Journal of proteome research

pubs.acs.org/jpr Article

MAGPIE: A Machine Learning Approach to Decipher Protein–Protein Interactions in Human Plasma

Published as part of Journal of Proteome Research special issue "Canadian Proteomics".

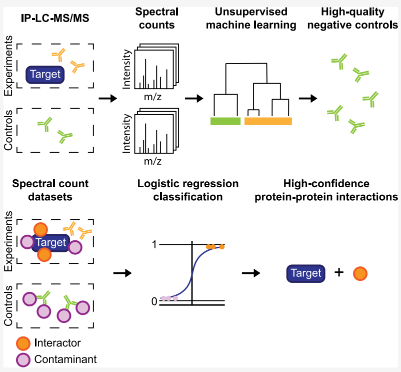
Emily Hashimoto-Roth, Diane Forget, Vanessa P. Gaspar, Steffany A. L. Bennett, Marie-Soleil Gauthier, Benoit Coulombe, and Mathieu Lavallée-Adam*

Cite This: J. Proteome Res. 2025, 24, 383–396 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

Downloaded via UNIV OF OTTAWA on May 21, 2025 at 04:03:13 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

ABSTRACT: Immunoprecipitation coupled to tandem mass spectrometry (IP-MS/MS) methods are often used to identify protein–protein interactions (PPIs). While these approaches are prone to false positive identifications through contamination and antibody nonspecific binding, their results can be filtered using negative controls and computational modeling. However, such filtering does not effectively detect false-positive interactions when IP-MS/MS is performed on human plasma samples. Therein, proteins cannot be overexpressed or inhibited, and existing modeling 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/MS, which leverages negative controls that include antibodies targeting proteins not expected to be present in human plasma. A set of negative controls used for false positive interaction modeling is first constructed. MAGPIE then assesses the reliability of PPIs detected in IP-MS/MS experiments using antibodies that target known plasma proteins. When applied to five IP-MS/MS experiments as a proof of concept, our algorithm identified 68 PPIs with an FDR of 20.77%. MAGPIE significantly outperformed a state-of-the-art PPI discovery tool and identified known and predicted PPIs. Our approach provides an unprecedented ability to detect human plasma PPIs, which enables a better understanding of biological processes in plasma.



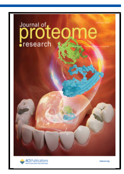
KEYWORDS: *proteomics, plasma, protein–protein Interactions, machine learning, mass spectrometry, supervised learning, immunoprecipitation, affinity purification, antibody, artificial intelligence*

INTRODUCTION

Characterizing protein–protein interactions can reveal much about the function of proteins, the complexes in which they form, and the biological processes in which they are involved. To this end, groups have mapped large-scale protein–protein interaction (PPI) networks in yeast (sp. *Saccharomyces cerevisiae*).[1−3] For example, both a protein kinase and phosphatase interactome have been mapped.[4] Similarly, the PPI network in nematodes (sp. *Caenorhabditis elegans*) has also been largely mapped.[5] In humans, a proximity-dependent interaction network, named Cell Map, has been made available.[6] Quantitative proteomics has also been used to define the interactome network topology of HeLa cell lines, proposing the importance of interaction stoichiometry for creating complete networks.[7,8] Finally, the BioPlex protein–protein interaction network is one of the largest networks for interactions characterized in humans, composed of 118,162 interactions from 14,586 at the time of publication.[9,10]

Nowadays, tandem mass spectrometry (MS/MS) is typically coupled to different strategies for screening PPIs. Among these strategies, we note immunoprecipitation (IP) coupled to tandem mass spectrometry (IP-MS/MS).[11] Therein, the tandem affinity purification (TAP)[12,13] and the FLAG tags[14] are commonly used to isolate an affinity-tagged protein of interest (often referred to as the bait) with its interactors (preys). Further progress in the development of IP-MS/MS systems has spawned proximity labeling methods, such as BioID,[15] for the identification of proximal proteins. Another useful approach for detecting direct PPIs is cross-linking (XL) coupled to MS/MS (XL-MS/MS).[16,17] With this technique, interacting proteins are covalently bound prior to enzymatic digestion. The cross-linked peptides can then be enriched, and the resulting data provides information not only about the direct physical contact of the interacting proteins but also insights into the structural biology of these interaction pairs.[18]

Received: February 29, 2024
Revised: September 2, 2024
Accepted: October 29, 2024
Published: January 7, 2025



© 2025 American Chemical Society

https://doi.org/10.1021/acs.jproteome.4c00160
J. Proteome Res. 2025, 24, 383–396

The choice of purification system influences experimental design and these methods can be further optimized to target historically difficult proteins to characterize, such as transmembrane proteins.[19]

Both the immunoprecipitation strategy and mass spectrometry portions of IP-MS/MS generate false-positive protein−protein interaction identifications, which can be filtered by computational approaches. False positives arising from MS/MS peptide identifications are typically controlled using a target-decoy database search that estimates an identification false discovery rate (FDR).[20] Much like any mass spectrometry experiments, IP experiments are susceptible to contamination by exogenous proteins artifactually introduced to samples, such as keratins, bovine serum albumin, Protein A, which includes the IgG binding domain of the TAP tag, or the tryptic enzymes themselves used for digestion during MS/MS sample preparation.[21] These contaminating proteins are well-characterized in public repositories, such as the contaminant repository for affinity purification-mass spectrometry data (CRAPome),[22] and thus can be easily excluded from IP-MS/MS results. A more confounding source of false-positive identifications stems from antibody nonspecific binding of proteins to the antibody of the IP system or the molecular tag expressed by the protein of interest. This issue becomes more prominent as the biological matrix becomes more complex. While these nonspecific protein bindings are often correct mass spectrometry identifications, they are unlikely to be biologically relevant interactors of the protein of interest. These false positive PPIs have been shown to be effectively filtered out by combining the use of experimental negative controls and computational modeling.[23−26] Examples of negative controls for an IP experiment would be to perform the antibody purification in a system not expressing the affinity tag, expressing the affinity tag alone without bait, or expressing a protein foreign to (or irrelevant to) the organism fused with the affinity tag.[27] It is assumed that proteins purified in these negative controls are examples of antibody nonspecific binding.

To filter nonspecific binding, computational approaches for IP-MS generally assess the confidence of a putative PPI, such that it is deemed to be a bona fide interactor if its prey is present at a significantly higher level than in negative controls. This logic therefore emphasizes the importance of choosing negative controls that are highly unlikely to result in the purification of bona fide interactions and that are only composed of nonspecific bindings. Many of these approaches use label-free quantification MS/MS data to assign a confidence score to a successfully purified protein−protein interaction,[28] either using precursor ion intensity or spectral count as quantification measures. Among the prominent algorithms to identify bona fide protein−protein interactions from MS/MS data, we note the Significance Analysis of INTeractome (SAINT) algorithm.[29] SAINT uses a mixture model and either precursor intensity or spectral count to estimate the posterior probability that a putative interactor is true and uses these probabilities to estimate an FDR for the unique identifications. The computation of posterior probabilities for putative bait−prey pairs and the implementation of Bayesian inference to identify bona fide interactions have been used in multiple algorithms.[30] Decontaminator uses Mascot database search engine scores[31] to assess the confidence of putative bait−prey pairs.[32] CompPASS[33] considers Z-scores of spectral count data and novel D-scores to assess the uniqueness of a putative interactor for a given bait and its reproducibility across biological replicates. Finally, the MasterMap system takes advantage of changes in precursor peptide intensity across sequentially diluted samples containing a given protein of interest.[34] PPI confidence can be also evaluated through the use of network topology.[35] The latter approaches stipulate that a given interaction pair can be supported by other biologically relevant interactions in a given organism.

Despite these advances, assessing PPI confidence in human plasma remains challenging. Many systemic molecular pathways observed in circulating PPIs remain uncharacterized due to a lack of the necessary experimental and computational approaches for reliable PPI identifications. For instance, proprotein convertase subtilisin/kexin type 9 serine protease (PCSK9) is present at varying levels in human plasma and plays an important role in hypercholesterolemia and various other cardiovascular disease phenotypes.[36,37] One major challenge is the modeling of background contamination in human plasma. Since affinity tagging of bait proteins with techniques such as TAP or FLAG is impossible, the ability to generate representative negative control experiments, as in standard IP-MS/MS experiments, is minimal. While difficult, attempts have nevertheless been made in the past decade to confidently identify PPIs in human plasma. A 2019 study attempted to assess the selectivity of antibodies for their target by directly immunoprecipitating proteins with their respective antibodies and systematically evaluating each antibody's enrichment for its target, by computing Z-scores of the label-free quantification (LFQ) intensities of each purification.[38] There, the authors considered an antibody to be enriched for its target if the quantified protein obtained a Z-score ≥3. One drawback of this approach is that such a Z-score threshold is arbitrary. Besides this approach, current tools for filtering out false-positive protein−protein interactions from IP-MS/MS data are designed only for experiments whose negative controls can be easily produced and cannot be readily applied in the context of human plasma analysis.

To address this challenge, we present MAGPIE, a Machine learning Assessment with loGistic regression of Protein−protein IntEractions, to address the challenge of assessing PPI confidence in human plasma. MAGPIE uses a novel two-phase computational approach for discriminating between putative PPIs and antibody nonspecific binding in human plasma. The first phase identifies a set of experimental negative controls to use for modeling contamination and antibody nonspecific binding. These controls are identified from a set of antibody purifications targeting proteins not expected, with high confidence, to be present in human plasma. Proteins identified with these antibodies are used to model human plasma nonspecific binding abundance. The second phase is a supervised machine learning algorithm that predicts whether a putative PPI—detected in an antibody purification targeting a protein expected to be present in human plasma—is biologically relevant or the result of nonspecific binding. We show here in a proof of concept that MAGPIE outperforms a state-of-the-art PPI confidence assessment software package, SAINT, and that it identifies biologically relevant PPIs documented in protein−protein interaction repositories[39] and predicted by AlphaFold 3.[40]

https://doi.org/10.1021/acs.jproteome.4c00160
J. Proteome Res. 2025, 24, 383−396

EXPERIMENTAL METHODS

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.[36,41] Antibody information and sources are provided in Table S1. Iodoacetamide, DTT, and glucagon were from Sigma, sequencing grade modified trypsin was from Promega, and HBS-EP buffer was from GE Healthcare. HPLC-grade water, trifluoroacetic acid, and acetonitrile were purchased from Fisher.

Protein Affinity Capture

Commercially acquired pooled human plasma samples from 100 individuals with different sexes, ages, and ethnicities preserved in EDTA (250 μL) (Zenbio) were diluted with 175 μL of HBS-EP buffer and dispensed in 96-well robotic PCR plates (Abgene). Protein affinity capture was automated by 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 × 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 −20 °C until digestion.

Sample Preparation for LC-MS/MS

Samples were reconstituted in 45 μL of 4 M 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 37 °C for 30 min on a ThermoMixer (Thermo Fisher Scientific) set at 350 rpm. After cooling at room temperature for 5 min, 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 a 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 20 h on a ThermoMixer set at 350 rpm and 37 °C, 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 peptide carrier.

LC-MS/MS Analysis

Peptide samples were loaded into a 75 μm i.d. × 150 mm Self-Pack C18 column with 5 μm particles 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% over 100 min and then from 40 to 85% B over 10 min. The HPLC system was coupled to an Orbitrap Fusion mass spectrometer (Thermo Scientific) via 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 and a target value of 3×10^5. The 25 most intense ions were fragmented in the HCD collision cell and analyzed in the linear ion trap with a target value of 2×10^4 and normalized collision energy at 28. Target ions selected for fragmentation were dynamically excluded for 20 s.

Computational Identification of Proteins

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 identification was performed by searching the UniProtKB/SwissProt human protein sequence database (downloaded in July 2016)[42] with Mascot 2.6 (Matrix Science).[31] The mass tolerances for precursor and fragment ions were set to 10 ppm and 0.6 Da, respectively. Tryptic peptides, allowing up to two missed cleavages, were searched by the algorithm. Cysteine carbamidomethylation was specified as a fixed modification and methionine oxidation as a variable modification. Data interpretation was performed using Scaffold (version 4.8),[43] Mascot, and Qual Browser (Xcalibur, Thermo Fischer Scientific), and protein identifications were reported at an FDR of <1%.

IP-MS/MS Data Sets Details and Availability

The immunoprecipitation experiments were performed using 12 antibodies, targeting 10 proteins, and copurifying a total of 226 unique plasma proteins (protein identification FDR < 1%). 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 (CNDP1 (×2) (CNDP1_HUMAN), KLK6 (KLK6_HUMAN), SNCA (SYUA_HUMAN), PCSK9 (PCSK9_HUMAN)). These proteins were selected based on the availability of good quality antibodies and to represent a mix of proteins that were fairly well characterized and others for which the interactions in plasma were largely unknown. Conversely, the remaining seven immunoprecipitation experiments were performed using antibodies targeting proteins not expected to be present in human plasma (FLAG, HA, LC3B, METTL23, RPAP2). These tentative experimental negative controls were initially assessed to determine if they confidently produced useable examples of antibody nonspecific binding. The resulting data were then used for the construction of a machine-learning classifier. The MS/MS-based proteomics data have been deposited to the ProteomeXchange Consortium[44] via the PRIDE[45] partner repository with the data set identifier PXD050230.

Identifying a Set of Experimental Negative Controls Using Unsupervised Learning

MS/MS data sets were analyzed to evaluate whether a subset of experimental negative controls captured a population of antibody nonspecific binding partners. The rationale is that nonspecific bindings are more likely to occur with abundant plasma proteins and that such bindings should be somewhat similar for different antibodies. Conversely, an antibody that uniquely purifies a set of nonspecific proteins would not be useful to model contamination events that can occur in empirical experiments. Controls showing a similar set of purified proteins could, therefore, be used to build a model of nonspecific binding. The protein spectral counts in each IP-MS/MS experiment were normalized. Specifically, the spectral count of a protein p, was normalized against the total spectral count for all proteins purified in a given IP-MS/MS experiment b.

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

where $x_{b,p}$ is the spectral count of p when purified in b, $s_{b,p}$ is its normalized value and x_b is the spectral count sum of all purified proteins in i.

Bootstrapped hierarchical clustering was performed using the complete linkage algorithm and Euclidean distance to investigate the similarity between proteins purified by the different antibodies, while assessing the robustness of the resulting clusters. This analysis was implemented in R (version 3.6.1), using the *pvclust* package (version 2.2.0).[46] Principal component analysis (PCA) was used to further confirm this clustering analysis. PCA was implemented in Python (version 3.7), using the Scikit-learn package (version 0.24.0).[47]

Discriminating Bona Fide PPIs from Antibody Nonspecific Binding Using MAGPIE

We developed MAGPIE, a supervised learning algorithm, which is trained to output the probability that a given putative PPI is a true one given a subset of negative controls. Most proteins were not detected in all negative controls; however, the nonidentification of a protein by MS/MS does not equate to its absence in the analyzed sample. For any protein identified in at least one negative control or IP-MS/MS experiment of known plasma proteins, spectral pseudocount values were attributed to the same protein in negative controls when it failed to be detected. This provided a conservative assessment of the PPI reliability. Pseudocounts were assigned by uniformly and randomly sampling a value from the bottom 10% of nonzero spectral count values belonging to the respective negative control.

Training Set and Classifying Features. MAGPIE is implemented as a logistic regression classifier for the binary classification of the putative PPIs. The negative training set was built from proteins detected in the negative controls. There are, however, no clear-cut criteria for building a positive training set. Nevertheless, Fredolini et al. showed that a Z-score can represent a good selection criterion for true PPIs in plasma.[38] For each purified protein p in an IP-MS/MS experiment targeting a plasma bait protein b, MAGPIE therefore calculates a Z-score $z_{b,\,p}$ as follows:

$$\text{Z-score}_{b,p} = \frac{s_{b,p} - \mu_p}{\sigma_p}$$

where $s_{b,\,p}$ is the normalized spectral count of p, purified in IP-MS/MS experiment b, μ_p is the mean normalized spectral count of p across all controls, and σ_p is the standard deviation of the normalized spectral count of p across all controls. Fold-change, $fc_{b,\,p}$, of a putative interacting protein, p, was calculated as follows,

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

To train this classifier, the criterion of a Z-score greater than or equal to 3 defined likely high-confidence protein−protein interactions, which in turn were used as our positive training examples. Since there were many more negative examples than positive examples, a subset of negative examples was randomly sampled to create a 1:1 class-balanced training set. Normalized spectral count data were then mined to generate the classifying features. These included the normalized spectral count of the protein in the IP-MS/MS experiment, the average, standard deviation, and maximum normalized spectral count of the protein in the negative control experiments, and the normalized spectral count fold-change. The features are listed and further defined in Table 1.

Table 1. 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.

Logistic Regression Classifier to Detect Bona Fide PPIs. MAGPIE implements a logistic regression model for classification, which was trained to output the probability that a given putative protein−protein interaction constitutes a true interaction. The logistic regression model was implemented using the Scikit-learn package (version 0.24.0), and its hyperparameters were optimized to minimize our model's false discovery rate as much as possible. 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" to apply the Newton conjugate gradient algorithm for optimization.

Evaluating MAGPIE's Performance for Detecting Bona Fide PPIs. Due to the lack of existing ground truth knowledge about the PPIs purified in our human plasma study, true positive PPIs cannot be easily estimated. However, false-positive PPIs can be estimated from the set of proteins detected in the control experiments. Such proteins would obtain a probability from the logistic regression classifier that would be higher than a specified confidence threshold. This information allows for the estimation of an FDR at a given probability threshold, t (i.e., FDR(t)). Similar to a leave-one-out cross-validation procedure, MAGPIE implements a leave-one-control-out (LOCO) strategy to estimate this FDR. This strategy functions by executing MAGPIE k times, where k is the number of negative control IP-MS/MS experiments. At each iteration, classification features are re-engineered, by omitting the spectral count data of the i^{th} negative control, where $i = 1, ..., k$. In other words, the spectral count average,

https://doi.org/10.1021/acs.jproteome.4c00160
J. Proteome Res. 2025, 24, 383−396

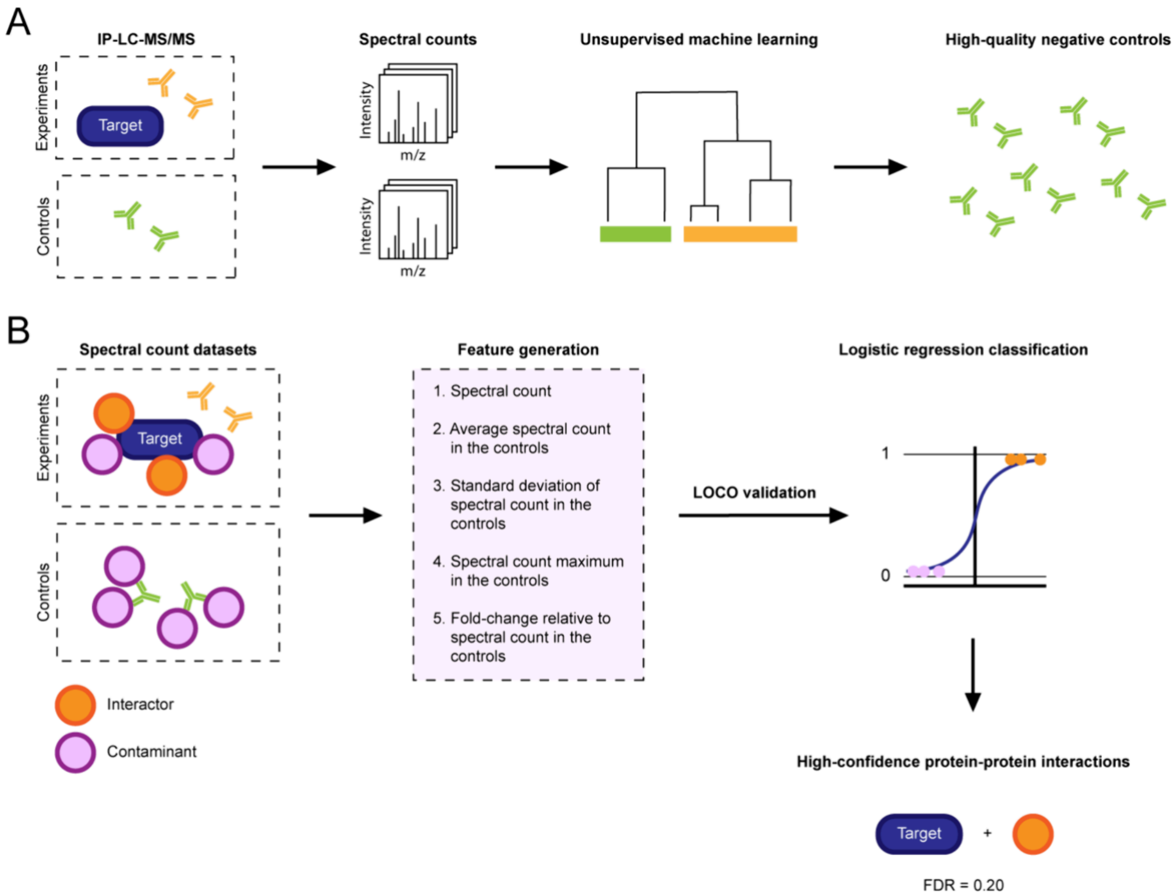


Figure 1. Workflow schematic. (A) The mass spectral count data sets of empirical and negative control IP-MS/MS experiments performed in human plasma samples were analyzed by unsupervised machine learning algorithms (hierarchical clustering and principal component analysis) to identify a set of high-quality negative controls. (B) Features for supervised machine learning were engineered and assigned to all training and testing examples. A logistic regression classifier was constructed and trained to output the probability that a given putative protein−protein interaction was true, from which an FDR is derived.

standard deviation, maximum, fold-change, and Z-scores were recomputed for each LOCO iteration. The purified plasma protein detections belonging to the omitted negative control, of a given iteration, were used as the testing data set for predicting the number of false-positive identifications, as such a protein deemed true by MAGPIE represents nonspecific binding. Hence, for every IP-MS/MS experiment b^C from the set of negative control experiments B^C, MAGPIE's logistic regression computes the probability $prob(p_{b^C})$ for all proteins $p_{b^C} \in P_{b^C}$, which is the set of proteins detected in the control experiment b^C. Similarly, MAGPIC computes $prob(p_{b^E})$ for all proteins $p_{b^E} \in P_{b^E}$, the set of proteins detected in an experiment targeting plasma proteins b^E from the set of all experiments targeting plasma proteins B^E. In other words, FDRs are estimated by computing the fraction of the number of plasma proteins in the left-out controls in each LOCO iteration that obtained a confidence score greater than or equal to the threshold t (i.e., false-positive interactions) against the number of plasma proteins in the actual experiments that obtained a confidence score greater than or equal to t (i.e., true interactions). This allows our method to estimate an FDR for a given value of t as follows:

$$\mathrm{FDR}(t) = \frac{\dfrac{\sum_{b^C \in B^C} \sum_{p_b{}^C \in P_b{}^C} 1_{prob}(p_{b^C}) \geq t}{|B^C \times P_b{}^C|}}{\dfrac{\sum_{b^E \in B^E} \sum_{p_b{}^E \in P_b{}^E} 1_{prob}(p_{b^E}) \geq t}{|B^E \times P_b{}^E|}}$$

where 1_a is an indicator function yielding 1 if a is true and 0 otherwise.

A monotonic transformation is applied to the estimated FDRs when predicting the number of putative protein−protein interactions at a given FDR, calculated as follows,

$$\mathrm{FDR}(t) = min(\mathrm{FDR}(t), \mathrm{FDR}(t + inc))$$

where t is the probability 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, where the numerators and denominators become very small. With the results of the LOCO runs, 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 runs. Finally, because there are two instances of random sampling in each run (spectral pseudocounts data imputation and training set assembly), the robustness of MAGPIE's performance was evaluated by executing it 1000 times, without seed setting, to ascertain more confidence in its results.

https://doi.org/10.1021/acs.jproteome.4c00160
J. Proteome Res. 2025, 24, 383−396

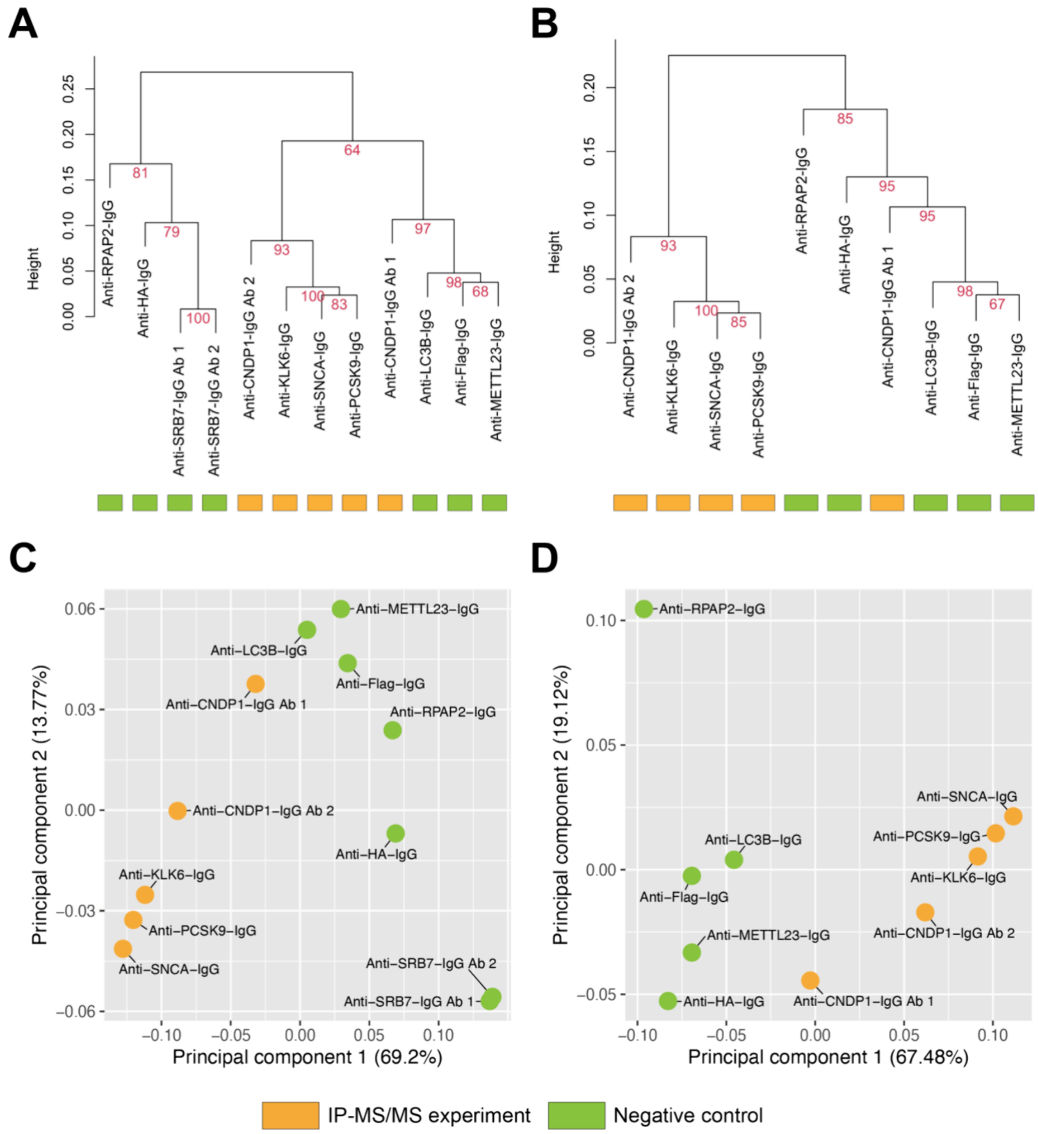


Figure 2. Hierarchical clustering and principal component analysis of IP-MS/MS experiments and negative controls. (A) Complete linkage analysis on the normalized spectral count data prior to excluding the Anti-SRB7-IgG Ab 1 and Anti-SRB7-IgG Ab 2 experiments. (B) Complete linkage analysis on the normalized spectral count data after excluding the Anti-SRB7-IgG Ab 1 and Anti-SRB7-IgG Ab 2 experiments. (A, B) Both analyses were implemented using Euclidean distance and with 10,000 bootstrapping iterations, whose values are denoted by the red numbers under 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 color-coded at the bottom of each leaf. (C) 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. (D) Principal component analysis on the normalized spectral count data after excluding the Anti-SRB7-IgG Ab 1 and Anti-SRB7-IgG Ab 2 experiments. (C, D) Variance in the data explained by the given principal component is indicated in the axis labels.

Benchmarking MAGPIE against SAINT

MAGPIE was benchmarked against SAINT, a leading algorithm for assessing the confidence of putative protein−protein interactions. To prepare input data for SAINT, UniProt Swiss-Prot and TrEMBL (both accessed May 20, 2020) were locally downloaded to retrieve protein sequence lengths, which are necessary in SAINT's input. To execute SAINT, the *saint-spc-ctrl* program was executed with recommended parameters *nburn* = 2000, *niter* = 10,000 (both Gibbs sampling parameters), *lowMode* = 1, *minFold* = 1, and *normalize* = 0. SAINT input spectral count data was imputed using pseudocounts using the same method MAGPIE's input data was imputed. To perform a fair comparison, SAINT was run while implementing the same LOCO strategy as MAGPIE. In other words, SAINT was run 5 times, wherein 5 is the number of negative control IP-MS/MS experiments. In each run, the spectral count data of a different 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 estimation and monotonic transformation were applied, though using the SAINT probability as the threshold, *t*, instead of the logistic regression probability. The results of MAGPIE and SAINT could therefore be fairly compared. This FDR value is different from the BFDR derived from SAINT, which could not be used here to perform a fair comparison.

https://doi.org/10.1021/acs.jproteome.4c00160
J. Proteome Res. 2025, 24, 383−396

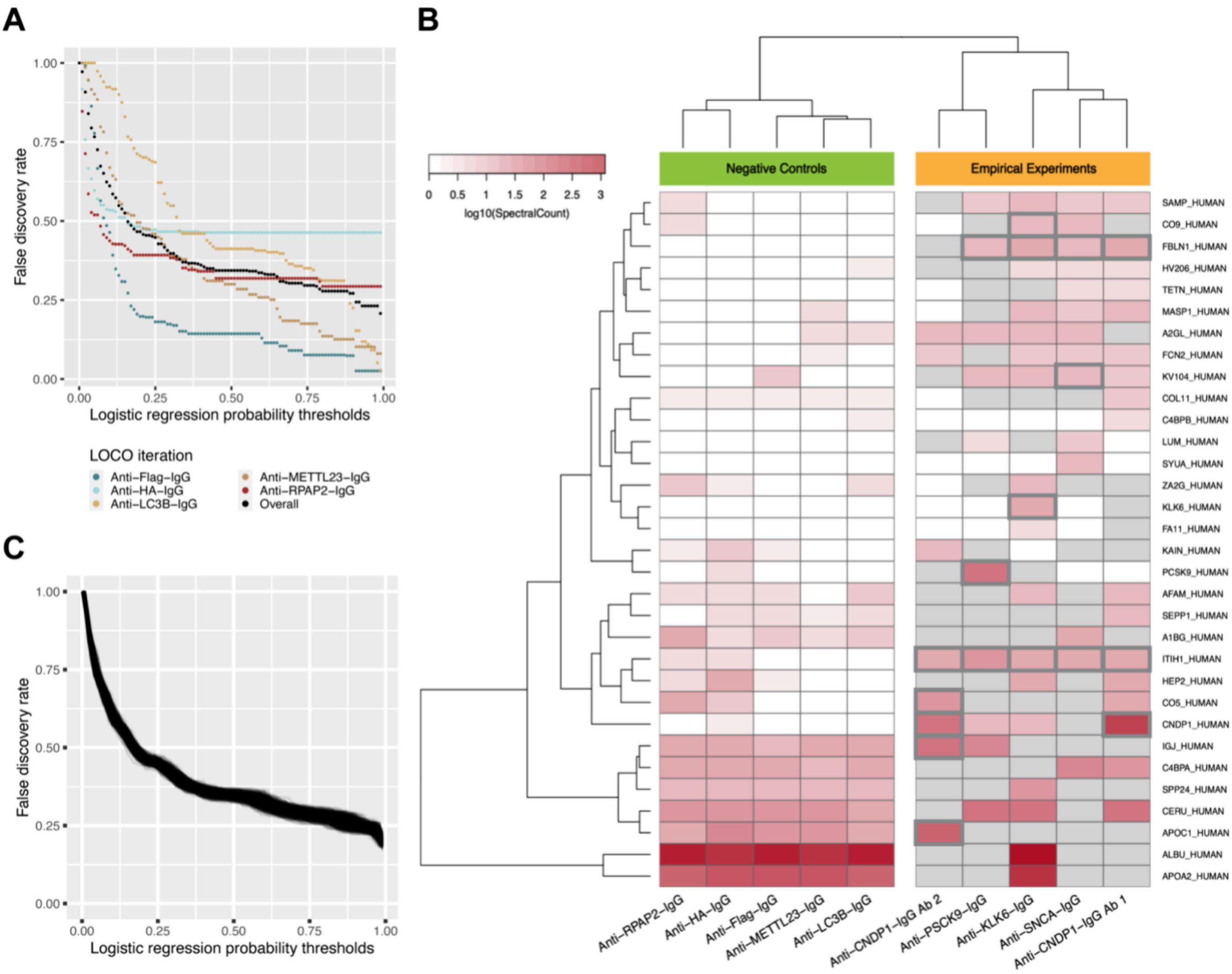


Figure 3. MAGPIE classifying performance evaluation and the 68 PPIs it deemed high confidence. (A) FDRs estimated at logistic regression probability thresholds for each of the five leave-one-control-out cross-validation runs corresponding to a given negative control experiment. An overall FDR, derived from the results of the five leave-one-control-out runs, is superimposed. (B) Heatmap of the log10-transformed spectral count data belonging to the 68 high-confidence PPIs across all IP-MS/MS experiments and negative controls. Comparison of spectral count abundance for high-confidence PPIs, classified by MAGPIE (logistic regression probability ≥ 0.99, FDR = 20.77%). Gray cells represent successfully purified proteins that were identified by MS/MS, but that are not deemed high-confidence interactors in the IP-MS/MS experiments. Red colored cells were confidently detected by MAGPIE, while red cells with a dark gray outline represent common PPIs detected by both MAGPIE and SAINT. Dendrograms were generated using complete linkage and the Euclidean distance (C) Evaluating the robustness of MAGPIE. FDRs estimated for 1,000 randomized runs of MAGPIE as a function of the logistic regression probability thresholds, assessing the effects of the spectral pseudocount stochastic data imputation and random selection of negative training examples.

Known Protein−Protein Interactions Present within Our Data Sets

Protein interactions detected by our IP-MS/MS experiments were further validated against the STRING protein−protein interaction repository (version 11.0)[39] using the UniProt IDs of interacting proteins. STRING interactions considered were known and predicted protein−protein interactions with a STRING-derived medium confidence interaction score (score ≥ 0.4), which correspond to STRING's default query setting. Subnetworks of the known and predicted protein−protein interactions detected in our IP-MS/MS experiments were created using Cytoscape (version 3.8.0)[48] and annotated to denote if MAGPIE classified these interactions with high confidence. MAGPIE's high-confidence PPIs were also compared to those stored in the BioGRID (version 4.4.235) PPI database.[49]

AlphaFold 3 PPI Predictions

AlphaFold 3's web server[40] was used to predict PPIs between the proteins involved in the high confidence PPIs detected by MAGPIE. Sequences for each protein were downloaded from UniProtKB/SwissProt in August 2024. AlphaFold 3 was executed with the seed set to auto. A predicted template modeling (pTM) score ≥0.5 was used to indicate a valid PPI prediction.

Software Availability

MAGPIE is implemented as an open-source platform-independent Python (version 3.7) software package and is available for download at: https://github.com/LavalleeAdamLab/MAGPIE. The full outputs of both MAGPIE's and SAINT's analyses presented in this study are

https://doi.org/10.1021/acs.jproteome.4c00160
J. Proteome Res. 2025, 24, 383−396

provided in the Supporting Information, Files S1 and S2, respectively.

■ RESULTS

Overview of the Approach

A series of IP-MS/MS experiments using antibodies targeting proteins known to be present in human plasma were performed. Conversely, a series of IP-MS/MS experiments using antibodies targeting proteins not expected to be present in human plasma was performed to generate a negative control data set. Hierarchical clustering and principal component analysis (PCA) were used to identify a set of representative negative control experiments. This was followed by the construction of MAGPIE (Machine learning Assessment with loGistic regression of Protein−protein IntEractions), a supervised learning-based algorithm for assessing the confidence of putative PPIs in human plasma. Our experimental workflow is graphically depicted in Figure 1. MAGPIE was then benchmarked against a leading algorithm for assessing PPI confidence, SAINT. MAGPIE's high-confidence identifications were further validated using external repositories of PPIs.

Unsupervised Learning Reveals a Set of Experimental Negative Controls with Similar Nonspecific Bindings

The bootstrapped hierarchical clustering analysis, performed on the normalized spectral count data, revealed that experiments targeting known plasma proteins mostly cluster separately from the negative controls (Figure 2A,B). 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 (data not shown). Figure 2A shows these antibodies clustering with two other negative controls (antibody IDs: Anti-HA-IgG, Anti-RPAP2-IgG). Figure 2B, excluding the two outlier Anti-SRB7-IgG experiments, shows one cluster composed almost exclusively of negative controls and a second cluster of exclusively encompassing experiments of plasma proteins. Furthermore, the bootstrapping results are relatively higher in the two dominating clusters of Figure 2B versus those of Figure 2A, indicating more robust clusters in Figure 2B. PCA was run on the same normalized spectral count data set (Figure 2C,D) and also showed that the two Anti-SRB7-IgG experiments would likely be unreliable for modeling background noise in the MS/MS data, given the uniqueness of their profiles (Figure 2C). This analysis revealed the same clustering of negative controls separated from most of the experiments targeting plasma proteins upon removal of both Anti-SRB7-IgG experiments (Figure 2D).

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), that showed purification profiles different from those of the five antibodies that purified target plasma proteins: CNDP1 (×2), KLK6, SNCA, and PCSK9 (antibody IDs: Anti-CNDP1-IgG Ab 1, Anti-CNDP1-IgG Ab 2, Anti-KLK6-IgG, Anti-SNCA-IgG, and Anti-PCSK9-IgG, respectively). These experiments suggest that specific protein−protein interactions, differing from the background, are captured by these antibodies. Table S2 also shows that these antibodies are able to effectively purify their targets in plasma. The mass spectrometry data from these experiments were therefore used for all further analyses.

MAGPIE Identifies Plasma PPIs while Controlling for the False Discovery Rate

In constructing the training data sets, the positive training data set was composed of putative protein−protein interactions whose Z-scores were greater than or equal to three, totaling 226 positive training examples. To ensure a class-balanced training data set, 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 S1, showing that the resulting 226 negative training examples roughly represent all five negative control experiments equally. With the completed training data sets, MAGPIE's logistic regression model was trained, and its performance was evaluated through the implementation of our leave-one-control out (LOCO) cross-validation methodology (see Experimental Methods). Because MAGPIE implements a logistic regression model, the weight associated with each feature could be easily extracted. The normalized spectral count fold-change was attributed the most weight in the model (Table S3). False discovery rates were computed at given logistic regression probability thresholds for the false-positive identifications in each of the negative control experiments (Figure 3A). An overall false discovery rate was then derived, demonstrating that MAGPIE achieves a false discovery rate of 20.77% at the 0.99 logistic regression probability threshold (Figure S2). At this false discovery rate, MAGPIE identifies 68 high-confidence protein−protein interactions out of 882 putative protein−protein interactions tested (Table S4). Annotated spectra for the 68 interactions detected with a spectral count ≤50 are provided in the Supporting Information, File S3. Of importance, all five plasma proteins targeted by our five antibodies were identified with high confidence in their respective experiments, representing strong positive controls. Figure 3B presents 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.

MAGPIE's Predictions Are Robust to Differences in Training Sets

Given the stochasticity associated with MAGPIE's training, the addition of pseudocount values in the negative control experiments, and the random sampling of negative training examples from controls, we evaluated the robustness of MAGPIE's predictions. Figure 3C shows the resulting false discovery rates after MAGPIE was run 1000 times. Figure S3 shows the coefficient of variation of the FDR for these runs. Little variation is observed in the false discovery rates at any probability threshold, and a standard deviation of 0.01 was observed at a probability threshold of 0.99, indicating that MAGPIE's reported performances are minimally influenced by the randomly added pseudocount values or randomly sampled negative training examples. We investigated the minimum training set size necessary to maintain MAGPIE's performances. Figure S4 shows that four negative controls achieve prediction performances similar to those of five negative controls in terms of FDR control. However, training with three negative controls sees a large increase of FDR values, showing that MAGPIE's modeling is not as accurate and that four

https://doi.org/10.1021/acs.jproteome.4c00160
J. Proteome Res. 2025, 24, 383−396

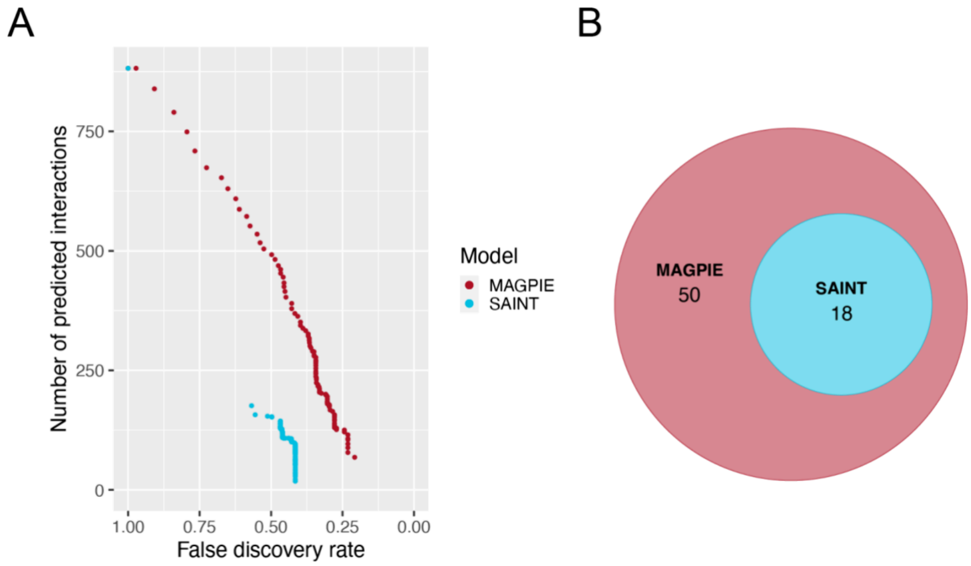


Figure 4. Benchmarking of MAGPIE against SAINT. (A) Comparing the number of predicted true interactions at estimated FDRs. (B) Venn diagram of high-confidence interactions (MAGPIE: logistic regression classifier probability ≥ 0.99, FDR = 20.77%; SAINT: probability ≥ 0.99, FDR = 41.53%) as predicted by each model.

negative controls appear to be the minimum needed to achieve reliable predictions.

MAGPIE Outperforms SAINT when Detecting Plasma PPIs

When comparing their identifications, MAGPIE outperforms SAINT in terms of both the number of interactions identified and their reliability. When applying our LOCO cross-validation methods, SAINT identified 18 high-confidence interactions at a probability threshold of 0.99 (the most stringent threshold), corresponding to a false discovery rate of 41.5% (Figure 4). At its 0.99 logistic regression probability threshold, MAGPIE identified more high-confidence interactions (Figure 4A), including the 18 identifications made by SAINT (Figure 4B). Additionally, SAINT fails to identify SNCA, targeted by the Anti-SNCA-IgG antibody (i.e., positive control), as one of its high-confidence interactions. We also show the level of overlap as a function of the output probabilities of both SAINT and MAGPIE (Figure S2). From this figure, we notice that the overlap gradually increases as probability increases. It is worth mentioning that, while SAINT is a leading algorithm for protein−protein interaction assessment, it was never designed to deal with human plasma samples or to be used with the type of negative controls employed in this investigation. This further emphasizes the need for a tailored approach to study and characterize plasma.

MAGPIE's PPI Identifications Are Corroborated by Interactions in PPI Repositories

Since there does not exist any experimentally validated large-scale protein−protein interaction data sets for the proteins targeted directly by the purifying antibodies used in our empirical IP-MS/MS experiments, an indirect method was used to validate MAGPIE's results. Of the five targeted proteins, two had interactions in human cell lines that were documented (known or predicted) in the STRING repository (PCSK9 and SNCA) (Figure 5).

Twenty-five of these known or predicted protein interactors of PCSK9 were found to be identified by mass spectrometry in our data set (Figure 5A). Of these, MAGPIE classified one, ceruloplasmin (CP), with a probability of 0.993, corresponding to an FDR of 20.77%. Five more known or predicted PCSK9 interactors were identified at a high probability, greater than or equal to 0.95, but less than 0.99 (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 data set (Figure 5B). One of these was identified as a high-confidence interactor, the serum amyloid P-component (APCS). The second STRING interactor of SNCA was not detected, kallikrein-6 (KLK6). However, the KLK6 protein was another one of the targets by a purifying antibody, Anti-KLK6-IgG, within our data set. 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 data set provides evidence that this previously unreported interaction may take place in human plasma. Another of MAGPIE's high-confidence interactions was corroborated by the BioGRID PPI repository. The interaction between SNCA and A1BG was previously reported by yeast-two-hybrid.[50] While this overlap with BioGRID may seem low at first glance, it is worth noting that the overwhelming majority of BioGRID's PPIs are derived from cell lines. This biological context is very different from that of plasma, where the PPIs detected by MAGPIE are taking place. Finally, Fredolini et al. reported an enrichment of ITIH1, when an antibody targeted CNDP1, that achieved a Z-score of 2.54, which was just south of their threshold of 3.[38] While this interaction is not formally reported in their paper, it was detected to be of high confidence in our analysis.

MAGPIE's PPI Identifications Are Corroborated by AlphaFold 3's PPI Predictions

Since the literature on PPIs taking place in human plasma is limited, we used AlphaFold 3's PPI predictions[51] to further validate the high-confidence PPIs detected by MAGPIE. 52 out of the 68 high-confidence PPIs from MAGPIE (76.5%) were confidently predicted by AlphaFold 3 (pTM score ≥ 0.5; Table S4). This fraction is in line with the FDR of 20.77%

https://doi.org/10.1021/acs.jproteome.4c00160
J. Proteome Res. 2025, 24, 383−396

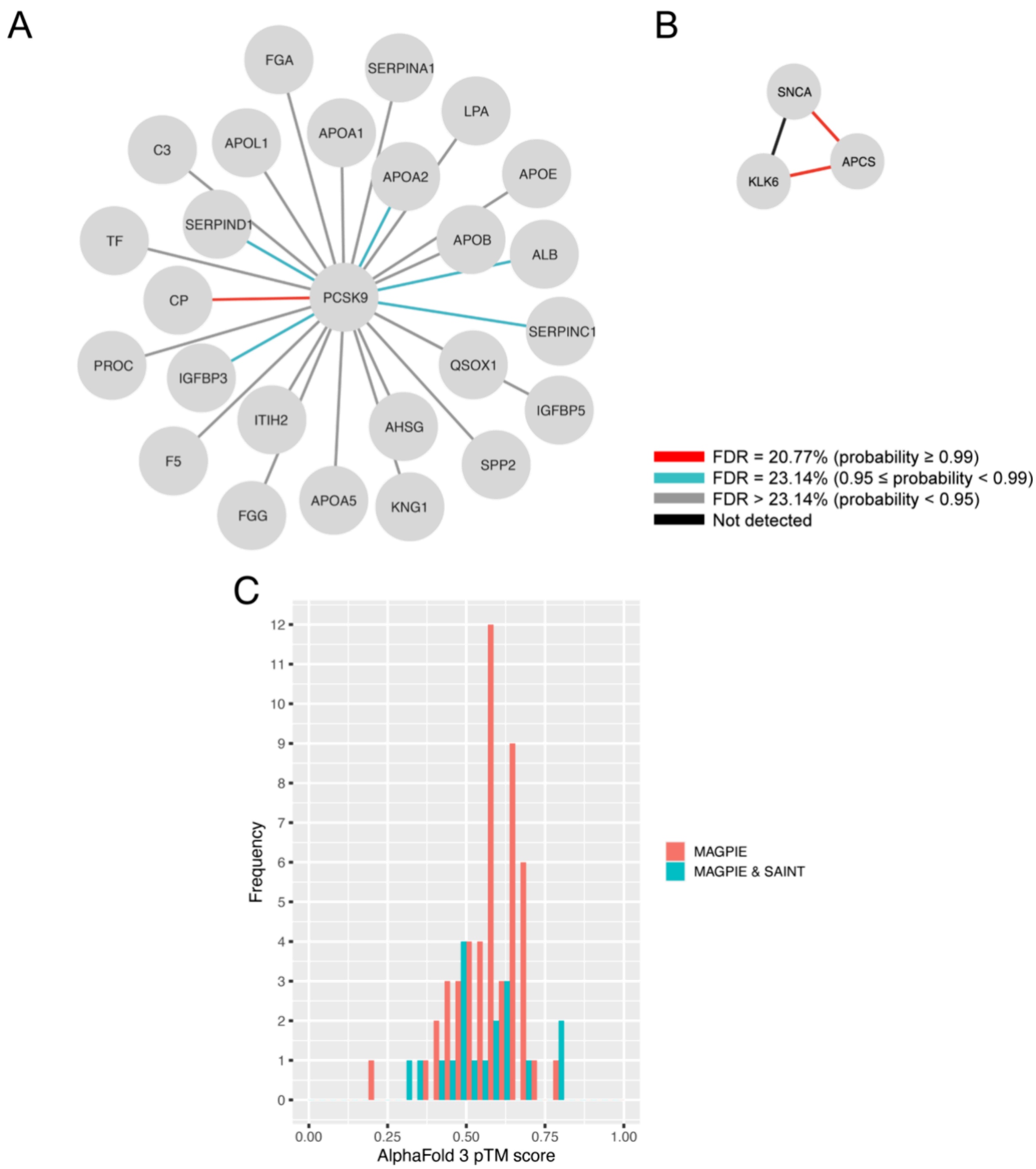


Figure 5. Protein−protein interaction subnetworks as identified by STRING. (A) Comparing the 25 known or predicted protein−protein interactions involving PCSK9, as identified by STRING, to the classification results outputted by MAGPIE. (B) Comparing the two known or predicted protein−protein interactions involving SNCA, as identified by STRING, to the classification results outputted by MAGPIE (B). (A, B) Interaction confidence level is color-coded. (C) Distribution of AlphaFold 3 pTM prediction confidence scores for the high confidence PPIs solely detected by MAGPIE (red) and by both MAGPIE and SAINT (blue).

achieved by these 68 PPIs. As for MAGPIE PPIs not validated by AlphaFold 3, it is worth noting that AlphaFold 3 predicts only direct PPIs. However, our approach may detect PPIs that are indirect. The interactions could need cofactors to occur; however, this scenario is not easily modeled using AlphaFold 3, given the large number of protein combinations possible. Interestingly, the ratio of PPIs validated by AlphaFold 3 was greater for the 50 PPIs solely detected by MAGPIE (82%) than for the 18 PPIs detected by both SAINT and MAGPIE (61%). This hints at the fact that the rate of true positives among PPIs solely predicted by MAGPIE is likely to be as high if not higher than those also detected by SAINT. Similarly, Figure 5C shows that the distribution of the confidence scores of the AlphaFold 3 predictions between these two sets of PPIs is very similar, again highlighting that PPIs solely detected by MAGPIE are likely as correct as those also detected by SAINT.

https://doi.org/10.1021/acs.jproteome.4c00160
J. Proteome Res. 2025, 24, 383−396

■ DISCUSSION

Negative Control Selection

Our proof-of-concept study highlighted that different antibodies targeting different proteins not expected to be in human plasma can indeed purify a similar set of proteins, constituting a model for background contamination and nonspecific binding. It also demonstrated that while these purification profiles are similar to each other, they differ enough from those of antibodies targeting known plasma proteins. This suggests that such antibodies tend to purify proteins that could be considered "sticky" (i.e., likely to be purified by many antibodies) or very abundant. This behavior enabled the creation of our MAGPIE algorithm and the confidence assessment of many novel human plasma PPIs. However, we acknowledge that our proof-of-concept remains at a small scale; thus, it is very likely that the antibodies we selected do not comprehensively cover all nonspecific plasma protein binding. This fact is probably one of the reasons why false positives were detected in our data sets, as spurious nonspecific binding events may not all be well modeled.

To improve results moving forward, performing a comprehensive survey of antibodies targeting proteins not expected to be present in human plasma may enable a more complete representation of nonspecific binding in plasma. Such modeling would likely more efficiently capture spurious contamination events.

Limitations of the Supervised Machine Learning Model

The training and testing data sets that could be produced from our data set facilitated the construction of a supervised machine learning model with moderately good performance (FDR of 20.77%). However, the negative control experiments targeted both proteins that are known to exist in other compartments in humans, such as the intracellular RNA polymerase II-associated protein 2 (antibody ID: Anti-RPAP2-IgG), and proteins that do not exist in humans, such as the synthetic molecular FLAG tag (antibody ID: Anti-Flag-IgG). While, in theory, the negative control antibodies were not targeting known human plasma proteins, there is a chance that they had some degree of affinity for plasma proteins. Indeed, proteins may share domains with the intended target of the antibody. We recognize that protein isoforms may also enter circulation. For instance, the protein−protein interaction network for RPAP2 has been extensively characterized in human cell lines.[52,53] Hemagglutinin (HA) is the most abundant surface glycoprotein of the influenza A virus[54] and the human host response to both the viral infection and its vaccination have been extensively characterized.[55] Should there be any affinity between the negative control antibodies and circulating plasma proteins, then the resulting protein purifications may not make the best possible examples of contamination and antibody nonspecific binding. Ideally, the proteins purified in the negative controls would solely constitute the background noise in the spectral count data for computational modeling, but this remains difficult to assess without some degree of uncertainty.

In its current state, MAGPIE's model requires at least 3 negative controls to be executed. In the proof-of-concept presented here, we showed that negative controls originating from different antibodies can capture the level of protein nonspecific binding in plasma. While our approach does not use standard replicates, these control act as such, with the advantage that negative controls using different antibodies capture a wider range of proteins than controls performed with the same antibody. They also likely make the standard deviation of the background protein spectral counts larger, which provides greater specificity of our approach by making it more challenging for a protein to be classified as a high confidence interaction. In the future, our model could, however, be modified to integrate standard replicate negative controls. Similarly, features could be added and modified in the logistic regression model to account for replicates of experiments using antibodies targeting plasma proteins. For instance, the model could incorporate a D-score, as used in the CompPASS PPI confidence assessment software to capture reproducibility.[26]

MAGPIE's machine learning approach focuses on maximizing the number of protein−protein interactions detected while minimizing the FDR. Minimizing FDR, although a conservative approach, may come at the expense of generating results with more false negatives. PPIs such as the one involving KLK6 and A1AT, which are known to interact in some biological fluids,[40] have been missed by our approach and may represent such a false negative. Given that the fold-change of A1AT with respect to its presence in negative controls is low (1.155), it is also possible that the antibody used to target KLK6 interfered with this potential interaction.

Indirect Validation of High-Confidence Interactions Using STRING

Because there do not exist experimentally validated data sets for protein−protein interactions in human plasma for the four proteins in our data set, external public repositories were 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.[56] While this approach was not an absolute method for validation, it provided insights into the confidence and potential relevancy of MAGPIE's identifications. It is worth noting that, because 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. The small overlap between MAGPIE's results and STRING is therefore expected, as most of the interactions in the repository are based on experiments in cell lines.

■ CONCLUSIONS

In this paper, we present MAGPIE, a novel machine learning-based tool for assessing the confidence of putative protein−protein interactions in human plasma. The methodologies developed and MAGPIE are the first of their kind in the pursuit of characterizing the human plasma interactome and providing a better understanding of the biological processes taking place in it.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.4c00160.

(Figure S1) Proportion of negative examples randomly sampled from each negative control experiment for supervised machine learning training; (Figure S2) number of PPIs detected by SAINT and MAGPIE and their overlap as a function of their respective probability thresholds; (Figure S3) coefficient of variation of the false discovery rates of MAGPIE as a function of its probability threshold for 1000 executions; (Figure S4) false discovery rates derived by MAGPIE as a function of its probability threshold using all possible combinations of 3 negative controls and 4 negative controls (PDF)

(Table S1) Antibody information and sources; (Table S2) normalized spectral counts of plasma proteins targeted by antibodies for all antibodies; (Table S3) logistic regression model weights computed for all five of MAGPIE's features; (Table S4) all candidate PPIs assessed by MAGPIE along with their classifying features, probabilities, FDR values, SAINT probabilities and FDR values, and results from BioGRID search and AlphaFold 3 predictions (XLSX)

(File S1) MAGPIE output files in a compressed folder; (File S2) SAINT output file; (File S3) annotated spectra for high-confidence protein−protein interactions in a compressed folder (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Mathieu Lavallée-Adam − *Department of Biochemistry, Microbiology and Immunology and Ottawa Institute of Systems Biology, Faculty of Medicine, University of Ottawa, Ottawa, Ontario K1H 8M5, Canada;* orcid.org/0000-0003-2124-3872; Phone: 1-613-562-5800 ext. 8224; Email: mathieu.lavallee@uottawa.ca

Authors

Emily Hashimoto-Roth − *Department of Biochemistry, Microbiology and Immunology and Ottawa Institute of Systems Biology, Faculty of Medicine, University of Ottawa, Ottawa, Ontario K1H 8M5, Canada*

Diane Forget − *Translational Proteomics Laboratory, Institut de recherches cliniques de Montréal, Québec H2W 1R7, Canada*

Vanessa P. Gaspar − *Translational Proteomics Laboratory, Institut de recherches cliniques de Montréal, Québec H2W 1R7, Canada*

Steffany A. L. Bennett − *Department of Biochemistry, Microbiology and Immunology and Ottawa Institute of Systems Biology, Faculty of Medicine, University of Ottawa, Ottawa, Ontario K1H 8M5, Canada; Department of Chemistry and Biomolecular Sciences, Centre for Catalysis and Research Innovation, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada*

Marie-Soleil Gauthier − *Translational Proteomics Laboratory, Institut de recherches cliniques de Montréal, Québec H2W 1R7, Canada*

Benoit Coulombe − *Translational Proteomics Laboratory, Institut de recherches cliniques de Montréal, Québec H2W 1R7, Canada; Département de biochimie et médecine moléculaire, Faculté de médecine, Université de Montréal, Québec H3C 3J7, Canada;* orcid.org/0000-0003-1702-0692

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.4c00160

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge funding from the following sources: Natural Sciences and Engineering Research Council (NSERC) of Canada Discovery grants to M.L.A. and S.A.L.B. E.H.R. was funded by a stipend from the NSERC CREATE in Metabolomics Advanced Training and International Exchange (MATRIX) Program. This research was enabled in part by support provided by an IVADO and Génome Québec grant to M.L.A and B.C. The authors are grateful to the team of the Institut de recherches cliniques de Montréal (IRCM) Mass Spectrometry and Proteomics platform for their assistance in sample processing and mass spectrometric data acquisition.

■ ABBREVIATIONS

IP, immunoprecipitation; PPI, protein−protein Interaction; FDR, false discovery rate; PCA, principal component analysis; LC, liquid chromatography; MS, mass spectrometry; MS/MS, tandem mass spectrometry; XL, cross-linking; MAGPIE, machine learning assessment with logistic regression of protein−protein interactions; LOCO, leave-one-control-out; SAINT, significance analysis of interactome

■ REFERENCES

(1) Schwikowski, B.; Uetz, P.; Fields, S. A network of protein−protein interactions in yeast. *Nat. Biotechnol.* **2000**, *18* (12), 1257−1261.

(2) Krogan, N. J.; Cagney, G.; Yu, H.; Zhong, G.; Guo, X.; Ignatchenko, A.; Li, J.; Pu, S.; Datta, N.; Tikuisis, A. P.; Punna, T.; Peregrín-Alvarez, J. M.; Shales, M.; Zhang, X.; Davey, M.; Robinson, M. D.; Paccanaro, A.; Bray, J. E.; Sheung, A.; Beattie, B.; Richards, D. P.; Canadien, V.; Lalev, A.; Mena, F.; Wong, P.; Starostine, A.; Canete, M. M.; Vlasblom, J.; Wu, S.; Orsi, C.; Collins, S. R.; Chandran, S.; Haw, R.; Rilstone, J. J.; Gandi, K.; Thompson, N. J.; Musso, G.; St Onge, P.; Ghanny, S.; Lam, M. H. Y.; Butland, G.; Altaf-Ul, A. M.; Kanaya, S.; Shilatifard, A.; O'Shea, E.; Weissman, J. S.; Ingles, C. J.; Hughes, T. R.; Parkinson, J.; Gerstein, M.; Wodak, S. J.; Emili, A.; Greenblatt, J. F. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature.* **2006**, *440* (7084), 637−643.

(3) Pitre, S.; North, C.; Alamgir, M.; Jessulat, M.; Chan, A.; Luo, X.; Green, J. R.; Dumontier, M.; Dehne, F.; Golshani, A. Global investigation of protein−protein interactions in yeast Saccharomyces cerevisiae using re-occurring short polypeptide sequences. *Nucleic Acids Res.* **2008**, *36* (13), 4286−4294.

(4) Breitkreutz, A.; Choi, H.; Sharom, J. R.; Boucher, L.; Neduva, V.; Larsen, B.; Lin, Z. Y.; Breitkreutz, B. J.; Stark, C.; Liu, G.; Ahn, J.; Dewar-Darch, D.; Reguly, T.; Tang, X.; Almeida, R.; Qin, Z. S.; Pawson, T.; Gingras, A. C.; Nesvizhskii, A. I.; Tyers, M. A Global Protein Kinase and Phosphatase Interaction Network in Yeast. *Science (1979).* **2010**, *328* (5981), 1043−1046.

(5) Simonis, N.; Rual, J. F.; Carvunis, A. R.; Tasan, M.; Lemmens, I.; Hirozane-Kishikawa, T.; Hao, T.; Sahalie, J. M.; Venkatesan, K.; Gebreab, F.; Cevik, S.; Klitgord, N.; Fan, C.; Braun, P.; Li, N.; Ayivi-Guedehoussou, N.; Dann, E.; Bertin, N.; Szeto, D.; Dricot, A.; Yildirim, M. A.; Lin, C.; de Smet, A. S.; Kao, H. L.; Simon, C.; Smolyar, A.; Ahn, J. S.; Tewari, M.; Boxem, M.; Milstein, S.; Yu, H.; Dreze, M.; Vandenhaute, J.; Gunsalus, K. C.; Cusick, M. E.; Hill, D. E.; Tavernier, J.; Roth, F. P.; Vidal, M. Empirically controlled

mapping of the Caenorhabditis elegans protein-protein interactome network. *Nat. Methods.* **2009**, *6* (1), 47−54.
(6) Go, C. D.; Knight, J. D. R.; Rajasekharan, A.; Rathod, B.; Hesketh, G. G.; Abe, K. T.; Youn, J. Y.; Samavarchi-Tehrani, P.; Zhang, H.; Zhu, L. Y.; Popiel, E.; Lambert, J. P.; Coyaud, É; Cheung, S. W. T.; Rajendran, D.; Wong, C. J.; Antonicka, H.; Pelletier, L.; Palazzo, A. F.; Shoubridge, E. A.; Raught, B.; Gingras, A. C. A proximity-dependent biotinylation map of a human cell. *Nature.* **2021**, *595* (7865), 120−124.
(7) Hein, M. Y.; Hubner, N. C.; Poser, I.; Cox, J.; Nagaraj, N.; Toyoda, Y.; Gak, I. A.; Weisswange, I.; Mansfeld, J.; Buchholz, F.; Hyman, A. A.; Mann, M. A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell.* **2015**, *163* (3), 712−723.
(8) Minton, K. Strength in numbers. *Nat. Rev. Mol. Cell Biol.* **2015**, *16* (12), 702−703.
(9) Huttlin, E. L.; Ting, L.; Bruckner, R. J.; Gebreab, F.; Gygi, M. P.; Szpyt, J.; Tam, S.; Zarraga, G.; Colby, G.; Baltier, K.; Dong, R.; Guarani, V.; Vaites, L. P.; Ordureau, A.; Rad, R.; Erickson, B. K.; Wühr, M.; Chick, J.; Zhai, B.; Kolippakkam, D.; Mintseris, J.; Obar, R. A.; Harris, T.; Artavanis-Tsakonas, S.; Sowa, M. E.; De Camilli, P.; Paulo, J. A.; Harper, J. W.; Gygi, S. P. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell.* **2015**, *162* (2), 425−440.
(10) Huttlin, E. L.; Bruckner, R. J.; Navarrete-Perea, J.; Cannon, J. R.; Baltier, K.; Gebreab, F.; Gygi, M. P.; Thornock, A.; Zarraga, G.; Tam, S.; Szpyt, J.; Gassaway, B. M.; Panov, A.; Parzen, H.; Fu, S.; Golbazi, A.; Maenpaa, E.; Stricker, K.; Guha Thakurta, S.; Zhang, T.; Rad, R.; Pan, J.; Nusinow, D. P.; Paulo, J. A.; Schweppe, D. K.; Vaites, L. P.; Harper, J. W.; Gygi, S. P. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell.* **2021**, *184* (11), 3022−3040.
(11) Vasilescu, J.; Figeys, D. Mapping protein−protein interactions by mass spectrometry. *Curr. Opin Biotechnol.* **2006**, *17* (4), 394−399.
(12) Rigaut, G.; Shevchenko, A.; Rutz, B.; Wilm, M.; Mann, M.; Séraphin, B. A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **1999**, *17* (10), 1030−1032.
(13) Li, Y. Commonly used tag combinations for tandem affinity purification. *Biotechnol Appl. Biochem.* **2010**, *55* (2), 73−83.
(14) Hopp, T. P.; Prickett, K. S.; Price, V. L.; Libby, R. T.; March, C. J.; Pat Cerretti, D.; Urdal, D. L.; Conlon, P. J. A Short Polypeptide Marker Sequence Useful for Recombinant Protein Identification and Purification. *Bio/Technology.* **1988**, *6* (10), 1204−1210.
(15) Roux, K. J.; Kim, D. I.; Burke, B.; May, D. G. BioID: A Screen for Protein-Protein Interactions. *Curr. Protoc Protein Sci.* **2018**, *91* (1), 19.
(16) Young, M. M.; Tang, N.; Hempel, J. C.; Oshiro, C. M.; Taylor, E. W.; Kuntz, I. D.; Gibson, B. W.; Dollinger, G. High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proceedings of the National Academy of Sciences.* **2000**, *97* (11), 5802−5806.
(17) 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.* **2016**, *41* (1), 20−32.
(18) Leitner, A.; Walzthoeni, T.; Kahraman, A.; Herzog, F.; Rinner, O.; Beck, M.; Aebersold, R. Probing Native Protein Structures by Chemical Cross-linking, Mass Spectrometry, and Bioinformatics. *Molecular & Cellular Proteomics.* **2010**, *9* (8), 1634−1649.
(19) Liu, Q.; Zheng, J.; Sun, W.; Huo, Y.; Zhang, L.; Hao, P.; Wang, H.; Zhuang, M. A proximity-tagging system to identify membrane protein−protein interactions. *Nat. Methods.* **2018**, *15* (9), 715−722.
(20) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods.* **2007**, *4* (3), 207−214.
(21) Keller, B. O.; Sui, J.; Young, A. B.; Whittal, R. M. Interferences and contaminants encountered in modern mass spectrometry. *Anal. Chim. Acta* **2008**, *627* (1), 71−81.
(22) Mellacheruvu, D.; Wright, Z.; Couzens, A. L.; Lambert, J. P.; St-Denis, N. A.; Li, T.; Miteva, Y. V.; Hauri, S.; Sardiu, M. E.; Low, T. Y.; Halim, V. A.; Bagshaw, R. D.; Hubner, N. C.; Al-Hakim, A.; Bouchard, A.; Faubert, D.; Fermin, D.; Dunham, W. H.; Goudreault, M.; Lin, Z. Y.; Badillo, B. G.; Pawson, T.; Durocher, D.; Coulombe, B.; Aebersold, R.; Superti-Furga, G.; Colinge, J.; Heck, A. J. R.; Choi, H.; Gstaiger, M.; Mohammed, S.; Cristea, I. M.; Bennett, K. L.; Washburn, M. P.; Raught, B.; Ewing, R. M.; Gingras, A. C.; Nesvizhskii, A. I. The CRAPome: a contaminant repository for affinity purification−mass spectrometry data. *Nat. Methods.* **2013**, *10* (8), 730−736.
(23) Choi, H.; Larsen, B.; Lin, Z. Y.; Breitkreutz, A.; Mellacheruvu, D.; Fermin, D.; Qin, Z. S.; Tyers, M.; Gingras, A. C.; Nesvizhskii, A. I. SAINT: probabilistic scoring of affinity purification−mass spectrometry data. *Nat. Methods.* **2011**, *8* (1), 70−73.
(24) Lavallée-Adam, M.; Cloutier, P.; Coulombe, B.; Blanchette, M. Modeling contaminants in AP-MS/MS experiments. *J. Proteome Res.* **2011**, *10* (2), 886−895.
(25) Lavallée-Adam, M.; Rousseau, J.; Domecq, C.; Bouchard, A.; Forget, D.; Faubert, D.; Blanchette, M.; Coulombe, B. Discovery of cell compartment specific protein-protein interactions using affinity purification combined with tandem mass spectrometry. *J. Proteome Res.* **2013**, *12* (1), 272−281.
(26) Sowa, M. E.; Bennett, E. J.; Gygi, S. P.; Harper, J. W. Defining the human deubiquitinating enzyme interaction landscape. *Cell.* **2009**, *138* (2), 389−403.
(27) Dunham, W. H.; Mullin, M.; Gingras, A. C. Affinity-purification coupled to mass spectrometry: Basic principles and strategies. *Proteomics.* **2012**, *12* (10), 1576−1590.
(28) Altelaar, A. F. M.; Munoz, J.; Heck, A. J. R. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* **2013**, *14* (1), 35−48.
(29) Choi, H.; Larsen, B.; Lin, Z. Y.; Breitkreutz, A.; Mellacheruvu, D.; Fermin, D.; Qin, Z. S.; Tyers, M.; Gingras, A. C.; Nesvizhskii, A. I. SAINT: probabilistic scoring of affinity purification−mass spectrometry data. *Nat. Methods.* **2011**, *8* (1), 70−73.
(30) Sardiu, M. E.; Cai, Y.; Jin, J.; Swanson, S. K.; Conaway, R. C.; Conaway, J. W.; Florens, L.; Washburn, M. P. Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. *Proceedings of the National Academy of Sciences.* **2008**, *105* (5), 1454−1459.
(31) 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.* **1999**, *20* (18), 3551−3567.
(32) Lavallée-Adam, M.; Cloutier, P.; Coulombe, B.; Blanchette, M. Modeling contaminants in AP-MS/MS experiments. *J. Proteome Res.* **2011**, *10* (2), 886−895.
(33) Sowa, M. E.; Bennett, E. J.; Gygi, S. P.; Harper, J. W. Defining the Human Deubiquitinating Enzyme Interaction Landscape. *Cell.* **2009**, *138* (2), 389−403.
(34) Rinner, O.; Mueller, L. N.; Hubálek, M.; Müller, M.; Gstaiger, M.; Aebersold, R. An integrated mass spectrometric and computational framework for the analysis of protein interaction networks. *Nat. Biotechnol.* **2007**, *25* (3), 345−352.
(35) Cloutier, P.; Al-Khoury, R.; Lavallée-Adam, M.; Faubert, D.; Jiang, H.; Poitras, C.; Bouchard, A.; Forget, D.; Blanchette, M.; Coulombe, B. High-resolution mapping of the protein interaction network for the human transcription machinery and affinity purification of RNA polymerase II-associated complexes. *Methods.* **2009**, *48* (4), 381−386.
(36) Gauthier, M. S.; Pérusse, J. R.; Awan, Z.; Bouchard, A.; Tessier, S.; Champagne, J.; Krastins, B.; Byram, G.; Chabot, K.; Garneau, P.; Rabasa-Lhoret, R.; Faubert, D.; Lopez, M. F.; Seidah, N. G.; Coulombe, B. 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.* **2015**, *81*, 66−73.

https://doi.org/10.1021/acs.jproteome.4c00160
J. Proteome Res. 2025, 24, 383−396

(37) Caselli, C.; Del Turco, S.; Ragusa, R.; Lorenzoni, V.; De Graaf, M.; Basta, G.; Scholte, A.; De Caterina, R.; Neglia, D. Association of PCSK9 plasma levels with metabolic patterns and coronary atherosclerosis in patients with stable angina. *Cardiovasc Diabetol.* **2019**, *18* (1), 144.

(38) Fredolini, C.; Byström, S.; Sanchez-Rivera, L.; Ioannou, M.; Tamburro, D.; Pontén, F.; Branca, R. M.; Nilsson, P.; Lehtiö, J.; Schwenk, J. M. Systematic assessment of antibody selectivity in plasma based on a resource of enrichment profiles. *Sci. Rep.* **2019**, *9* (1), 8324.

(39) Jensen, L. J.; Kuhn, M.; Stark, M.; Chaffron, S.; Creevey, C.; Muller, J.; Doerks, T.; Julien, P.; Roth, A.; Simonovic, M.; Bork, P.; von Mering, C. STRING 8 - A global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* **2009**, *37* (SUPPL. 1), 412−416.

(40) Bloemberg, D.; Nguyen, T.; MacLean, S.; Zafer, A.; Gadoury, C.; Gurnani, K.; Chattopadhyay, A.; Ash, J.; Lippens, J.; Harcus, D.; Pagé, M.; Fortin, A.; Pon, R. A.; Gilbert, R.; Marcil, A.; Weeratna, R. D.; McComb, S. A High-Throughput Method for Characterizing Novel Chimeric Antigen Receptors in Jurkat Cells. *Mol. Ther Methods Clin Dev.* **2020**, *16*, 238−254.

(41) Gauthier, M. S.; Awan, Z.; Bouchard, A.; Champagne, J.; Tessier, S.; Faubert, D.; Chabot, K.; Garneau, P. Y.; Rabasa-Lhoret, R.; Seidah, N. G.; Ridker, P. M.; Genest, J.; Coulombe, B. 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.* **2018**, *12* (4), 1027−1038.

(42) UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51* (D1), D523−D531.

(43) Searle, B. C. Scaffold: A bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics.* **2010**, *10* (6), 1265−1269.

(44) Deutsch, E. W.; Bandeira, N.; Perez-Riverol, Y.; Sharma, V.; Carver, J. J.; Mendoza, L.; Kundu, D. J.; Wang, S.; Bandla, C.; Kamatchinathan, S.; Hewapathirana, S.; Pullman, B. S.; Wertz, J.; Sun, Z.; Kawano, S.; Okuda, S.; Watanabe, Y.; MacLean, B.; MacCoss, M. J.; Zhu, Y.; Ishihama, Y.; Vizcaíno, J. A. The ProteomeXchange consortium at 10 years: 2023 update. *Nucleic Acids Res.* **2023**, *51* (D1), D1539−D1548.

(45) Perez-Riverol, Y.; Bai, J.; Bandla, C.; García-Seisdedos, D.; Hewapathirana, S.; Kamatchinathan, S.; Kundu, D. J.; Prakash, A.; Frericks-Zipper, A.; Eisenacher, M.; Walzer, M.; Wang, S.; Brazma, A.; Vizcaíno, J. A. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* **2022**, *50* (D1), D543−D552.

(46) Suzuki, R.; Shimodaira, H. Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics.* **2006**, *22* (12), 1540−1542.

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

(48) Shannon, Paul; Markiel, Andrew; Ozier, Owen; Baliga, Nitin S.; Wang, Jonathan T.; Ramage, Daniel; Amin, Nada; Schwikowski, Benno; Ideker, Trey Cytoscape: A Software Environment for Integrated Models. *Genome Res.* **1971**, *13* (22), 426.

(49) Oughtred, R.; Rust, J.; Chang, C.; Breitkreutz, B.; Stark, C.; Willems, A.; Boucher, L.; Leung, G.; Kolas, N.; Zhang, F.; Dolma, S.; Coulombe-Huntington, J.; Chatr-aryamontri, A.; Dolinski, K.; Tyers, M. The < scp > BioGRID</scp> database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* **2021**, *30* (1), 187−200.

(50) Vinayagam, A.; Stelzl, U.; Foulle, R.; Plassmann, S.; Zenkner, M.; Timm, J.; Assmus, H. E.; Andrade-Navarro, M. A.; Wanker, E. E. A Directed Protein Interaction Network for Investigating Intracellular Signal Transduction. *Sci. Signal.* **2011**, *4* (189), rs8.

(51) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C. C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Žídek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature.* **2024**, *630* (8016), 493−500.

(52) Jeronimo, C.; Forget, D.; Bouchard, A.; Li, Q.; Chua, G.; Poitras, C.; Thérien, C.; Bergeron, D.; Bourassa, S.; Greenblatt, J.; Chabot, B.; Poirier, G. G.; Hughes, T. R.; Blanchette, M.; Price, D. H.; Coulombe, B. Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme. *Mol. Cell* **2007**, *27* (2), 262−274.

(53) Cloutier, P.; Al-Khoury, R.; Lavallée-Adam, M.; Faubert, D.; Jiang, H.; Poitras, C.; Bouchard, A.; Forget, D.; Blanchette, M.; Coulombe, B. High-resolution mapping of the protein interaction network for the human transcription machinery and affinity purification of RNA polymerase II-associated complexes. *Methods.* **2009**, *48* (4), 381−386.

(54) Chen, X.; Liu, S.; Goraya, M. U.; Maarouf, M.; Huang, S.; Chen, J. L. Host Immune Response to Influenza A Virus Infection. *Front Immunol.* **2018**, *9*, 9.

(55) Krammer, F. The human antibody response to influenza A virus infection and vaccination. *Nat. Rev. Immunol.* **2019**, *19* (6), 383−397.

(56) 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.* **2005**, *21* (16), 3360−3368.

■ NOTE ADDED AFTER ASAP PUBLICATION

A partially corrected version of this paper was published on January 7, 2025. Corrections were made throughout the document, and the fully corrected version of the paper was published ASAP on January 14, 2025.