hence, mutual rank (MR) values computed for the bait and prey proteins, as reported by COXPRESdb (version 7.3)[95], for our putative protein-protein interaction pairs were analyzed. These values are computed as the geometric mean of rankings based on the Pearson's correlation coefficient of gene A to gene B

and gene B to gene A. The human gene co-expression data, composed of the union of RNA-seq and microarray data of different cell types and tissues, was locally downloaded (accessed June 11, 2021)[96] and used to map MR values to putative interaction bait and prey proteins. Following mapping, the cumulative distribution of MR values were computed for scores between 0 and 20,000 in bins of 1,000. This was done separately for high-confidence putative protein-protein interactions, whose logistic regression probability ≥ 0.95, and the remaining lesser and non-confident putative protein-protein interactions. The distribution of MR value cumulative frequencies was plotted to evaluate the difference in gene co-expression between sets of putative protein-protein interactions. Finally, to evaluate whether the two sets of MR values came from the same distribution, a Kolmogorov-Smirnov test was implemented, using the *dgof* package (version 1.2) in R, to compute a D_{max} and *p*-value for statistical significance assessment.

3 Results

3.1 Unsupervised machine learning can be used to identify experimental negative controls

The confident identification of a set of experimental negative controls was required prior to constructing the classifying model for putative protein-protein interactions in human plasma. We hypothesize that antibodies used in conditions where the target proteins are not present could be used as experimental negative controls, since the vast majority of the proteins purified with these antibodies are likely to bind the antibody itself and not its target. This notion was tested by running two unsupervised machine learning algorithms on the spectral counts of the set of proteins purified in the empirical experiments and the negative controls (*see Methods 2.2.2 – 2.2.4*). The rationale behind this analysis is that negative controls, which cluster together in both supervised analyses, with little to no empirical experiments, are likely to be composed of mostly non-specific protein bindings. Antibodies targeting proteins not expected to be present in human plasma do not purify their target and instead show similar sets of proteins in their purification, highlighting that they likely purify high abundance contaminant proteins from the human plasma samples. The bootstrapped hierarchical clustering analysis, performed on the normalized spectral count data, revealed that empirical experiments mostly cluster separately from negative controls (Figure 6). Two negative controls (Antibody IDs: Anti-SRB7-IgG Ab 1, Anti-SRB7-IgG Ab 2) had noticeably higher spectral count abundance than all other negative controls. Figure 6A shows these antibodies clustering with two other negative controls (Antibody IDs: Anti-HA-IgG, Anti-RPAP2-IgG). Figure 6B, excluding the two outlier Anti-SRB7-IgG experiments, shows one cluster composed almost exclusively of negative controls and a second cluster of exclusively empirical experiments.

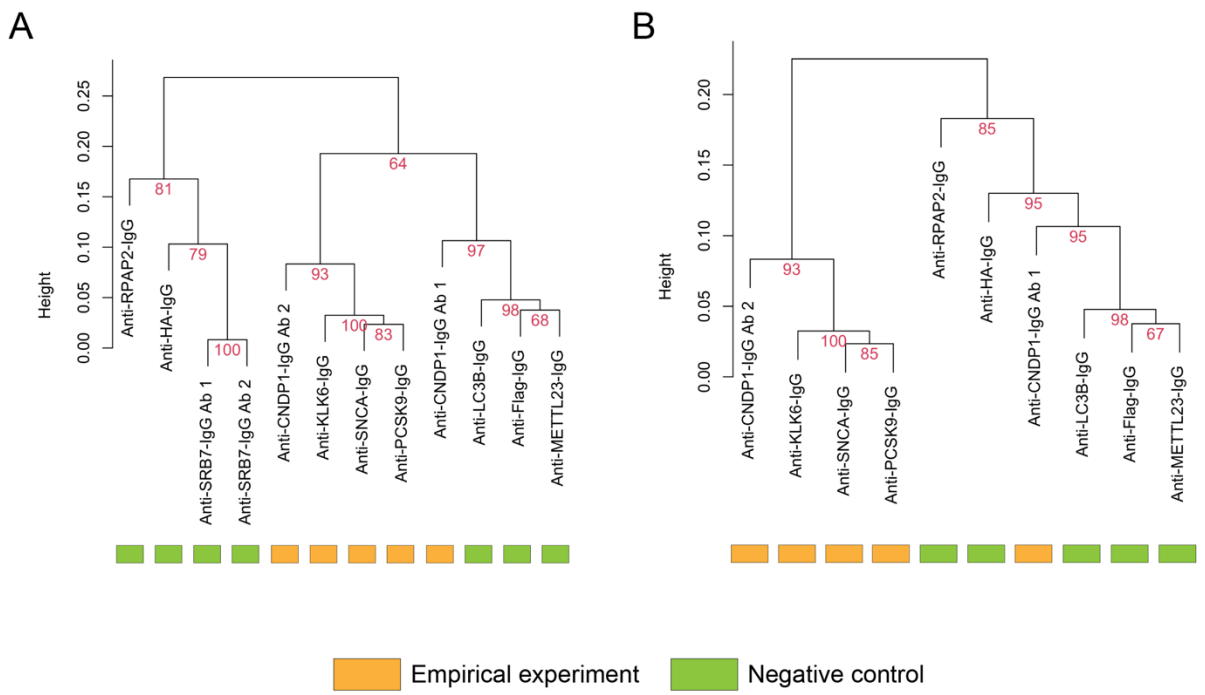

Figure 6. Hierarchical clustering of empirical experiments and negative controls. Complete linkage analysis on the normalized spectral count data prior to excluding the Anti-SRB7-IgG Ab 1 and Anti-SRB7-IgG Ab 2 experiments (A). Complete linkage analysis on the normalized spectral count data after excluding the Anti-SRB7-IgG Ab 1 and Anti-SRB7-IgG Ab 2 experiments (B). Both analyses were implemented with 10,000 bootstrapping iterations, whose results are denoted by the red numbers next to the inner nodes of the dendrogram. Branch height represents the relative distance between the normalized spectral count profiles of the different experiments. Designations of experiments are colour-coded at the bottom of each leaf.

Furthermore, the bootstrapping results are relatively higher in the two dominating clusters of Figure 6B versus those of Figure 6A, indicating more robust clusters in Figure 6B. The principal component analysis was run on the same normalized spectral count dataset to confirm the hierarchical clustering analysis results (Figure 7) and showed that the two Anti-SRB7-IgG experiments would be unreliable for modeling background noise in the mass spectrometry data (Figure 7A). This analysis revealed the same clustering of negative controls separated from most of the empirical experiments (Figure 7B). As such, interpreting the results from both the bootstrapped hierarchical clustering analysis and principal component analysis allowed for the confident identification of a set of negative control that appear to purify the same population of proteins and could potentially be used for further analysis and supervised machine learning model construction.

These analyses revealed five negative control experiments involving the antibodies targeting Flag, HA, LC3B, METTL23, and RPAP2 (Antibody IDs: Anti-Flag-IgG, Anti-HA-IgG, Anti-LC3B-IgG, Anti-METTL23-IgG, and Anti-RPAP2-IgG, respectively). Additionally, five experimental antibodies that successfully purified their target proteins were identified, targeting CNDP1 (x2), KLK6, SNCA, and PCSK9 (Antibody IDs: Anti-CNDP2-IgG Ab 1, Anti-CNDP1-IgG Ab 2, Anti-KLK6-IgG, Anti-SNCA-IgG, and Anti-PCSK9-IgG, respectively). These experiments were chosen because they mostly clustered independently of the negative controls, apart from Anti-CNDP1-IgG Ab 1. This highlights that the proteins purified with these antibodies do not resemble those purified with the negative controls. Therefore, they may contain some specific protein-protein interactions related to the protein targeted by these antibodies. The mass spectrometry data from these empirical experiments was used for all further analyses.

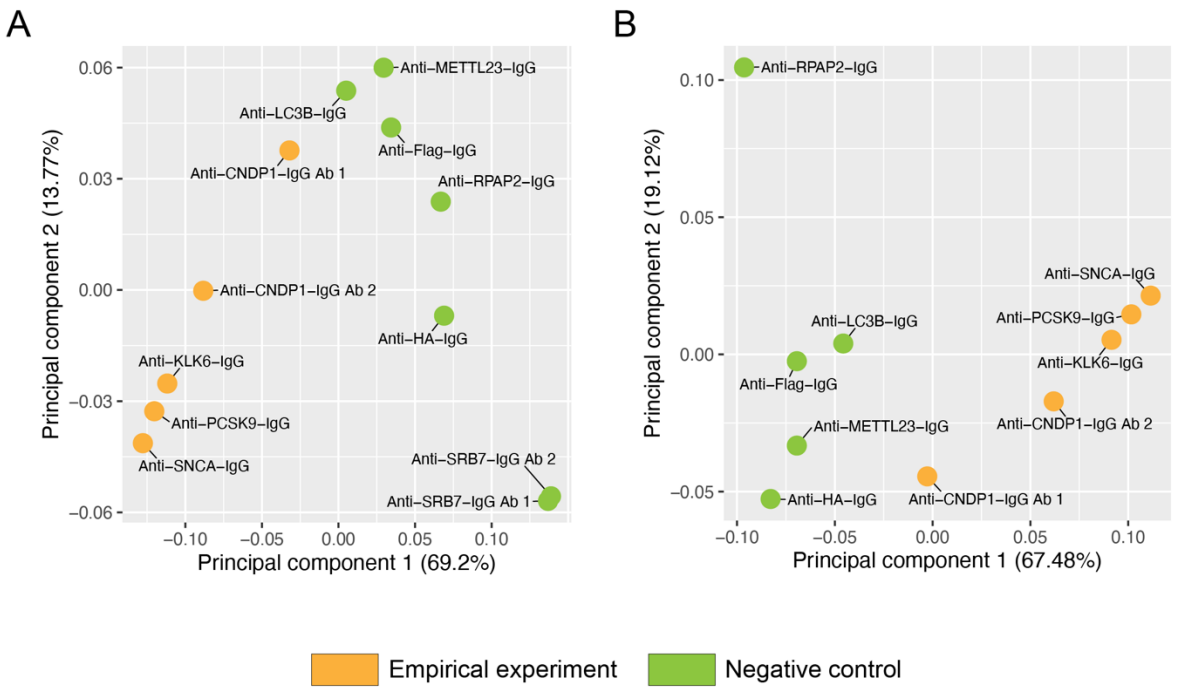

Figure 7. Principal component analysis on empirical experiments and negative controls.

Principal component analysis on the normalized spectral count data prior to excluding the Anti-SRB7-IgG Ab 1 and Anti-SRB7-IgG Ab 2 experiments (A). Principal component analysis on the normalized spectral count data after excluding the Anti-SRB7-IgG Ab 1 and Anti-SRB7-IgG Ab 2 experiments (B). Amount of variance in the data explained by the principal component is indicated in the axis labels. Experiments are colour-coded based on whether they are empirical experiments or controls.

3.2 Z-scores identify likely high-confidence protein-protein interactions

The biological relevance of protein-protein interactions of the five experimental antibodies was measured using Z-scores and fold-change calculation assessments. Z-scores and fold-change values were calculated for both the empirical antibody experiments and negative controls. We call the Z-scores and fold-change values our confidence scores, as we used them to assess confidence in the purified proteins within the empirical experiments. The distribution of the confidence scores calculated for purified proteins in the empirical experiments were plotted separately. The Z-scores centered around zero (Figure 8A) and the fold-change values centered around one (Figure 8B), confirming that the spectral counts belonging to many of the purified proteins in the empirical experiments resemble the background contamination. Following computation, these scores were used as thresholds, for which false discovery rates were calculated (*see Methods 2.3.2*). The false discovery rates were plotted, separately, as a function of the Z-score (Figure 8C) and fold-change (Figure 8D) thresholds.

Figures 8C and 8D demonstrate that as these score thresholds increase, their respective false discovery rates decrease. In other words, as the criteria to be considered a high-confidence putative protein-protein interaction becomes more stringent, the fewer false-positive identifications occur, which is the expected behaviour of an adequate confidence score. The Z-score assessment performed reasonably well, with a false discovery rate of 25.21% at a Z-score threshold of 3. This suggested that there was a reasonably small proportion of false-positive protein-protein interaction identifications when the criteria to be considered a high-confidence interaction was having a Z-score greater than or equal to 3. This is the same threshold that was used by Fredolini et al.[75] in their study of antibody specificity. At the false discovery rate of 25.21%, there are 226 high-confidence protein-protein interactions.

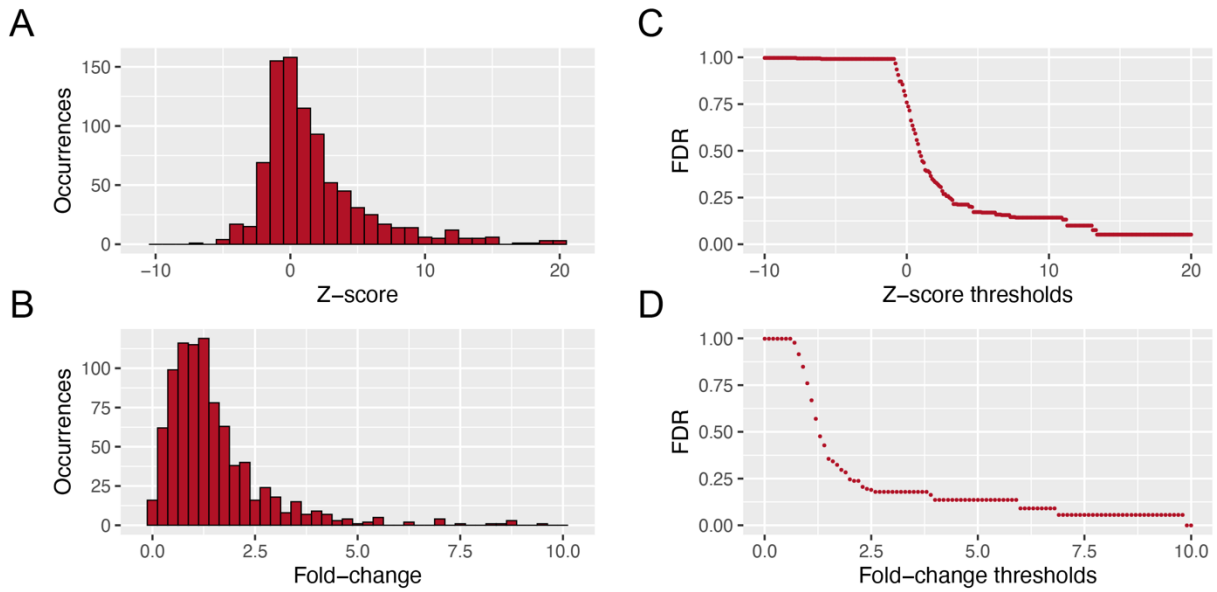

Figure 8. Analysis of Z-scores and fold-change values. Distribution of calculated Z-scores (A) and fold-change values (B) for all putative protein-protein interactions detected in empirical experiments. Estimated false discovery rates (FDRs) calculated at each Z-score (C) and fold-change value (D) threshold.

This suggested that approximately 56 of these identified interactions were false-positive interactions. Therefore, the Z-score threshold of greater than or equal to 3 was used to identify the high-confidence putative protein-protein interactions for training a supervised machine learning classifier.

Additionally, the fold-change analysis also performed reasonably well, with a false-discovery rate of 24.70% at a fold-change threshold of 2. At this false discovery rate, there are 193 high-confidence protein-protein interactions, suggesting that approximately 48 of these identified interactions were false-positives interactions. With similar performances, the Z-score

threshold for identifying high-confidence putative protein-protein interactions for training a supervised machine learning classifier was chosen because of additional literature support (*see Discussion 4.2*).

3.3 MAGPIE identifies protein-protein interactions with a reasonable false discovery rate

It was hypothesized that machine learning could be used to computationally model the contaminating non-specific binding of antibodies with circulating plasma proteins. As such, a supervised logistic regression model was constructed to classify putative protein-protein interactions and contaminating non-specific binding in the experiments carried out in human plasma samples (s*ee Methods 2.3.3 – 2.3.4*). In constructing the training datasets, the positive training dataset was composed of putative protein-protein interactions whose Z-scores were greater than or equal to three, totalling 226 positive training examples. To ensure a class-balanced training dataset, 226 purifications in the negative control experiments were randomly sampled to yield the negative training examples. The proportion of negative training examples belonging to each negative control was plotted in Figure 9, showing that the resulting 226 negative training examples roughly represent all five negative control experiments equally. With the completed training datasets, MAGPIE's logistic regression model was trained, and its performance was diligently evaluated through the implementation of our leave-one-control out (LOCO) cross-validation methodology (*see Methods 2.3.5*). False discovery rates were computed at given logistic regression probability thresholds for the false positive identifications in each of the negative control experiments (Figure 10). An overall false discovery rate was then derived, demonstrating that MAGPIE performs reasonably well with a false discovery rate of 20.77% at the 0.99 logistic regression probability threshold.

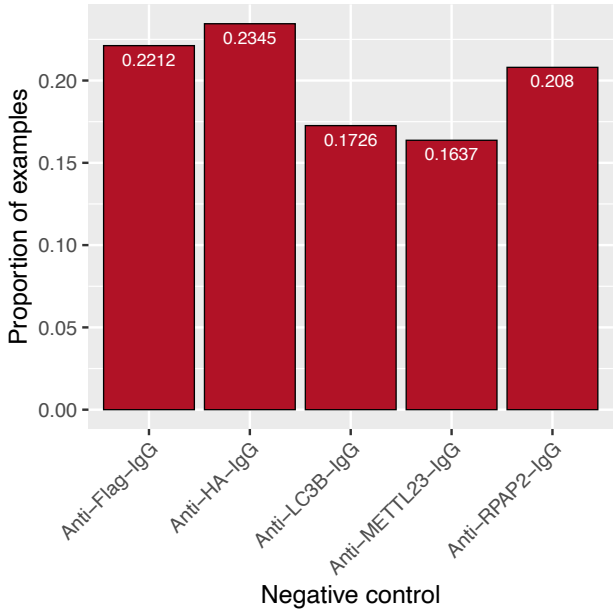

Figure 9. Proportion of negative example random sampling for supervised machine learning training. Proportion of random sampling from each negative control experiment when constructing a negative example training dataset. Exact proportions are denoted in white in the appropriate bars.

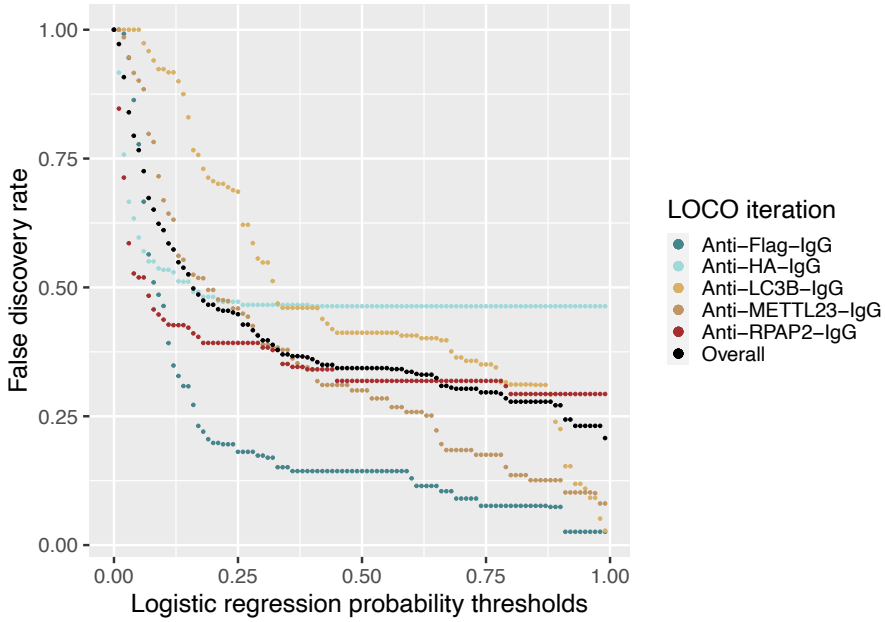

Figure 10. Classifier FDR estimation from leave-one-control-out cross-validation scheme. FDRs estimated at logistic regression probability thresholds, for each of the five leave-one-control-out cross-validation iterations corresponding to a given negative control experiment (Anti-Flag-IgG: dark blue, Anti-HA-IgG: light blue, Anti-LC3B-IgG: yellow, Anti-METTL23-IgG: brown, Anti-RPAP2-IgG: maroon). An overall FDR, derived from the results of the five leave-one-control-out iterations, is superimposed (Overall: black).

At this false discovery rate, there are 68 high-confidence protein-protein interactions, suggesting that approximately 14 of these identified interactions were false-positives interactions. However, we find that there are 19 false-positive identifications at this stringent logistic regression probability threshold, derived from the LOCO cross-validation. Table 3 summarizes the number of false-positive identifications at the logistic regression probability thresholds ranging from 0.90 to 0.99. In essence, these are the number of purifications in the negative control experiments that

were falsely classified with high a probability of being a true interaction at the aforementioned probability thresholds.

Table 3. False-positive protein-protein interaction identifications across probability thresholds.

Logistic probability threshold	Cumulative number of false-positive identifications	False discovery rate (%)
0.99	19	20.77
0.98	22	23.14
0.97	22	23.14
0.96	22	23.14
0.95	26	23.14
0.94	27	23.14
0.93	27	23.14
0.92	27	24.37
0.91	30	24.37
0.90	30	27.12

Similarly, Table 4 summarizes the number of false-positive identifications made at the 0.99 logistic regression probability threshold across the five LOCO cross-validation iterations, wherein the purifications from the left out negative control experiment were used to compute the false discovery rate of that given LOCO iteration. This table highlights that a large number of false-positive identifications originate from the Anti-HA-IgG experiment. This suggests that the modeling of contaminants was less effective to capture contamination events for this experiment. The same applies to a lesser extent for the experiment involving the antibody targeting RPAP2. Figure 11 shows that, once again, the number of purifications in the negative controls that were randomly sampled to be used as negative training examples are evenly shared across the remaining negative control experiments in that LOCO iteration.

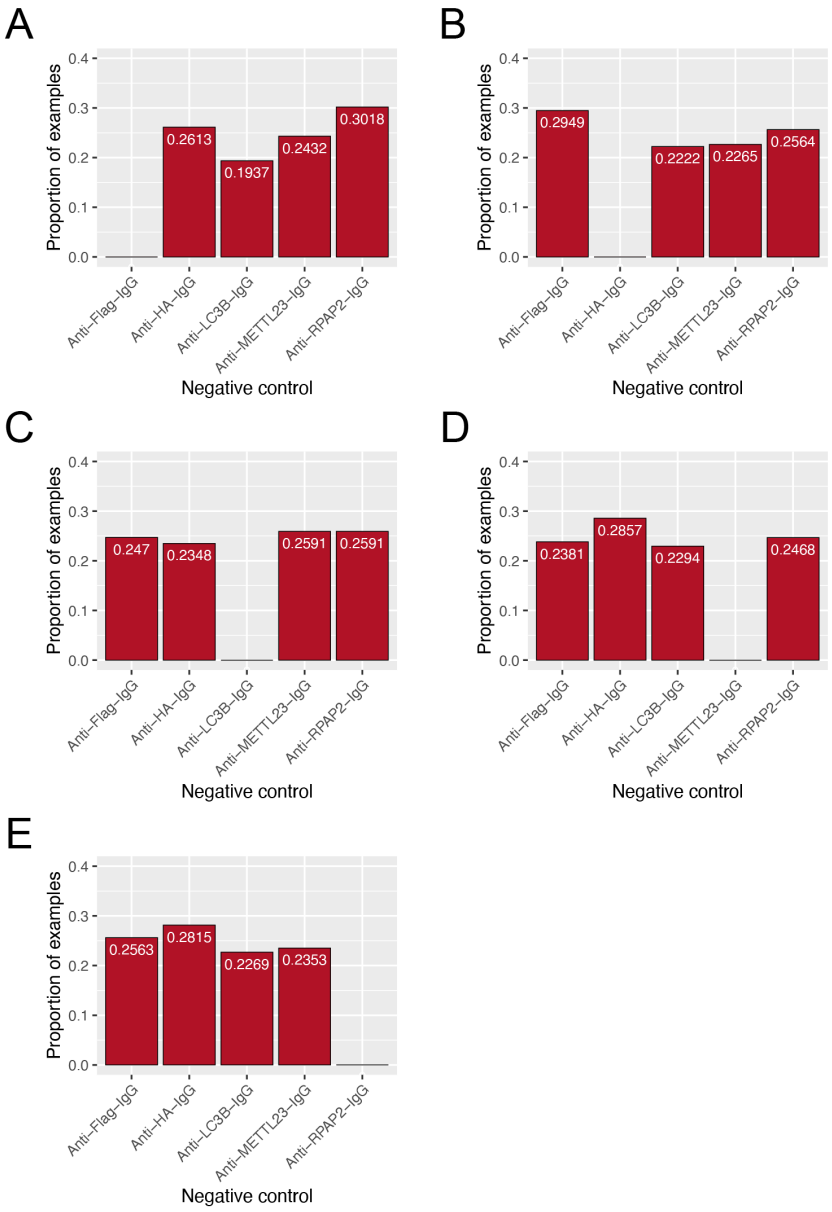

Figure 11. Proportion of negative example random sampling for supervised machine learning training for each LOCO cross-validation iteration. Analyzing the proportion of random sampling from each remaining negative control experiment when constructing a negative example training data, after leaving out the Anti-Flag-IgG (A), Anti-HA-IgG (B), Anti-LC3B-IgG (C), Anti-METTL23-IgG (D), and Anti-RPAP2-IgG (E) negative controls, respectively. Exact proportions are denoted in white in the appropriate bars.

Table 4. False-positive protein-protein interaction identifications in each cross-validation iteration at the logistic regression probability threshold of 0.99.

Leave-one-control-out cross-validation iteration	Number of false-positive identifications
Anti-Flag-IgG	1
Anti-HA-IgG	19
Anti-LC3B-IgG	1
Anti-METTL23-IgG	3
Anti-RPAP2-IgG	10

Because MAGPIE implements a logistic regression model, the weight associated to each feature could easily be extracted. Fold-change relative to spectral count in the controls was given notably more weight than all other features, with a coefficient of -1.35. Table 5 summarizes all features weights, determined by the model's optimization during training. Upon training MAGPIE with each feature individually, training with the most discriminative feature led to virtually identical results as training with all five. The other four features led to severely underfit models that each failed to identify any high-confidence interaction. Despite these largely different weights, and subsequent poor models, all five features were kept, for generalizability.

Table 5. Logistic regression model weights computed for all five of MAGPIE's features.

Feature	Feature weight
Spectral count	-0.0817
Average spectral count in the controls	-0.00824
Standard deviation of spectral count in the controls	0.03
Spectral count maximum in the controls	0.0287
Spectral count fold-change relative to the controls	-1.35

All possible combinations of three negative control experiments were used to construct MAGPIE to evaluate how a reduction in available data for modeling background noise may affect MAGPIE's performance. Figure 12A shows that the false discovery rates computed across the logistic regression probability thresholds of all three negative control combinations perform more poorly than the original full negative control dataset. Subsequently, Figure 12B demonstrates that, at these poorer false discovery rates, fewer putative protein-protein interactions are classified with high confidence.

3.4 MAGPIE's algorithm is robust despite its instances of stochasticity

Throughout MAGPIE's algorithm, there are two instances of random sampling. The first is during the addition of pseudocount values in the negative control experiments (*see Methods 2.3.1*) and the second is when negative training examples are randomly selected (*see Methods 2.3.3*). Because of these instances of stochasticity, it was necessary to evaluate the robustness of MAGPIE's algorithm. Figure 13 shows the resulting false discovery rates after running MAGPIE 1000 times without taking steps to ensure reproducible results (i.e., the instances of random sampling were truly random). Little variation is observed in the resulting false discovery rates, indicating that MAGPIE's reported performance is not influenced by the randomly added pseudocount values or randomly sampled negative training examples.

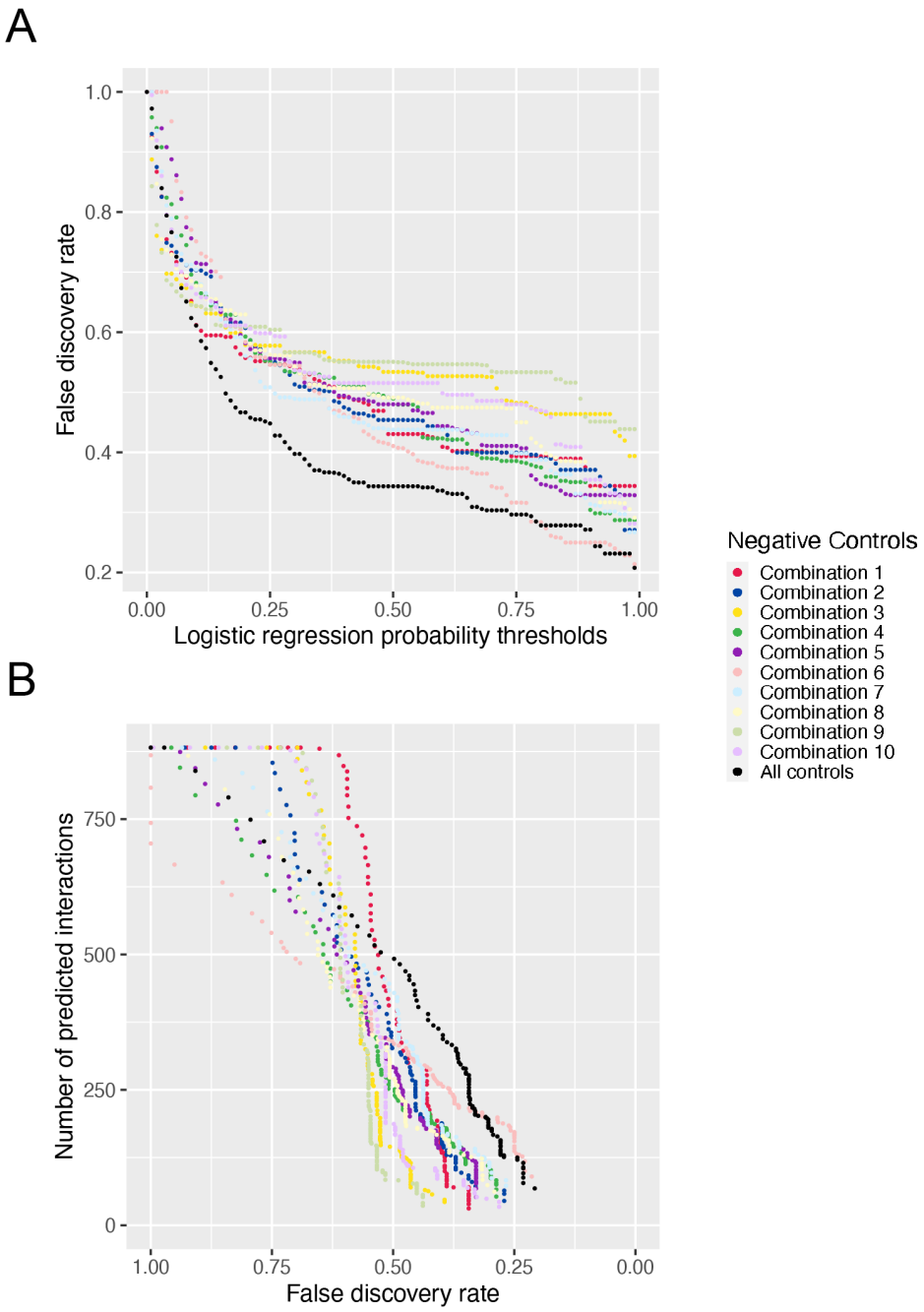

Figure 12. Evaluating MAGPIE's performance when training on different combinations of negative controls. Comparison of FDRs at given logistic regression probability thresholds for all possible combinations of three experimental negative controls (A). Comparison of the number of interactions identified at given FDR thresholds for all possible combinations of three experimental negative controls (B). Combination 1: Anti-Flag-IgG, Anti-HA-IgG, Anti-RPAP2-

IgG; Combination 2: Anti-Flag-IgG, Anti-METTL23-IgG, Anti-RPAP2-IgG; Combination 3: Anti-HA-IgG, Anti-METTL23-IgG, Anti-RPAP2-IgG; Combination 4: Anti-LC3B-IgG, Anti-METTL23-IgG, Anti-RPAP2-IgG; Combination 5: Anti-HA-IgG, Anti-LC3B-IgG, Anti-METTL23-IgG; Combination 6: Anti-Flag-IgG, Anti-LC3B-IgG, Anti-METTL23-IgG; Combination 7: Anti-Flag-IgG, Anti-HA-IgG, Anti-METTL23-IgG; Combination 8: Anti-Flag-IgG, Anti-LC3B-IgG, Anti-RPAP2-IgG; Combination 9: Anti-HA-IgG, Anti-LC3B-IgG, Anti-RPAP2-IgG; Combination 10: Anti-Flag-IgG, Anti-HA-IgG, Anti-LC3B-IgG.

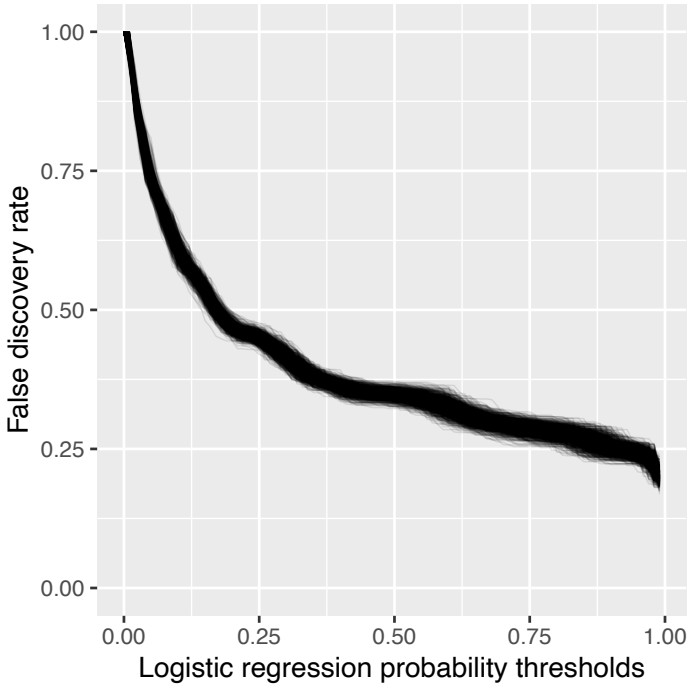

Figure 13. Evaluating the robustness of MAGPIE's algorithm. FDRs estimated for 1000 randomized runs of MAGPIE as a function of the logistic regression probabilities, assessing the effect of random sampling instances within the algorithm.

3.5 MAGPIE identifies plasma protein-protein interactions with high-confidence

We hypothesized that supervised machine learning could be used to computationally model the contaminating non-specific binding occurring with circulating human plasma proteins. This was tested using MAGPIE, which implements a logistic regression model to classify putative protein-protein interactions and instances of antibody non-specific binding from experiments taking place in human plasma samples (*see Methods 2.3.3 – 2.3.4*). The output of MAGPIE's testing runs is the probabilities for the 882 putative protein-protein interactions combinations, predicting whether were high-confidence interactions or antibody non-specific

binding. At the logistic regression probability threshold of greater than or equal to 0.99, MAGPIE reports 68 high confidence protein-protein interactions, listed in Table 6. Of importance, all five of the positive controls present within the input spectral count dataset were identified with high confidence, which we would expect with a well-performing classification model. Briefly, the positive controls are the successfully purified proteins targeted by antibodies in the empirical IP experiments. Figure 14 is a heatmap of the $\log_{10}$-transformed spectral counts belonging to these high-confidence interactions. This visualization reveals that MAGPIE identifies putative protein-protein interactions that range from low to high spectral count abundance. It further reveals that several high-confidence interactions were quantified with spectral count abundances that mirror those of the proteins purified in the negative control experiments.

Table 6. High-confidence protein-protein interactions identified by MAGPIE (N = 68) with a probability ≥ 0.99 and an FDR of 20.77%.

Purifying antibody	Purified protein interactor
Anti-CNDP1-IgG Ab 1	CERU_HUMAN
Anti-CNDP1-IgG Ab 1	CNDP1_HUMAN*
Anti-CNDP1-IgG Ab 1	C4BPA_HUMAN
Anti-CNDP1-IgG Ab 1	CO5_HUMAN
Anti-CNDP1-IgG Ab 1	HEP2_HUMAN
Anti-CNDP1-IgG Ab 1	ITIH1_HUMAN
Anti-CNDP1-IgG Ab 1	AFAM_HUMAN
Anti-CNDP1-IgG Ab 1	FBLN1_HUMAN
Anti-CNDP1-IgG Ab 1	SEPP1_HUMAN
Anti-CNDP1-IgG Ab 1	MASP1_HUMAN
Anti-CNDP1-IgG Ab 1	SAMP_HUMAN

Anti-CNDP1-IgG Ab 1	COL11_HUMAN
Anti-CNDP1-IgG Ab 1	HV206_HUMAN
Anti-CNDP1-IgG Ab 1	FCN2_HUMAN
Anti-CNDP1-IgG Ab 1	TETN_HUMAN
Anti-CNDP1-IgG Ab 1	C4BPB_HUMAN
Anti-CNDP1-IgG Ab 1	KV104_HUMAN
Anti-KLK6-IgG	APOA2_HUMAN
Anti-KLK6-IgG	CERU_HUMAN
Anti-KLK6-IgG	CNDP1_HUMAN
Anti-KLK6-IgG	HEP2_HUMAN
Anti-KLK6-IgG	SPP24_HUMAN
Anti-KLK6-IgG	ITIH1_HUMAN
Anti-KLK6-IgG	AFAM_HUMAN
Anti-KLK6-IgG	CO9_HUMAN
Anti-KLK6-IgG	FBLN1_HUMAN
Anti-KLK6-IgG	A2GL_HUMAN
Anti-KLK6-IgG	MASP1_HUMAN
Anti-KLK6-IgG	SAMP_HUMAN
Anti-KLK6-IgG	ZA2G_HUMAN
Anti-KLK6-IgG	HV206_HUMAN
Anti-KLK6-IgG	FCN2_HUMAN
Anti-KLK6-IgG	KLK6_HUMAN*
Anti-KLK6-IgG	FA11_HUMAN
Anti-KLK6-IgG	KV104_HUMAN
Anti-SNCA-IgG	ALBU_HUMAN
Anti-SNCA-IgG	C4BPA_HUMAN
Anti-SNCA-IgG	A1BG_HUMAN
Anti-SNCA-IgG	ITIH1_HUMAN
Anti-SNCA-IgG	CO9_HUMAN
Anti-SNCA-IgG	FBLN1_HUMAN

Anti-SNCA-IgG	A2GL_HUMAN
Anti-SNCA-IgG	MASP1_HUMAN
Anti-SNCA-IgG	SAMP_HUMAN
Anti-SNCA-IgG	HV206_HUMAN
Anti-SNCA-IgG	FCN2_HUMAN
Anti-SNCA-IgG	TETN_HUMAN
Anti-SNCA-IgG	LUM_HUMAN
Anti-SNCA-IgG	SYUA_HUMAN*
Anti-SNCA-IgG	KV104_HUMAN
Anti-PCSK9-IgG	CERU_HUMAN
Anti-PCSK9-IgG	CNDP1_HUMAN
Anti-PCSK9-IgG	IGJ_HUMAN
Anti-PCSK9-IgG	ITIH1_HUMAN
Anti-PCSK9-IgG	PCSK9_HUMAN*
Anti-PCSK9-IgG	FBLN1_HUMAN
Anti-PCSK9-IgG	A2GL_HUMAN
Anti-PCSK9-IgG	SAMP_HUMAN
Anti-PCSK9-IgG	LUM_HUMAN
Anti-PCSK9-IgG	KV104_HUMAN
Anti-CNDP1-IgG AB 2	APOC1_HUMAN
Anti-CNDP1-IgG AB 2	CNDP1_HUMAN*
Anti-CNDP1-IgG AB 2	IGJ_HUMAN
Anti-CNDP1-IgG AB 2	CO5_HUMAN
Anti-CNDP1-IgG AB 2	ITIH1_HUMAN
Anti-CNDP1-IgG AB 2	KAIN_HUMAN
Anti-CNDP1-IgG AB 2	A2GL_HUMAN
Anti-CNDP1-IgG AB 2	FCN2_HUMAN

* Confident identification of positive control (logistic regression probability ≥ 0.99).

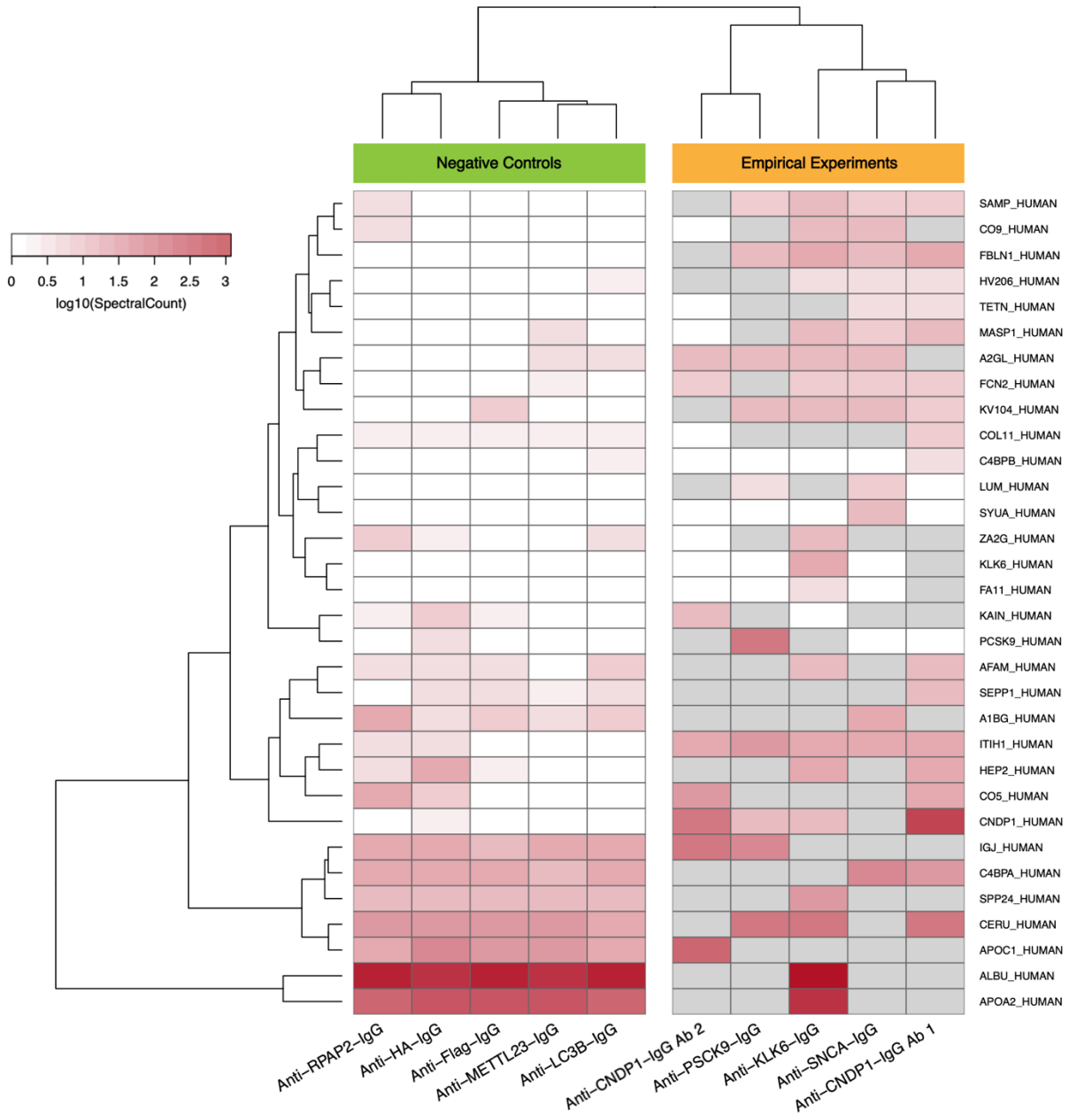

Figure 14. Heatmap of the log10-transformed normalized spectral count data belonging to the 68 high-confidence interactions across all experiments and negative controls.

Comparison of spectral count abundance for high-confidence protein-protein interaction identifications, classified by MAGPIE (logistic regression probability ≥ 0.99, FDR = 20.77%). Grey cells represent successfully purified proteins that were identified by mass spectrometry, but that are not deemed high confidence interactors in the empirical experiments.

3.5.1 MAGPIE outperforms SAINT when detecting protein-protein interactions

When comparing its identifications to SAINT, the leading algorithm for standard confidence assessment of putative protein-protein interactions, MAGPIE outperforms SAINT in several ways. After applying our LOCO cross-validation methods (*see Methods 2.3.5 – 2.3.6*), SAINT identified 18 high-confidence interactions at a probability threshold of 0.99, corresponding to a false discovery rate of 41.53% (Figure 15). At its 0.99 logistic regression probability threshold, MAGPIE identifies more high-confidence interactions (Figure 15A), including the 18 identifications made by SAINT (Figure 15B). Additionally, SAINT fails to identify one of the five positive controls, for Anti-SNCA-IgG, as one of its high-confidence interactions. It is worth mentioning that while SAINT is a leading algorithm for protein-protein interaction reliability assessment it was never designed to deal with human plasma samples nor to use the type of negative controls to detect contamination that I have established in this investigation. It is therefore not surprising to see MAGPIE outperform SAINT to this level and shows that its tailored approach to study plasma is relevant.

3.6 MAGPIE's identifications are corroborated by the STRING database

Because there does not exist any experimentally validated protein-protein interaction datasets for the proteins targeted directly by the purifying antibodies used in our empirical immunoprecipitation experiments, an indirect method was required to validate MAGPIE's results (*see Methods 2.4.1*). Two subnetworks of known or predicted protein-protein interactions in human cell lines could be extracted from STRING for our immunoprecipitation dataset, for the PCSK9 and SNCA proteins (Figure 16).

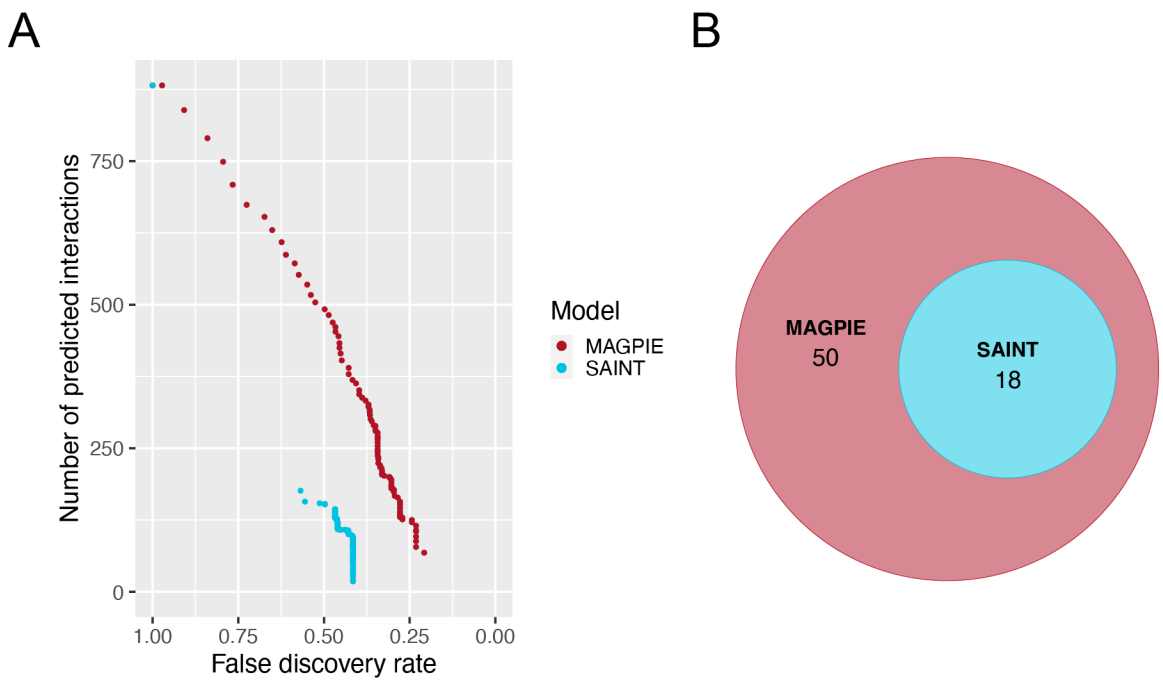

Figure 15. Benchmarking of MAGPIE against SAINT. Comparing the number of predicted true interactions at estimated FDRs (A). Venn diagram of high-confidence interactions (MAGPIE: logistic regression classifier probability ≥ 0.99, FDR = 20.77%; SAINT: probability ≥ 0.99, FDR = 41.53%) as predicted by each model (B).

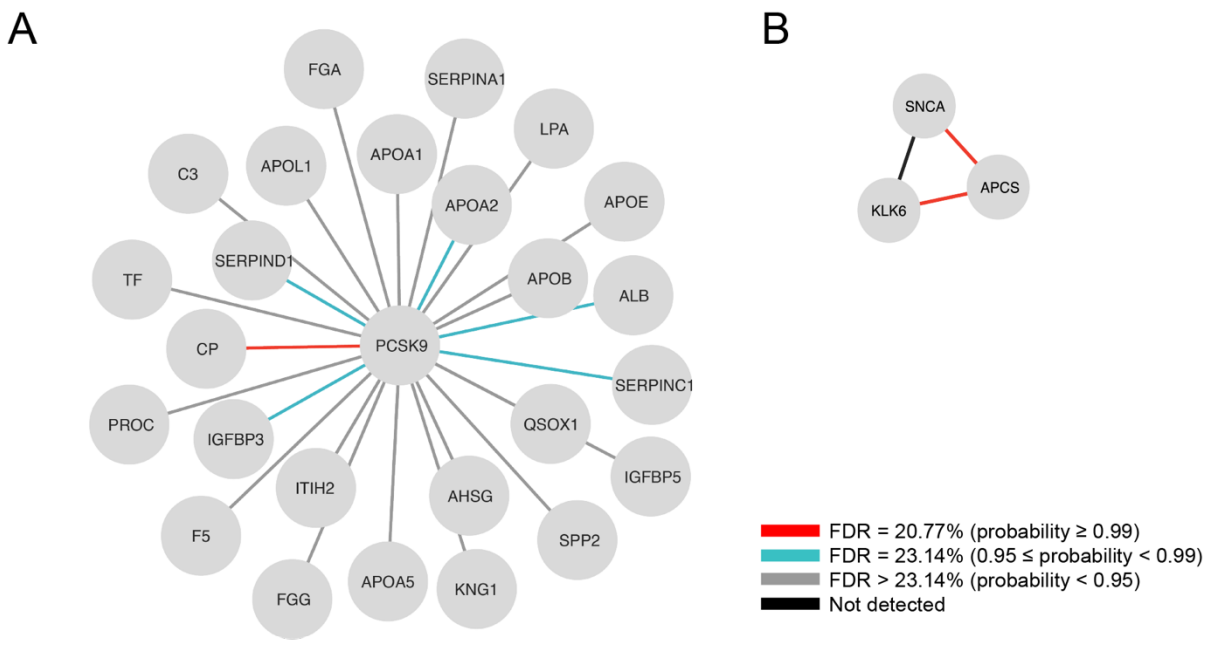

Figure 16. Protein-protein interaction subnetworks as identified by STRING. Comparing the 25 known or predicted protein-protein interactions involving PCSK9, as identified by STRING, to the classification results outputted by MAGPIE (A). Comparing the 2 known or predicted protein-protein interactions involving SNCA, as identified by STRING, to the classification results outputted by MAGPIE (B). Interaction confidence level is colour-coded.

After querying the STRING database, 25 known or predicted protein interactors of the PCSK9 protein were found to be present in our dataset (Figure 16A). Of these, MAGPIE classified one, ceruloplasmin (UniProt ID: CERU_HUMAN; gene: CP), with a logistic regression probability of greater than or equal to 0.99, corresponding to a false discovery rate of 20.77%. Five more known or predicted PCSK9 interactor were identified at a high probability, greater than or equal to 0.95, but less than 0.99 (UniProt IDs: ALBU_HUMAN, ANT3_HUMAN, APOA2_HUMAN, HEP2_HUMAN, and IBP3_HUMAN; genes: ALB, SERPINC1, APOA2, SERPIND1, and

IGFBP3), corresponding to a false discovery rate of 23.14%. Composing the second subnetwork, two known or predicted protein interactors of the SNCA protein were found to be present in our dataset (Figure 16B). One of these was identified as a high-confidence interactor of the SNCA protein, serum amyloid P-component (UniProt ID: SAMP_HUMAN, gene: APCS). The second known or predicted interactor of SNCA was not detected, kallikrein-6 (UniProt ID: KLK6_HUMAN, gene: KLK6). However, the KLK6 protein was another one of the targets by a purifying antibody, Anti-KLK6-IgG, within our dataset. Notably, MAGPIE classified the putative protein-protein interaction between KLK6 and APCS with high confidence. While STRING does not directly validate this interaction, the fact that two interactors of SNCA are interacting in our dataset provides confidence that this previously unreported interaction may likely take place in human plasma.

3.7 Gene Ontology semantic similarity between interacting protein pairs is insignificant

It was hypothesized that high-confidence protein-protein interactions would score higher Gene Ontology (GO) semantic similarity scores than the putative interactions classified with less to no confidence, since interacting proteins are more likely to share functional annotations than non-interacting proteins. To verify this, the cumulative distribution of the GO semantic similarity scores, for the two sets of interaction pairs were evaluated (Figure 17) for each of the three GO domains. However, the resulting two distributions were very similar, and no differences were observed to each other (*see Discussion 4.5*).

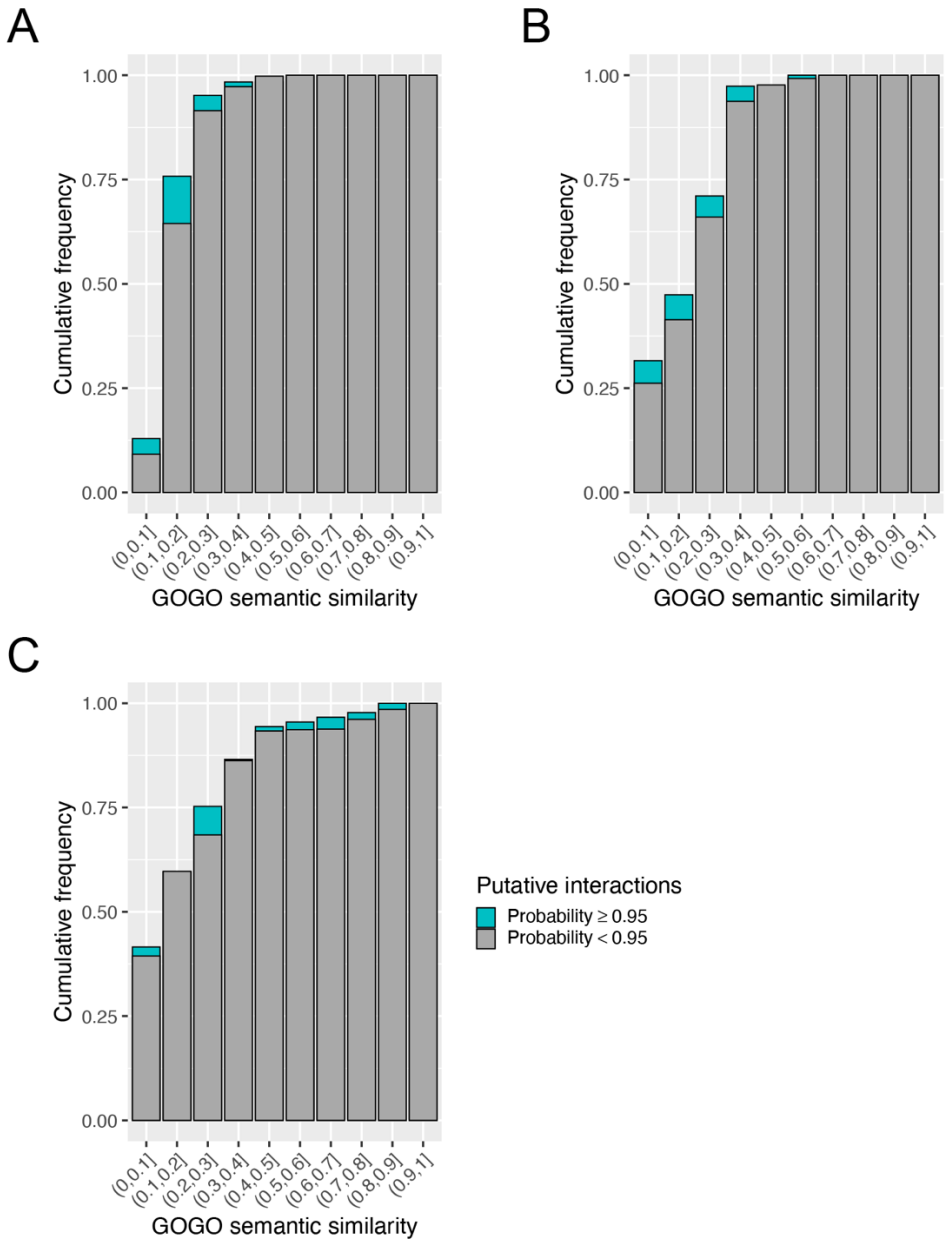

Figure 17. GO semantic similarity scores for pairs of interacting proteins. Comparing the distribution of GO semantic similarity score cumulative frequency, computed using GOGO[94], between high-confidence protein-protein interaction pairs (n = 100), as identified by MAGPIE (logistic regression probability ≥ 0.95) and remaining putative protein-protein interaction pairs (n = 777), for the biological process (A), cellular compartment (B), and molecular function (C) domains.

3.8 Gene co-expression between interacting protein pairs is minimally elevated for high-confidence interaction pairs

Similarly to evaluating GO semantic similarity, it was hypothesized that high-confidence protein-protein interaction pairs are more likely to be co-expressed at the RNA level than proteins that are not interacting. Hence the mutual rank (MR) values as computed by the COXPRESdb of interacting proteins would be lower than their putative interactions counterparts classified with less to no confidence. This notion was evaluated by plotting the distribution of MR value cumulative frequencies for the sets of high-confidence and remaining protein-protein interaction pairs (Figure 18). The two observed distributions were very similar in shape, both trending towards higher MR values. To confirm whether the MR values of the two interaction pair sets indeed came from the same distribution, a Kolmogorov-Smirnov test was performed on the sets of MR values. This evaluation yielded a D_{max} of 0.091945 with a *p*-value of 0.4844, confirming the two sets of MR values are likely coming from the same distribution.

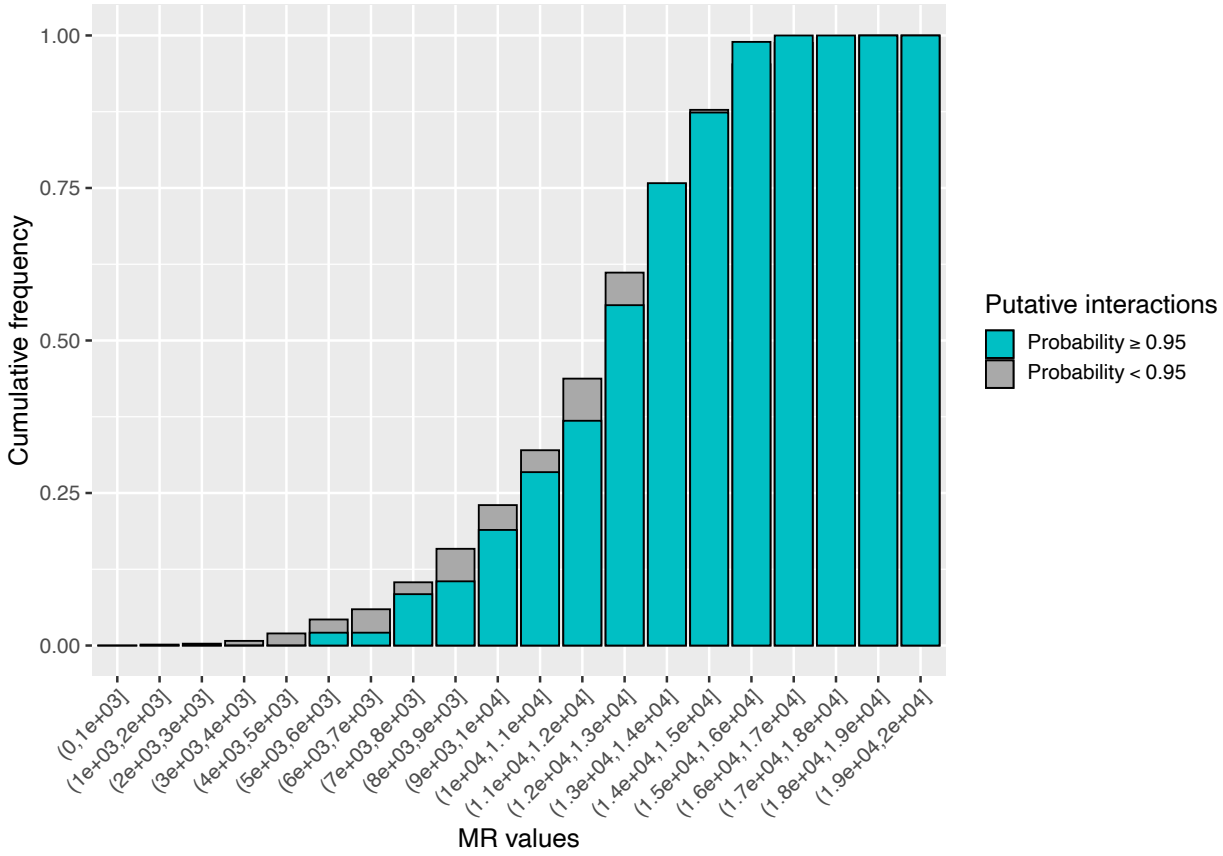

Figure 18. Gene co-expression MR scores between pairs of interacting proteins. Comparing the distribution of gene co-expression MR score cumulative frequency, reported by COXPRESdb, between high-confidence protein-protein interaction pairs (n =95), as identified by MAGPIE (logistic regression probability ≥ 0.95), and remaining putative protein-protein interaction pairs (n = 656). Kolmogorov-Smirnov test $D_{max} = 0.091945$ with *p*-value = 0.4844.

4 Discussion

Until now, no algorithm designed to effectively model the contamination and antibody non-specific binding in human plasma had been proposed. This made it extremely difficult to identify protein-protein interactions in plasma and yielded an over-abundance of false-positive protein-protein interaction identifications when such experiments were attempted, which no computational method could effectively filter from immunoprecipitation experiments. Typically, in cell lines, experimental negative controls and computational modeling are used in combination to perform the filtering of false-positive identifications. However, because they often require a protein of interest to express a molecular tag on its surface through DNA recombination, these methodologies are incompatible for use in human plasma. In addition to a lack of an algorithm for computational modeling, this limitation also makes designing experimental negative controls difficult. To circumvent these limitations, our collaborators performed modified immunoenrichment strategies carried out in human plasma as a means to provide workable label-free quantification data for computational analysis. These strategies bypass the need for molecular tags by directly targeting a protein of interest with an antibody to isolate it and its interactors prior to quantification by mass spectrometry.

In this thesis, I proposed a two-fold computational methodology to first identify a set of reliable negative controls for modeling background contamination and antibody non-specific binding and then to classify putative protein-protein interactions as bona-fide or non-specific. My methodology begins with the hypothesis that experimental negative controls behave similarly to each other, whose antibodies were targeting proteins not expected to be present in human plasma, and differently from empirical experiments, whose antibodies were targeting

known human plasma proteins, which should have specific protein-protein interactions. With a set of confidently identified negative controls, I developed a supervised machine learning algorithm to identify bona fide protein-protein interactions and antibody non-specific binding in human plasma.

4.1 Identifying reliable sets of empirical and negative control experiments

Prior to constructing a supervised machine learning classifier, it was necessary to confidently identify a set of experimental negative controls that could be used to model the background contamination and antibody non-specific binding in the spectral count data. To diminish how arbitrary the choices were and to corroborate the principal component analysis (PCA) results, bootstrapping was implemented with the hierarchical clustering to report on the robustness of clusters. This allowed for the identification experimental negative controls that reproducibly captured examples of contamination and antibody non-specific binding. Similar methodologies have been seen in proteomics, such that PCAs are frequently used to assess proteomic profiles between different experimental groups. For instance, Sarawat et al.[97] published a study in 2019, demonstrating that a PCA could be used to cluster patients with low-grade intraductal papillary mucinous neoplasia from healthy control patients based on their proteomic profiles quantified by LC-MS/MS in human serum samples. In 2020, Poulos et al.[98] investigated the reproducibility of data independent acquisition (DIA)-MS proteomics experiments, quantifying proteins using high performance liquid chromatography coupled to a time of flight mass spectrometer in eight samples with known ratios of ovarian and prostate cancer tissue and yeast. Specifically, the authors were testing technical variation across MS instruments in the absence (or near absence) of sample variation. With data acquired from

running duplicate or triplicate experiments on six instruments, two PCAs were performed, one clustering data points based on sample and the other based on the instrument used. Included in these was a negative control composed of experiments performed with HEK293T cell lysates, which were found to cluster separately from all eight experimental samples. Our methodology differs from these examples, and many other like-examples in the field, because our negative controls are a set of pooled experiments targeting different proteins, as opposed to replicate experiments with the same target protein. Ultimately, this is novel logic for designing the negative controls for computational modelling, further confirmed by our clustering analyses.

While the clustering results highlighted that negative controls in our dataset did share a higher level of similarity than empirical experiments, it could have very well not been the case if the antibodies investigated would have had more differing populations of non-specific bindings, much like the antibody targeting SRB7. In that case, the bootstrapping results could have been poor (i.e., clusters formed with low bootstrapping frequency values) and/or the PCA could have failed to confirm the hierarchical clustering results. Additional analyses would have then been required. Such methods may have included a *t*-distributed stochastic neighbour embedding[99] (t-SNE) analysis or a uniform manifold approximation and projection[100] (UMAP) analysis. Both analyses may be used to reduce dimensionality non-linearly and cluster a dataset, as is often done for datasets with high dimensionality and/or high complexity. One of many such examples includes a 2020 study done by Meyer et al.[101], demonstrating the analysis of complex proteomics data acquired by direct infusion-shotgun proteome analysis (DI-SPA) by DIA-MS. The authors showed that they can use a UMAP analysis to cluster >45,000 proteins based on their treatment time (24 and 6-hour groups) and, further, the proteotypes within the 6-hour treatment group.

Therefore, if our spectral count data had been too complex to separate linearly, then the logical next step would have been to attempt non-linear separation.

4.2 Identifying likely high-confidence putative protein-protein interactions

In addition to identifying a reliable set of negative control experiments for computational modeling, it was also necessary that a criterion for identifying likely high-confidence putative protein-protein interactions be established. Putative interactions that satisfy this criterion would, in theory, be the best examples of bona fide protein-protein interactions in human plasma attainable in the current spectral count dataset. Thus, these would later be used as positive training examples for the machine learning model.

Of the two scores evaluated, Z-scores and fold-change values, Z-scores were selected as this criterion. The justification behind this choice was two-fold. First, this criterion provided the greater number of likely high-confidence protein-protein interaction examples, which allowed for the construction of a reasonably sized training dataset (226 positive training examples, 226 negative training examples, 452 training examples total). The second justification stems from literature support, specifically a 2019 study conducted by Fredolini et al.[75], which sought to evaluate the selectivity of various antibodies to directly enrich their target protein in human plasma samples. In their analysis, the authors targeted 120 proteins directly using immunoprecipitation mass spectrometry (IP-MS/MS) and measured the enrichment of a given antibody for its target protein using Z-scores. They deemed an antibody to be enriched for its target if its Z-score was greater than or equal to 3, which was calculated using the quantification intensity average and standard deviation of a given protein across all empirical antibodies that successfully purified it. It is important to note the difference between the Z-score calculation in

Fredolini et al.'s study versus that in this thesis, wherein the Z-scores computed for putative protein-protein interactions in our dataset are relative to the spectral count average and standard deviation across a set of negative controls. In Fredolini's work, all antibodies were targeting proteins likely or possibly present in plasma. Hence, the background intensity of a protein is likely to include signal from the detection of the protein in the context of a bona fide interaction with other proteins targeted by some of the 120 antibodies tested. On the other hand, our methodology facilitates the measurement of a purified protein against a set of experiments highly unlikely to contain bona fide interactions involving the protein.

4.2.1 Limitations of using Z-scores to identify likely high-confidence protein-protein interactions

The use of Z-scores to identify likely high-confidence protein-protein interactions may be hindered by several limitations. First and foremost, our dataset is relatively small. The experimental negative control spectral count dataset that is used to model the background abundance of contamination and antibody non-specific binding is composed of the purifications belonging to five experiments. Because of how they are calculated, a small sample size is a recognized limitation of Z-scores. With fewer pieces of data available to calculate an average and standard deviation, the quality of a Z-score therefore diminishes. This may further introduce bias for interactions in given experiments. For example, if the spectral count standard deviations are underestimated, putative protein-protein interactions will be assigned a larger Z-score, increasing sensitivity, but wrongfully labelling putative interactions as being of likely high confidence. Conversely, an overestimated standard deviation would result in smaller Z-scores. This would have the potential to hinder the ability to identify likely high-confidence interactions

and, therefore, decrease sensitivity. Both situations could harm the construction of a training dataset for supervised machine learning, by yielding a biased or incomplete set of positive training examples.

To improve results moving forward, the clearest next step would be to obtain a larger mass spectrometry spectral count dataset, composed of more empirical and negative control experiments, and re-run our algorithm. With more negative control experiments, Z-score calculations may more accurately represent the background abundance of contamination and antibody non-specific binding. They could also more efficiently capture spurious contamination events. Additionally, a more tailored score could be implemented that may better represent the background noise in the negative control spectral count data. This score could be one that is uniquely derived for the purpose of measuring dynamic deviations of putative protein-protein interactions relative to purifications across negative control experiments. For example, the CompPASS[51] algorithm for standard confidence assessment of protein-protein interactions uses a D-score, which the authors derived themselves. This score considers the spectral count of the observed interactor in a given experiment, the frequency of the observed interactor across experiments, and the reproducibility of the putative interaction across replicate experiments. Modifying our methodology to incorporate such a score may help to better identify likely high-confidence protein-protein interactions and, thus, build an improved machine learning training dataset.

4.3 Limitations of the supervised machine learning model

We hypothesized that a supervised machine learning model could be used to computationally model the background abundance of contamination and antibody non-specific

binding in human plasma, to filter out false-positive protein-protein interaction identifications and confidently identify bona fide ones. To test this hypothesis, we created MAGPIE which implements a logistic regression classification model, constructed using spectral count data belonging to various immunoprecipitation experiments, wherein antibodies were used to target proteins directly (*see Methods 2.3*). MAGPIE outputs the probability that a given putative protein-protein interaction is true or the result of non-specific binding.

The training and testing datasets that could be produced from our spectral count data facilitated the construction of a supervised machine learning model with moderately good performance (FDR = 20.77%). However, the negative control experiments targeted both proteins that are known to exist elsewhere in humans, such as RNA polymerase II associated protein 2 (Antibody ID: Anti-RPAP2-IgG), and synthetic molecular tags, such as the FLAG tag (Antibody ID: Anti-Flag-IgG). While, in theory, the negative control antibodies were not targeting known human plasma proteins, there is a significant chance that they had some degree of affinity for plasma proteins. Indeed, proteins may be sharing domains with the intended target of the antibody. Isoforms of the protein may also enter circulation. This notion may be particularly true for the proteins known to exist elsewhere in the human body. For instance, the protein-protein interaction network for RPAP2 has been extensively characterized in human cell lines[34,57]. Hemagglutinin (HA) is the most abundant surface glycoprotein of the influenza A virus[102], the seasonal flu, and the human host response to both the viral infection and its vaccination have been extensively characterized[103]. Should there be any affinity between the negative control antibodies and circulating plasma proteins, then the resulting protein purifications may not make the best examples of contamination and antibody non-specific binding. Ideally, the proteins

purified in the negative controls would solely constitute the background noise in the spectral count data for computational modeling.

4.4 Indirect validation of high-confidence interactions using external public repositories

Because there does not exist experimentally validated datasets for protein-protein interactions in human plasma for the four bait proteins in our spectral count dataset, external public repositories needed to be leveraged to indirectly validate the high-confidence interactions identified by MAGPIE. The STRING protein-protein interaction database has been used in the past for validating interactions identified by predicting interactions from sequences and structures [104]. Because STRING is a comprehensive database of both known and predicted protein interactions, it was used to query MAGPIE's high-confidence identifications against known or predicted interactions present in our dataset. While this approach was not an absolute method for validation, it provided more insight into the confidence and potential relevancy of MAGPIE's identifications. It is worth noting that since methodologies to map protein-protein interactions in plasma are still in their infancy, it is expected that very few of the interactions reported by MAGPIE would have been previously deposited in STRING. Hence, the small overlap between MAGPIE's results and STRING is expected, since most of the interactions in the repository are based on experiments in cell lines.

The validation using STRING was supplemented with the gene co-expression analysis using the COXPRESS database, as it is well understood that interacting proteins are often the downstream products of genes that are co-expressed[105,106]. Our methods failed to reveal an enrichment for high gene co-expression in high confidence interaction pairs, but instead revealed a statistically significant enrichment for low gene co-expression in low confidence interactions

pairs. While this is contradictory to the above assumption, there are several regulation steps between RNA transcription and the interaction of two proteins in plasma. Indeed, gene expression does not always correlate with protein expression[107]. Furthermore, to interact in plasma, proteins need to be transported and secreted together or at the same time, which also adds a layer of regulation. It is likely that such regulation mechanisms explain why no link between gene co-expression and high-confidence interactions were found. Lastly, the human COXPRESdb dataset used (*see Methods 2.4.3*) was not tissue specific. This may have diminished the co-expression signal by averaging measurements from experiments in various tissue types, further hindering the analysis. Another validation approach that could be used in the future would be to investigate whether protein co-occurrence or co-expression in different proteomic datasets stored in repositories such as MASSive[108] and PRIDE[109] is correlated with the detection of high-confidence interactions by MAGPIE. While such indirect validation does not take into account transport and secretion pathways, it addresses issues related to transcript regulation.

4.5 Limitations of the Gene Ontology (GO) analysis

Previous studies have proposed that GO semantic similarity scores could be used to score protein-protein interactions to aid in assessing their reliability because proteins interacting together are more likely to share functional annotations than those that do not[110,111]. Therefore, it was hypothesized that protein-protein interaction pairs identified with high confidence by MAGPIE would have higher GO semantic similarity scores than interaction pairs with lesser or no confidence. After observing that this was not the case for our dataset, the GO terms associated to our interaction pairs were investigated. Of note, only four GO terms were associated to the

CNDP1 protein after filtering out ambiguous terms (*see Methods 2.4.2*). Not only is this a small number of GO terms, but the terms themselves are all from the molecular function domain: GO:0004180 (carboxypeptidase activity), GO:0008237 (metallopeptidase activity), GO:0008238 (exopeptidase activity), and GO:0016805 (dipeptidase activity). This suggests that, not only can scores not be computed for the biological process and cellular compartment domains, but it would be very difficult for a putative interactor to earn a high semantic similarity score for the molecular function domain, as the terms are all related to peptidase activity. Moreover, a protein that is annotated with GO terms such as metallopeptidase could very well not interact with another like-protein (i.e., another metallopeptidase), but instead on some target protein, which is likely to have another function. In such a case, we would expect this interaction pair to share a GO term in the biological process or cellular compartment namespace, but not necessarily a molecular function. Two of the five empirical experiments employed antibodies targeting the beta-ala-his dipeptidase (CNDP1) protein (Anti-CNDP1-IgG Ab 1, Anti-CNDP1-IgG Ab 2), meaning there was a larger representation of putative interactions with the CNDP1 bait than the three remaining bait proteins. This further hindered the GO semantic similarity score analysis.

4.6 Future work

Future work to improve MAGPIE would ideally begin with a larger spectral count datasets for reasons described in *Discussion 4.2.1*. In producing this larger dataset, the negative control experiments could be composed of more immunoprecipitation experiments utilizing antibodies targeting synthetic molecular tags, such as a tandem affinity purification (TAP) tag antibody, to minimize affinity for any circulating plasma proteins. Therein, an improved negative control spectral count dataset may be produced for computational modeling of the background

contamination and antibody non-specific binding. A new and larger dataset would also allow for testing of the model's generalizability. Additional facets of MAGPIE's features could then be evaluated. For example, with protein purifications from more empirical experiments (i.e., more than five), there may be a reasonable number of putative protein-protein interactions to label as positive training examples identified by using fold-change as the criterion for identifying likely high-confidence interactions. A comparative analysis could then be conducted, evaluating model performance using spectral count Z-score relative to the controls as a feature versus fold-change. Entirely new features could also be explored. For instance, it is well-understood that interacting proteins often co-localize in cellular compartments[112] and various public repositories make retrieving this information accessible, such as CellMap[14]. A binary feature to indicate whether a pair of putatively interacting proteins are known to co-localize elsewhere in the human body or cell compartments could be added (i.e., 0 = no, 1 = yes). If future experiments to produce a larger dataset includes technical replicates, then the D-score from the CompPASS algorithm could also be included as a feature. It would additionally be interesting to apply the methods of this thesis to other bodily fluids, for which little is known of their protein-protein interaction networks, such as saliva, tears, mucus, and urine. Like plasma, these matrices are not compatible with the use of molecular tags. Therefore, the modified immunoprecipitation approach followed by the computational methods developed in this thesis may serve useful for mapping their interactomes.

Naturally, a future step could also be to explore a new supervised machine learning model. Artificial neural network models and ensemble methods have been proposed for identifying protein-protein interactions in human cell lines, specifically for imbalanced datasets[113]. Because of the nature of a human plasma protein-protein interaction dataset, therein

having significantly fewer bona fide interactions than false-positives, this may also be an avenue worth exploring after exhausting all potential improvements to the logistic regression model.

Altogether, the methodologies developed and MAGPIE are the first of their kind in the pursuit of characterizing the mysteries that lie within human plasma interactome.

References

1. Yates, J. R. Mass spectrometry: from genomics to proteomics. *Trends Genet.* **16**, 5–8 (2000).
2. Tyers, M. & Mann, M. From genomics to proteomics. *Nature* **422**, 193–197 (2003).
3. Vihinen, M. Bioinformatics in proteomics. *Biomol. Eng.* **18**, 241–248 (2001).
4. Blattmann, P. & Aebersold, R. The Advent of Mass Spectrometry-Based Proteomics in Systems Biology Research. in *Encyclopedia of Cell Biology* 166–176 (Elsevier, 2016). doi:10.1016/B978-0-12-394447-4.40030-1
5. Petricoin, E. F., Zoon, K. C., Kohn, E. C., Barrett, J. C. & Liotta, L. A. Clinical proteomics: translating benchside promise into bedside reality. *Nat. Rev. Drug Discov.* **1**, 683–695 (2002).
6. Pankow, S. *et al.* ΔF508 CFTR interactome remodelling promotes rescue of cystic fibrosis. *Nature* **528**, 510–516 (2015).
7. Macklin, A., Khan, S. & Kislinger, T. Recent advances in mass spectrometry based clinical proteomics: applications to cancer research. *Clin. Proteomics* **17**, 17 (2020).
8. Schwikowski, B., Uetz, P. & Fields, S. A network of protein–protein interactions in yeast. *Nat. Biotechnol.* **18**, 1257–1261 (2000).
9. Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**, 637–643 (2006).
10. Pitre, S. *et al.* Global investigation of protein–protein interactions in yeast Saccharomyces cerevisiae using re-occurring short polypeptide sequences. *Nucleic Acids Res.* **36**, 4286–4294 (2008).

11. Breitkreutz, A. *et al.* A Global Protein Kinase and Phosphatase Interaction Network in Yeast. *Science (80-.).* **328**, 1043–1046 (2010).

12. Simonis, N. *et al.* Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. *Nat. Methods* **6**, 47–54 (2009).

13. Forget, D. *et al.* The Protein Interaction Network of the Human Transcription Machinery Reveals a Role for the Conserved GTPase RPAP4/GPN1 and Microtubule Assembly in Nuclear Import and Biogenesis of RNA Polymerase II. *Mol. Cell. Proteomics* **9**, 2827–2839 (2010).

14. Go, C. D. *et al.* A proximity-dependent biotinylation map of a human cell. *Nature* **595**, 120–124 (2021).

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

16. Minton, K. Strength in numbers. *Nat. Rev. Mol. Cell Biol.* **16**, 702–703 (2015).

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

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

19. Rossi, F., Charlton, C. A. & Blau, H. M. Monitoring protein-protein interactions in intact eukaryotic cells by -galactosidase complementation. *Proc. Natl. Acad. Sci.* **94**, 8405–8410 (1997).

20. Remy, I. & Michnick, S. W. Clonal selection and in vivo quantitation of protein interactions with protein-fragment complementation assays. *Proc. Natl. Acad. Sci.* **96**, 5394–5399 (1999).

21. Wong, J. H. *et al.* A yeast two-hybrid system for the screening and characterization of small-molecule inhibitors of protein–protein interactions identifies a novel putative Mdm2-binding site in p53. *BMC Biol.* **15**, 108 (2017).

22. Jia, Y., Kowalski, P. & Lopez, I. Using yeast two-hybrid system and molecular dynamics simulation to detect venom protein-protein interactions. *Curr. Res. Toxicol.* **2**, 93–98 (2021).

23. Hollingsworth, R. & White, J. H. Target discovery using the yeast two-hybrid system. *Drug Discov. Today TARGETS* **3**, 97–103 (2004).

24. Stynen, B., Tournu, H., Tavernier, J. & Van Dijck, P. Diversity in Genetic In Vivo Methods for Protein-Protein Interaction Studies: from the Yeast Two-Hybrid System to the Mammalian Split-Luciferase System. *Microbiol. Mol. Biol. Rev.* **76**, 331–382 (2012).

25. Yates, J. R. Mass spectrometry and the age of the proteome. *J. Mass Spectrom.* **33**, 1–19 (1998).

26. Domon, B. & Aebersold, R. Mass Spectrometry and Protein Analysis. *Science (80-.).* **312**, 212–217 (2006).

27. Vasilescu, J. & Figeys, D. Mapping protein–protein interactions by mass spectrometry. *Curr. Opin. Biotechnol.* **17**, 394–399 (2006).

28. Pitt, J. J. Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. *Clin. Biochem. Rev.* **30**, 19–34 (2009).

29. Pelletier, A. R. *et al.* MealTime-MS: A Machine Learning-Guided Real-Time Mass Spectrometry Analysis for Protein Identification and Efficient Dynamic Exclusion. *J. Am. Soc. Mass Spectrom.* **31**, 1459–1472 (2020).

30. Rigaut, G. *et al.* A generic protein purification method for protein complex

characterization and proteome exploration. *Nat. Biotechnol.* **17**, 1030–1032 (1999).

31. Li, Y. Commonly used tag combinations for tandem affinity purification. *Biotechnol. Appl. Biochem.* **55**, 73–83 (2010).

32. Berggård, T., Linse, S. & James, P. Methods for the detection and analysis of protein–protein interactions. *Proteomics* **7**, 2833–2842 (2007).

33. Hopp, T. P. *et al.* A Short Polypeptide Marker Sequence Useful for Recombinant Protein Identification and Purification. *Bio/Technology* **6**, 1204–1210 (1988).

34. Jeronimo, C. *et al.* Systematic Analysis of the Protein Interaction Network for the Human Transcription Machinery Reveals the Identity of the 7SK Capping Enzyme. *Mol. Cell* **27**, 262–274 (2007).

35. van den Berg, D. L. C. *et al.* An Oct4-Centered Protein Interaction Network in Embryonic Stem Cells. *Cell Stem Cell* **6**, 369–381 (2010).

36. Shi, G. & Jin, Y. Role of Oct4 in maintaining and regaining stem cell pluripotency. *Stem Cell Res. Ther.* **1**, 39 (2010).

37. Roux, K. J., Kim, D. I., Burke, B. & May, D. G. BioID: A Screen for Protein-Protein Interactions. *Curr. Protoc. Protein Sci.* **91**, (2018).

38. Young, M. M. *et al.* High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proc. Natl. Acad. Sci.* **97**, 5802–5806 (2000).

39. Leitner, A., Faini, M., Stengel, F. & Aebersold, R. Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. *Trends Biochem. Sci.* **41**, 20–32 (2016).

40. Leitner, A. *et al.* Probing Native Protein Structures by Chemical Cross-linking, Mass

Spectrometry, and Bioinformatics. *Mol. Cell. Proteomics* **9**, 1634–1649 (2010).

41. Richards, A. L., Eckhardt, M. & Krogan, N. J. Mass spectrometry-based protein–protein interaction networks for the study of human diseases. *Mol. Syst. Biol.* **17**, (2021).

42. Liu, Q. *et al.* A proximity-tagging system to identify membrane protein–protein interactions. *Nat. Methods* **15**, 715–722 (2018).

43. Ho, C. S. *et al.* Electrospray ionisation mass spectrometry: principles and clinical applications. *Clin. Biochem. Rev.* **24**, 3–12 (2003).

44. Marquioni, V., Nunes, F. M. F. & Novo-Mansur, M. T. M. Protein Identification by Database Searching of Mass Spectrometry Data in the Teaching of Proteomics. *J. Chem. Educ.* **98**, 812–823 (2021).

45. Eng, J. K., McCormack, A. L. & Yates, J. R. 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 (1994).

46. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* **389**, 1017–1031 (2007).

47. Pavelka, N. *et al.* Statistical Similarities between Transcriptomics and Quantitative Shotgun Proteomics Data. *Mol. Cell. Proteomics* **7**, 631–644 (2008).

48. Mahboob, S. *et al.* Is isolation of comprehensive human plasma peptidomes an achievable quest? *J. Proteomics* **127**, 300–309 (2015).

49. Keller, B. O., Sui, J., Young, A. B. & Whittal, R. M. Interferences and contaminants encountered in modern mass spectrometry. *Anal. Chim. Acta* **627**, 71–81 (2008).

50. Mellacheruvu, D. *et al.* The CRAPome: a contaminant repository for affinity purification–

mass spectrometry data. *Nat. Methods* **10**, 730–736 (2013).

51. Sowa, M. E., Bennett, E. J., Gygi, S. P. & Harper, J. W. Defining the Human Deubiquitinating Enzyme Interaction Landscape. *Cell* **138**, 389–403 (2009).

52. Choi, H. *et al.* SAINT: probabilistic scoring of affinity purification–mass spectrometry data. *Nat. Methods* **8**, 70–73 (2011).

53. Lavallée-Adam, M., Cloutier, P., Coulombe, B. & Blanchette, M. Modeling Contaminants in AP-MS/MS Experiments. *J. Proteome Res.* **10**, 886–895 (2011).

54. Altelaar, A. F. M., Munoz, J. & Heck, A. J. R. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* **14**, 35–48 (2013).

55. Sardiu, M. E. *et al.* Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. *Proc. Natl. Acad. Sci.* **105**, 1454–1459 (2008).

56. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–67 (1999).

57. Cloutier, P. *et al.* High-resolution mapping of the protein interaction network for the human transcription machinery and affinity purification of RNA polymerase II-associated complexes. *Methods* **48**, 381–386 (2009).

58. Tian, B., Zhao, C., Gu, F. & He, Z. A two-step framework for inferring direct protein-protein interaction network from AP-MS data. *BMC Syst. Biol.* **11**, 82 (2017).

59. Gauthier, M.-S. *et al.* A semi-automated mass spectrometric immunoassay coupled to selected reaction monitoring (MSIA-SRM) reveals novel relationships between circulating PCSK9 and metabolic phenotypes in patient cohorts. *Methods* **81**, 66–73 (2015).

60. Hanash, S. M., Pitteri, S. J. & Faca, V. M. Mining the plasma proteome for cancer

biomarkers. *Nature* **452**, 571–579 (2008).

61. Caselli, C. *et al.* Association of PCSK9 plasma levels with metabolic patterns and coronary atherosclerosis in patients with stable angina. *Cardiovasc. Diabetol.* **18**, 144 (2019).

62. Yamazaki, Y., Zhao, N., Caulfield, T. R., Liu, C.-C. & Bu, G. Apolipoprotein E and Alzheimer disease: pathobiology and targeting strategies. *Nat. Rev. Neurol.* **15**, 501–518 (2019).

63. Corder, E. *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science (80-.).* **261**, 921–923 (1993).

64. Nielsen, H. M. *et al.* Peripheral apoE isoform levels in cognitively normal APOE ε3/ε4 individuals are associated with regional gray matter volume and cerebral glucose metabolism. *Alzheimers. Res. Ther.* **9**, 5 (2017).

65. Martínez-Morillo, E. *et al.* Total apolipoprotein E levels and specific isoform composition in cerebrospinal fluid and plasma from Alzheimer's disease patients and controls. *Acta Neuropathol.* **127**, 633–643 (2014).

66. Leeman, M., Choi, J., Hansson, S., Storm, M. U. & Nilsson, L. Proteins and antibodies in serum, plasma, and whole blood—size characterization using asymmetrical flow field-flow fractionation (AF4). *Anal. Bioanal. Chem.* **410**, 4867–4873 (2018).

67. Pernemalm, M. *et al.* In-depth human plasma proteome analysis captures tissue proteins and transfer of protein variants across the placenta. *Elife* **8**, (2019).

68. Seong, Y., Yoo, Y. S., Akter, H. & Kang, M.-J. Sample preparation for detection of low abundance proteins in human plasma using ultra-high performance liquid chromatography coupled with highly accurate mass spectrometry. *J. Chromatogr. B* **1060**, 272–280 (2017).

69. Pisanu, S., Biosa, G., Carcangiu, L., Uzzau, S. & Pagnozzi, D. Comparative evaluation of seven commercial products for human serum enrichment/depletion by shotgun proteomics. *Talanta* **185**, 213–220 (2018).

70. van der Vusse, G. J. Albumin as Fatty Acid Transporter. *Drug Metab. Pharmacokinet.* **24**, 300–307 (2009).

71. Baker, M. E. Albumin's role in steroid hormone action and the origins of vertebrates: is albumin an essential protein? *FEBS Lett.* **439**, 9–12 (1998).

72. Mahendhar, R. *et al.* Effect of Albumin Polymorphism on Thyroid Hormones: A Case Report and Literature Review. *Cureus* (2018). doi:10.7759/cureus.2903

73. Merlot, A. M., Kalinowski, D. S. & Richardson, D. R. Unraveling the mysteries of serum albuminâ€"more than just a serum protein. *Front. Physiol.* **5**, (2014).

74. Burkhart, J. M. *et al.* The first comprehensive and quantitative analysis of human platelet protein composition allows the comparative analysis of structural and functional pathways. *Blood* **120**, e73–e82 (2012).

75. Fredolini, C. *et al.* Systematic assessment of antibody selectivity in plasma based on a resource of enrichment profiles. *Sci. Rep.* **9**, 8324 (2019).

76. Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **58**, 236 (1963).

77. Krznaric, D. & Levcopoulos, C. Fast Algorithms for Complete Linkage Clustering. *Discrete Comput. Geom.* **19**, 131–145 (1998).

78. Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **2**, 37–52 (1987).

79. Ivosev, G., Burton, L. & Bonner, R. Dimensionality Reduction and Visualization in

Principal Component Analysis. *Anal. Chem.* **80**, 4933–4944 (2008).

80. Zhou, B. *et al.* Plasma proteomics-based identification of novel biomarkers in early gastric cancer. *Clin. Biochem.* **76**, 5–10 (2020).

81. Miller, J. W., Goodman, R. & Smyth, P. On loss functions which minimize to conditional expected values and posterior probabilities. *IEEE Trans. Inf. Theory* **39**, 1404–1408 (1993).

82. Moore, A. W. & Lee, M. S. Efficient Algorithms for Minimizing Cross Validation Error. in *Machine Learning Proceedings 1994* 190–198 (Elsevier, 1994). doi:10.1016/B978-1-55860-335-6.50031-3

83. Rodriguez, J. D., Perez, A. & Lozano, J. A. Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 569–575 (2010).

84. Gauthier, M. S. *et al.* A semi-automated mass spectrometric immunoassay coupled to selected reaction monitoring (MSIA-SRM) reveals novel relationships between circulating PCSK9 and metabolic phenotypes in patient cohorts. *Methods* **81**, 66–73 (2015).

85. Gauthier, M. S. *et al.* Posttranslational modification of proprotein convertase subtilisin/kexin type 9 is differentially regulated in response to distinct cardiometabolic treatments as revealed by targeted proteomics. *J. Clin. Lipidol.* **12**, 1027–1038 (2018).

86. Perkins, D. N., Pappin, D. J. C., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).

87. Searle, B. C. Scaffold: A bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics* **10**, 1265–1269 (2010).

88. Suzuki, R. & Shimodaira, H. Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**, 1540–1542 (2006).

89. Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V., and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P., and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and & Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

90. Jensen, L. J. *et al.* STRING 8 - A global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* **37**, 412–416 (2009).

91. Paul Shannon, 1 *et al.* Cytoscape: A Software Environment for Integrated Models. *Genome Res.* **13**, 426 (1971).

92. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).

93. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–9 (2000).

94. Zhao, C. & Wang, Z. GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms. *Sci. Rep.* **8**, 15107 (2018).

95. Obayashi, T., Kagaya, Y., Aoki, Y., Tadaka, S. & Kinoshita, K. COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. *Nucleic Acids Res.* **47**, D55–D62 (2019).

96. Obayashi, Takeshi, & Kinoshita, K. Gene Coexpression Data for Human (Version COXPRESdb ver 7.3) [Dataset]. *Zenodo* (2019).

97. Saraswat, M. *et al.* Label-free serum proteomics and multivariate data analysis identifies

biomarkers and expression trends that differentiate Intraductal papillary mucinous neoplasia from pancreatic adenocarcinoma and healthy controls. *Transl. Med. Commun.* **4**, 6 (2019).

98. Poulos, R. C. *et al.* Strategies to enable large-scale proteomics for reproducible research. *Nat. Commun.* **11**, 3793 (2020).

99. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

100. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* (2018).

101. Meyer, J. G., Niemi, N. M., Pagliarini, D. J. & Coon, J. J. Quantitative shotgun proteome analysis by direct infusion. *Nat. Methods* **17**, 1222–1228 (2020).

102. Chen, X. *et al.* Host Immune Response to Influenza A Virus Infection. *Front. Immunol.* **9**, (2018).

103. Krammer, F. The human antibody response to influenza A virus infection and vaccination. *Nat. Rev. Immunol.* **19**, 383–397 (2019).

104. Espadaler, J., Romero-Isart, O., Jackson, R. M. & Oliva, B. Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships. *Bioinformatics* **21**, 3360–3368 (2005).

105. Tirosh, I. & Barkai, N. Computational verification of protein-protein interactions by orthologous co-expression. *BMC Bioinformatics* **6**, (2005).

106. De Bodt, S., Proost, S., Vandepoele, K., Rouzé, P. & Van de Peer, Y. Predicting protein-protein interactions in Arabidopsis thaliana through integration of orthology, gene ontology and co-expression. *BMC Genomics* **10**, 288 (2009).

107. Gry, M. *et al.* Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics* **10**, 365 (2009).

108. Wang, M. *et al.* Assembling the Community-Scale Discoverable Human Proteome. *Cell Syst.* **7**, 412-421.e5 (2018).

109. Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).

110. Jain, S. & Bader, G. D. An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC Bioinformatics* **11**, 562 (2010).

111. Zhang, S.-B. & Tang, Q.-R. Protein–protein interaction inference based on semantic similarity of Gene Ontology terms. *J. Theor. Biol.* **401**, 30–37 (2016).

112. Kuriyan, J. & Eisenberg, D. The origin of protein interactions and allostery in colocalization. *Nature* **450**, 983–990 (2007).

113. Zhang, Y. *et al.* Using ensemble methods to deal with imbalanced data in predicting protein–protein interactions. *Comput. Biol. Chem.* **36**, 36–41 (2012).