nature methods

ARTICLES

https://doi.org/10.1038/s41592-021-01194-4

Meta-analysis defines principles for the design and analysis of co-fractionation mass spectrometry experiments

Michael A. Skinnider[1] and Leonard J. Foster[1,2] ✉

Co-fractionation mass spectrometry (CF-MS) has emerged as a powerful technique for interactome mapping. However, there is little consensus on optimal strategies for the design of CF-MS experiments or their computational analysis. Here, we reanalyzed a total of 206 CF-MS experiments to generate a uniformly processed resource containing over 11 million measurements of protein abundance. We used this resource to benchmark experimental designs for CF-MS studies and systematically optimize computational approaches to network inference. We then applied this optimized methodology to reconstruct a draft-quality human interactome by CF-MS and predict over 700,000 protein–protein interactions across 27 eukaryotic species or clades. Our work defines new resources to illuminate proteome organization over evolutionary timescales and establishes best practices for the design and analysis of CF-MS studies.

Biological function arises from the dynamic organization of proteins in networks of physical interactions. Charting the complete protein–protein interaction network (the 'interactome') has thus been a long-standing goal of the post-genomic era, with a view to understanding cellular physiology and its perturbation in disease. Systematic screens have produced large-scale interactome maps in humans and model organisms, primarily using affinity purification–mass spectrometry (AP–MS) or yeast two-hybrid (Y2H) assays[1–5]. However, these methods are laborious, difficult to scale or apply to non-model organisms and require the introduction of protein tags that can disrupt interactions or alter localization[6,7].

CF-MS has emerged as an alternative strategy for interactome mapping that addresses several shortcomings of conventional methods[8,9]. In particular, CF-MS is capable of interactome mapping in high throughput, under native cellular conditions, in species not amenable to genetic manipulation and even across multiple species simultaneously[10,11]. However, CF-MS is still a relatively young technique, and the field has yet to arrive at a consensus regarding the best way to carry out a CF-MS experiment or analyze the resulting data. Many different experimental and analytical approaches have been proposed, with relatively little agreement between laboratories. Notably, many of the analytical approaches that have been proposed have been tested in only a single dataset[12–14], raising the question of how well these methods may generalize to other datasets, given the diversity of experimental designs employed in the field. A related issue is that the development of a computational pipeline for CF-MS data entails a series of analytical decisions, encompassing strategies for protein quantification, quality control and preprocessing, scoring elution profile similarity and integrating data from multiple replicates. Benchmarks to date have compared entire pipelines as distributed without systematically dissecting the impact of each of these decisions in turn, precluding a deeper understanding of the optimal methodology for analysis of CF-MS data and raising the possibility that some are currently made in an arbitrary manner.

We hypothesized that a comprehensive reanalysis of all published CF-MS datasets could allow us to conclusively establish best practices for both design and analysis of CF-MS studies. We uniformly reprocessed a total of 206 published CF-MS experiments, collectively spanning >12,000 fractions, to produce a resource with >11 million protein quantifications. We then used this resource to retrospectively benchmark experimental workflows and systematically dissect computational approaches to network inference from CF-MS data. Finally, we applied optimized analytical strategies to predict consensus CF-MS interactomes for 27 species or phylogenetic clades throughout the eukaryotic evolutionary tree.

Results

A comprehensive resource of uniformly processed CF-MS data. A survey of the literature identified 206 published CF-MS experiments with raw data deposited to public proteomic databases (Fig. 1a and Supplementary Table 1). These experiments were both biologically and technically heterogeneous, occurring in 24 different species and employing disparate approaches to fractionation and protein quantification (Fig. 1b and Extended Data Fig. 1a). We reasoned that this heterogeneity could allow us to retrospectively benchmark each facet of experimental design, while simultaneously providing an ideal testbed for computational analysis strategies. However, the divergent approaches to database search, protein quantification, quality control and data pre-processing employed by authors of the original studies presented an obstacle to an integrated analysis. We therefore reanalyzed the entire resource of 12,683 fractions, corresponding to over 27 months of uninterrupted instrument time, using MaxQuant[15]. From a total of 644 million tandem mass spectra, over 151 million peptides were sequenced, yielding 11.7 million measurements of protein abundance. On average, 2,386 protein groups were quantified in each experiment, corresponding to 16% of the organismal proteome (Fig. 1c and Extended Data Fig. 1b,c).

To evaluate the completeness of the integrated dataset, we calculated the recall of protein complexes from the Comprehensive Resource of Mammalian Protein Complexes (CORUM) database[16]

[1] Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia, Canada. [2] Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, British Columbia, Canada. ✉e-mail: foster@msl.ubc.ca

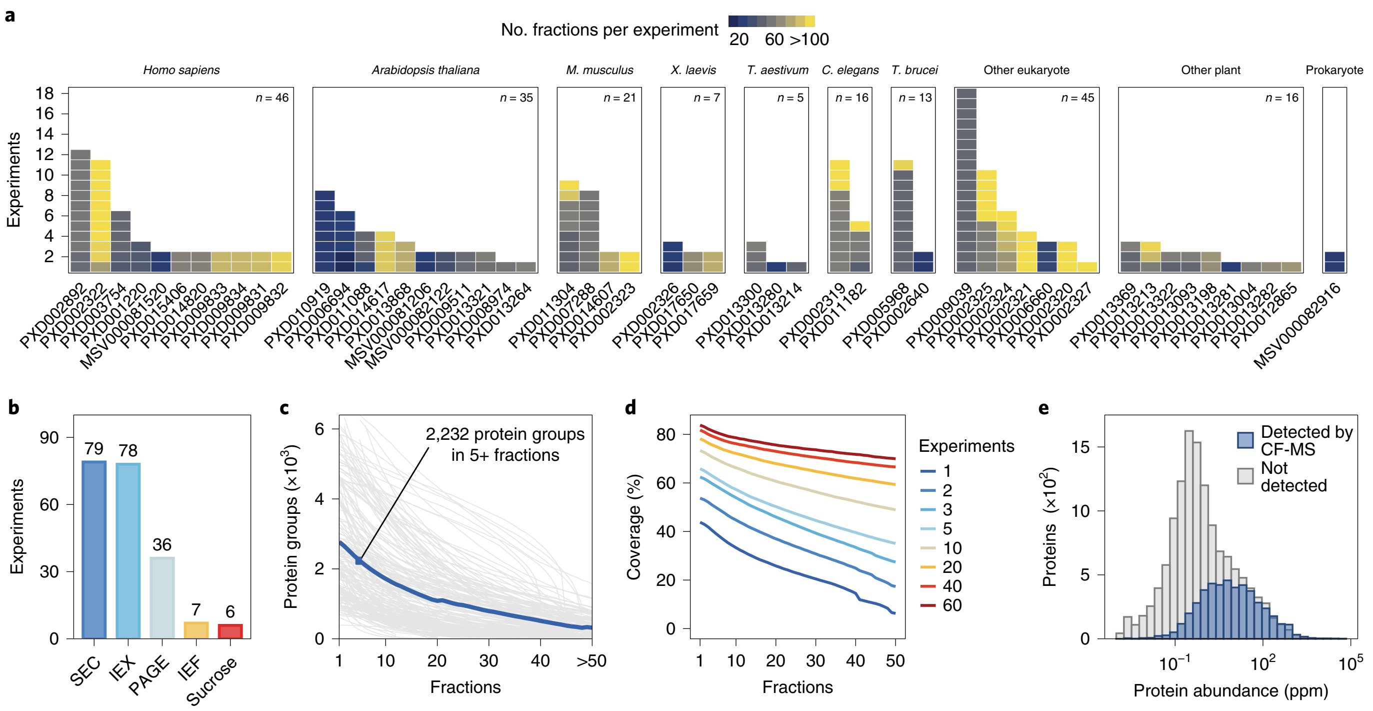


Fig. 1 | A comprehensive reanalysis of published CF-MS data. a, Overview of the 206 CF-MS experiments analyzed in this study. Each cell represents one CF-MS experiment. Cells are shaded by the number of fractions collected in each experiment. *M. musculus*, *Mus musculus*; *X. laevis*, *Xenopus laevis*; *T. aestivum*, *Triticum aestivum*; *C. elegans*, *Caenorhabditis elegans*. **b**, Fractionation approaches employed in published CF-MS experiments. IEF, isoelectric focusing; PAGE, polyacrylamide gel electrophoresis. **c**, Number of protein groups quantified in each CF-MS experiment (gray lines, individual datasets; blue line, mean across all datasets). **d**, Average coverage of CORUM protein complexes in random samples of 1–60 experiments from the subset of 67 human and mouse datasets. **e**, PaxDb[57] consensus protein abundance of human proteins detected and not detected by CF-MS.

in the subset of human and mouse experiments, finding that 84.1% of CORUM proteins were detected in at least one fraction (Fig. 1d). The remaining 15.9% of undetected proteins were enriched for Gene Ontology (GO) terms related to cellular signaling and proliferation, as well as transmembrane protein categories such as G protein-coupled receptors and ion channels, suggesting that more tailored methods may be necessary to interrogate the undetected complexes using CF-MS (Extended Data Fig. 1d). Proteins identified by CF-MS were also significantly more abundant than the proteome average (Fig. 1e and Extended Data Fig. 1e), confirming previous reports[1,17]. Studies employing longer liquid chromatography gradients achieved greater coverage of low-abundance proteins, but absolute coverage remained low, likely at least in part because low-abundance proteins are underrepresented among known protein complexes (Extended Data Fig. 1f–h).

For the subset of 178 experiments in which processed CF-MS chromatograms accompanied the original publication, we compared our own uniformly processed datasets to those analyzed by the authors of these studies. The number of high-quality chromatograms (that is, with proteins detected in at least five fractions) increased by an average of 24% compared to the original datasets, supporting our strategy of reanalyzing the raw data (Extended Data Fig. 1i). With nearly 12 million measurements of protein abundance spanning >12,000 fractions, our systematic reanalysis of published CF-MS data is among the largest existing collections of uniformly processed proteomic data[18–21], providing a resource to understand proteome architecture over evolutionary timescales.

A systematic benchmark optimizes analysis of individual CF-MS datasets. We next aimed to mine this comprehensive resource of CF-MS data to establish optimal strategies for experimental design and data analysis. Because the central premise of CF-MS is that protein complexes should co-elute across a separation gradient, we reasoned that superior experimental or analytical approaches should exhibit a greater ability to resolve known protein complexes. To formalize this notion, we quantified the degree to which known protein complexes could be recovered based on their observed patterns of co-abundance across CF-MS fractions using receiver operating characteristic (ROC) curve analysis[22] (Fig. 2a). In this framework, an area under the curve (AUC) of 1 reflects a perfect ability to identify known complexes from CF-MS data, whereas an AUC of 0.5 reflects random performance.

With this quantitative basis for comparison in hand, we first set out to establish an optimal pipeline for the analysis of individual CF-MS datasets. Perhaps the most central operation in the analysis of CF-MS data is to quantify the similarity of two protein chromatograms, which in turn provides a basis for the inference of protein–protein interactions. More than a dozen metrics have been used to quantify chromatographic similarity to date, with relatively little agreement between studies (Extended Data Fig. 2a). We therefore compared 24 measures of association for their ability to resolve known protein complexes in the 67 mouse and human datasets[23] (Fig. 2b and Supplementary Table 2a). Surprisingly, the two most ubiquitous metrics displayed starkly different trends: the Pearson correlation was among the top-performing metrics, but the Euclidean distance performed no better than random chance. The four top-performing metrics (mutual information, distance correlation, cosine distance and weighted cross-correlation) yielded nearly indistinguishable AUCs, raising the possibility of an upper limit to protein complex inference (Extended Data Fig. 3a and Supplementary Table 3a). To corroborate these trends, we performed a second analysis using GO annotations, rather than known

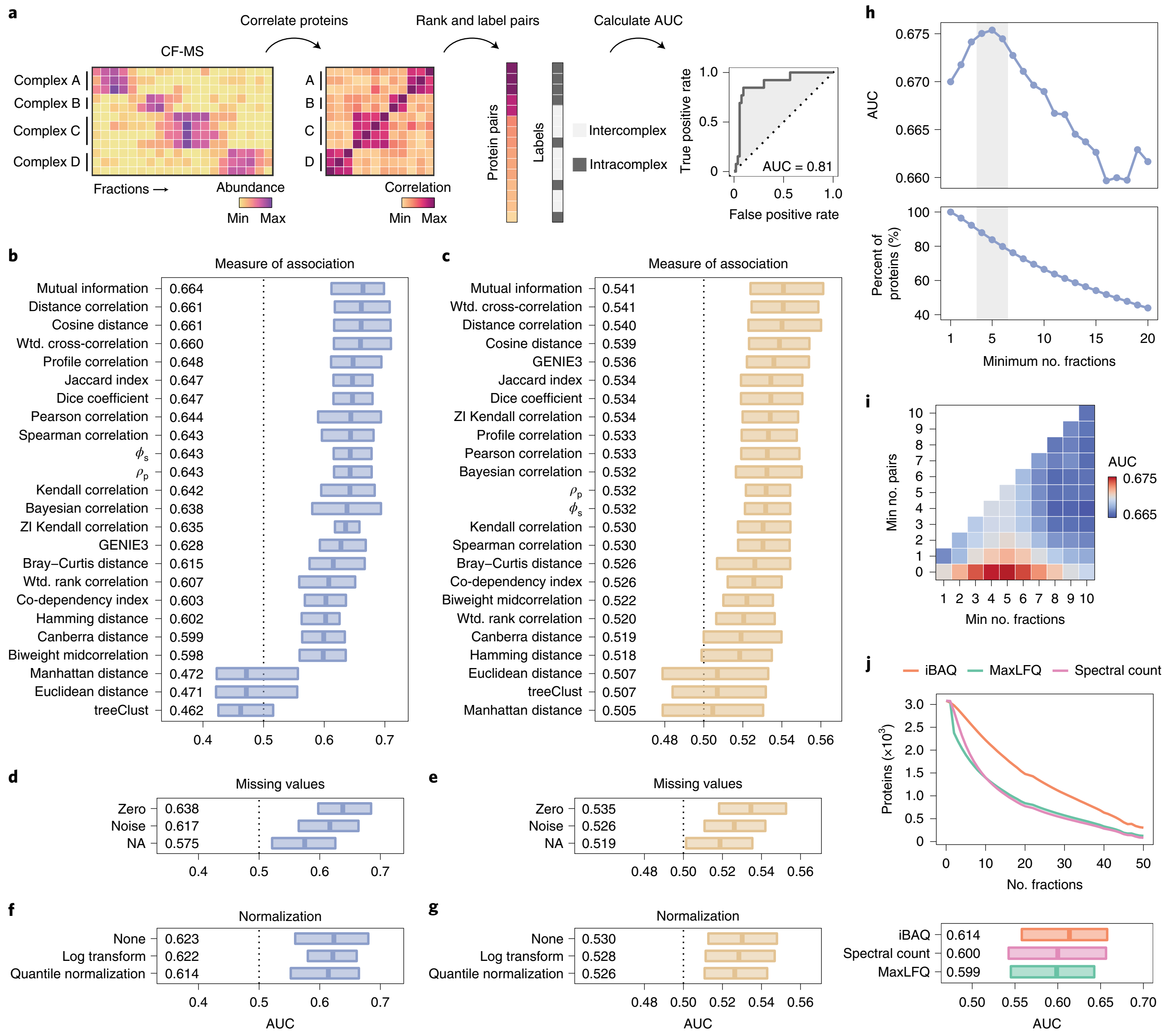


Fig. 2 | Benchmarking analysis of individual CF-MS experiments. a, Schematic overview of AUC calculation. Min, minimum; max, maximum. **b**, Recovery of known protein complexes in 67 mouse or human CF-MS datasets using 24 different measures of association. Inset text shows the median AUC across all CF-MS datasets for each measure of association. Wtd., weighted; ZI, zero-inflated. **c**, As in **b** but for proteins annotated with the same GO term in all 206 datasets. **d**, Recovery of known protein complexes when treating missing values as missing (NA), treating them as zeros, or imputing with near-zero noise. Inset text shows the median AUC across all CF-MS datasets for each treatment of missing values. **e**, As in **d** but for proteins annotated with the same GO term. **f**, Recovery of known protein complexes after log transformation, quantile normalization or no transformation of chromatograms. Inset text shows the median AUC across all CF-MS datasets for each chromatogram transformation. **g**, As in **f** but for proteins annotated with the same GO term. **h**, Recovery of known protein complexes (top) and percentage of originally quantified proteins (bottom) after filtering profiles not detected in a minimum number of fractions. **i**, Recovery of known protein complexes when filtering proteins not detected in a minimum number of fractions (*x* axis) and protein pairs not jointly detected in a minimum number of fractions (*y* axis). **j**, Mean number of protein groups identified (top) and recovery of known protein complexes (bottom) for three approaches to label-free quantification implemented in MaxQuant. Inset text shows the median AUC across all label-free CF-MS datasets for each approach to protein quantification.

protein complexes, as the ground truth. In lieu of membership in the same protein complex, we tested the ability of each metric to resolve protein pairs annotated with the same GO term, thereby allowing us to analyze all 206 datasets. Similar trends were apparent in this analysis, albeit with lower AUCs overall, as expected (Fig. 2c, Extended Data Fig. 3b and Supplementary Tables 2b and 3b). Notably, there was substantial variability in the top-performing methods within any individual dataset (Extended Data Fig. 2b). This observation suggests that small-scale analyses may fail to identify universally optimal methodologies and underscores the value of a comprehensive analysis.

Some published approaches deconvolve protein chromatograms into individual peaks, either to complement whole-chromatogram similarities or as the primary basis for analysis[24,25]. Peak-centric approaches may be required to resolve interactions between proteins that participate in multiple complexes. However, by considering fewer fractions, they necessarily lower the amount of evidence required to identify co-eluting proteins, which will expectantly increase the false positive rate. To quantify this trade-off, we compared whole-chromatogram similarities with a peak-centric approach in a subset of CF-MS datasets. The performance of the peak-centric approach was comparable to the Pearson correlation between whole chromatograms, suggesting an acceptable trade-off between false positives and false negatives, but was outperformed by several measures of whole-chromatogram similarity (Extended Data Fig. 2c–e). These observations suggest that, while deconvolution-based approaches may complement whole-chromatogram similarities, relying exclusively on individual peaks to score interactions entails an unnecessary loss of overall accuracy.

We next considered the effects of several preprocessing operations on protein complex inference. Missing values are ubiquitous in CF-MS data and can either be treated as such, replaced with zeros or imputed with near-zero noise[24]. We observed a marked increase in the AUC when treating missing values as zeros, suggesting that these overwhelmingly correspond to truly absent proteins (Fig. 2d,e, Extended Data Fig. 3c,d and Supplementary Table 3c,d). Similarly, some analysts have elected to normalize protein chromatograms to compensate for differences in absolute protein abundance. However, no increase in performance was observed after log transformation or quantile normalization of protein chromatograms, suggesting that normalization is generally unnecessary (Fig. 2f,g, Extended Data Fig. 3e,f and Supplementary Table 3e,f).

We also considered the possibility that these decisions might interact combinatorially. For instance, measures of association that assume a bivariate normal distribution might perform well only with quantile-normalized chromatograms. We therefore enumerated all 163 valid combinations of measures of association, missing-value handling and normalization strategies. The top-performing combinations were broadly consistent with the results of the joint analysis (Extended Data Figs. 2h,i, 3g,h and 4a,b and Supplementary Table 3e,f). However, a number of statistically significant interactions were observed (Supplementary Fig. 1 and Supplementary Table 4). For instance, the Euclidean distance achieved better-than-random performance only after log transformation of protein abundance, indicating that this metric is indeed capable of resolving protein complexes but only under specific conditions.

Computational pipelines for CF-MS data typically also include a step to filter low-quality chromatograms, with a range of more lenient to more stringent approaches proposed. We investigated the impact of removing chromatograms with less than some minimum number of observations. The AUC reached a peak when filtering proteins detected in less than four to five fractions, a threshold at which almost 90% of quantified proteins were retained (Fig. 2h and Extended Data Fig. 2f), suggesting that co-eluting protein complexes can be resolved using relatively few measurements. We also performed a similar analysis after filtering protein pairs not jointly detected in a minimum number of fractions but observed no improvement in performance (Fig. 2i).

Last, because most studies to date have employed label-free approaches to protein quantification (Extended Data Fig. 1a), we compared the three modes of label-free quantification implemented within MaxQuant. Surprisingly, the performance of the more sophisticated MaxLFQ normalization algorithm[26] was inferior to the much simpler paradigm of spectral counting[27] (Fig. 2j, Extended Data Figs. 2g and 3i,j and Supplementary Table 3i,j). This may suggest that the patterns of interest in CF-MS data are sufficiently coarse that relatively little quantitative resolution is required. Alternatively, the assumption of MaxLFQ that most proteins change minimally in abundance between fractions may render it inapplicable to CF-MS data. The iBAQ algorithm yielded the highest AUC, while also quantifying the greatest number of proteins. This finding suggests that iBAQ should be the method of choice for label-free CF-MS data, at least among those implemented in MaxQuant, although other approaches to summarizing peptide-level intensities have been described[28,29].

In sum, through a meta-analysis of over 200 CF-MS experiments, these analyses establish best practices for protein complex inference from individual CF-MS datasets, including label-free protein quantification, quality control, normalization, preprocessing and scoring interacting protein pairs.

Design of CF-MS experiments. Next, we turned our attention to the design of CF-MS experiments. We began by considering two of the most immediate questions that confront any investigator wishing to carry out a CF-MS study: how many biological replicates to collect and how many fractions to collect from each replicate. To address the latter question, we downsampled chromatograms from published datasets to a fixed number of fractions. As expected, the downsampling analysis indicated that collecting more fractions yielded a higher AUC (Fig. 3a). However, the rate of this increase plateaued rapidly with the addition of new fractions. Only marginal improvement was observed with more than ~40 fractions per replicate. This is remarkable, given that over 80% of experiments conducted to date have collected more than 40 fractions (Supplementary Table 1). We observed similar trends when varying the measure of association, using GO as the ground truth, dividing experiments by separation method and when sampling windows of adjacent fractions (Extended Data Fig. 5).

This finding implied that, rather than deeply profiling a small number of replicates, investigators with a fixed budget of mass spectrometry resources should consider profiling many replicates with fewer fractions. To quantitatively assess this trade-off, we repeated our downsampling analysis with fractions sampled from between one and five different replicates. Collecting additional replicates increased the AUC much faster than collecting additional fractions (Fig. 3b and Extended Data Fig. 6). Moreover, the effect had not saturated even with five biological replicates, suggesting that network inference overwhelmingly benefits from collecting many independent pictures of the same biological system, rather than a smaller number of high-resolution pictures.

Although most published CF-MS studies have employed label-free approaches to protein quantification, a handful have used metabolic or chemical labeling strategies[8,17,30–33], among which stable isotope labeling using amino acids in cell culture (SILAC) has predominated. For the subset of SILAC datasets, we compared protein complex recovery using SILAC ratios to the iBAQ intensities from individual isotopolog channels. As expected, SILAC labeling did indeed increase the AUC, but the effect was modest and variable, falling short of statistical significance ($P=0.17$, paired t-test; Fig. 3c and Extended Data Fig. 7a). Conversely, the requirement of protein detection in both isotopolog channels substantially decreased the number of proteins that could be quantified ($P=1.07\times10^{-10}$; Fig. 3c and Extended Data Fig. 7a). Thus, the increase in quantitative accuracy afforded by metabolic labeling in CF-MS may not justify the concomitant loss of proteome coverage.

Finally, we asked whether our retrospective analysis provided strong support for any particular fractionation technique. Size exclusion chromatography (SEC) and native polyacrylamide gel electrophoresis (N-PAGE) emerged as the top-performing methods, with the former also affording the greatest proteome coverage (Fig. 3d, Extended Data Fig. 7b–h and Supplementary Table 5). Isoelectric focusing and ion-exchange chromatography (IEX) exhibited less ability to resolve known protein complexes. We obtained similar results when computing the AUC for each individual complex in turn, instead of over all intracomplex and intercomplex interactions jointly (Extended Data Fig. 7i). However,

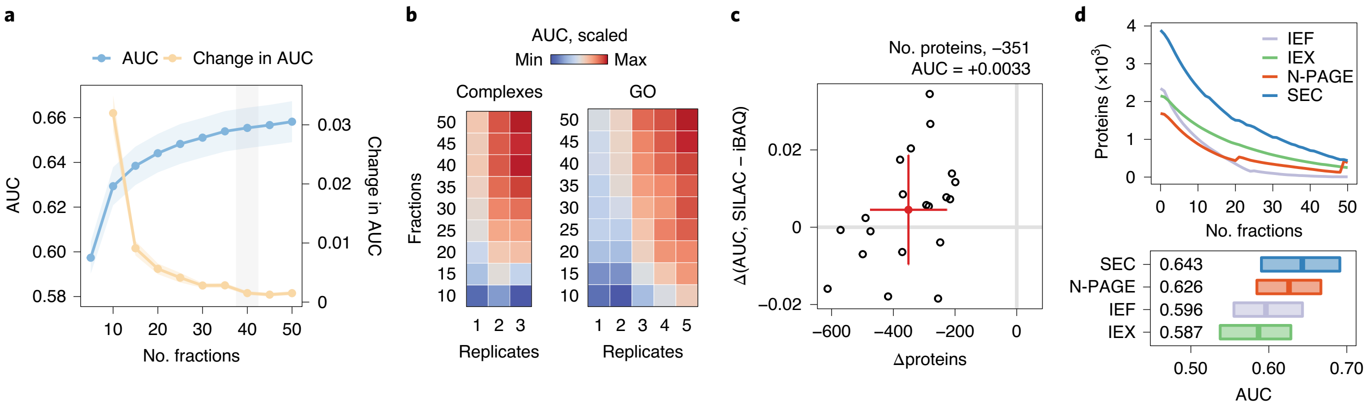


Fig. 3 | Design of CF-MS experiments. a, Recovery of known protein complexes after downsampling CF-MS chromatograms to a fixed number of fractions. Shaded area shows the standard error. **b**, Recovery of known protein complexes (left) and proteins annotated with the same GO term (right) in downsampled CF-MS chromatograms with fractions sampled from between one and five replicates. **c**, Comparison of protein complex recovery and proteome coverage between SILAC ratios and iBAQ intensities from individual isotopolog channels in 20 SILAC-labeled datasets. Caption and error bars show mean and s.d. of the difference in the number of protein groups quantified and the AUC between SILAC ratios and iBAQ intensities. **d**, Mean number of protein groups quantified (top) and recovery of known protein complexes (bottom) in published CF-MS experiments grouped by fractionation method. Inset text shows median AUCs for each fractionation method.

many complexes were consistently resolved best by specific techniques, suggesting substantial complementarity between methods (Extended Data Fig. 7j,k).

In sum, these analyses reveal best practices for the design of CF-MS studies, emphasizing the importance of biological replication and supporting SEC and N-PAGE as chromatographic approaches.

Integration of CF-MS replicates. Our analyses to this point have focused on individual CF-MS experiments. In practice, however, investigators generally seek to combine information from multiple CF-MS replicates during network inference. Supervised machine learning has emerged within the field as the standard strategy to this end. In this paradigm, a classifier is trained to identify interacting protein pairs, using features computed from each replicate as input and a training set constructed from known protein complexes. The classifier is trained in cross-validation to avoid leaking information between the training and test data and to allow for the possibility that some known complexes may not be assembled in a given dataset. This workflow entails a number of analytical decisions, including the structure of the cross-validation procedure, the number and identities of features computed for each replicate and the choice of classifier. We sought to dissect the contribution of each decision to network inference.

We first contemplated the design of the cross-validation procedure itself. Protein complexes can be split into folds that are disjoint with respect to either pairwise interactions or individual protein subunits, as illustrated in Extended Data Fig. 8a. In human CF-MS datasets, splitting by protein pairs consistently yielded an improved AUC compared to splitting by proteins in cross-validation, but this improvement vanished in a held-out set of protein complexes (Fig. 4a and Extended Data Fig. 8b). This observation strongly suggests that splitting by protein pairs leads to inflated estimates of network quality in cross-validation. We thus split the reference complexes by proteins for all remaining experiments.

We next sought to confirm that our comparison of measures of association in individual datasets (Fig. 2) would generalize to the integration of multiple CF-MS datasets. We performed network inference from combinations of two to six CF-MS datasets, using features derived from each of the 24 measures of association in turn. Measures of association exhibited similar performance in the single-dataset and multi-dataset settings, albeit with some discrepancies: for instance, our single-dataset analysis appeared to overestimate the performance of the mutual information, while underestimating that of the Pearson correlation (Extended Data Fig. 9a).

We also asked whether we could identify combinations of features with non-additive (that is, synergistic or antagonistic) contributions to classifier performance. We trained classifiers on all possible pairs of features and identified several measures of association with reproducible synergistic or antagonistic interactions (Extended Data Fig. 9b–f). Many of the synergistic interactions combined measures of correlation in protein abundance with metrics based on protein co-occurrence in overlapping fractions, suggesting that these two categories of features provide largely complementary sources of information. We also identified a strong synergistic interaction between the Pearson correlation and the Euclidean distance, which may explain the success of previous studies that have employed both measures for network inference (Extended Data Fig. 2a), despite the Euclidean distance itself yielding rankings that are little better than random (Fig. 2b).

To better understand the impact of the features provided to the classifier as input, we compared classifiers trained on either a selection of the top-performing features in individual replicates (Extended Data Fig. 4) or an equivalent number of randomly selected features. As expected, classifiers trained on top-performing features exhibited superior performance (Fig. 4b and Extended Data Fig. 8c). However, the difference was attenuated with large numbers of features or datasets when using a random forest classifier (Extended Data Fig. 8d), implying that, in data-rich regimes, more powerful classifiers can recover the same biological signal from noisier inputs. Interestingly, classifiers trained on the features used by PrInCE[24] achieved performance comparable to that of those trained on an equivalent number of top-performing features, suggesting that the precise identity of the features is not critical for network inference as long as a sufficient number of reasonably discriminative features are provided.

We also asked whether improved performance could be achieved by merging all CF-MS experiments into a single combined matrix before calculating features but found that this had a uniformly negative effect (Extended Data Fig. 8e). Calculating features separately for each replicate may allow the classifier to assign higher weights to higher-quality datasets while downweighting noisy data. We additionally investigated the effect of imputing missing values in

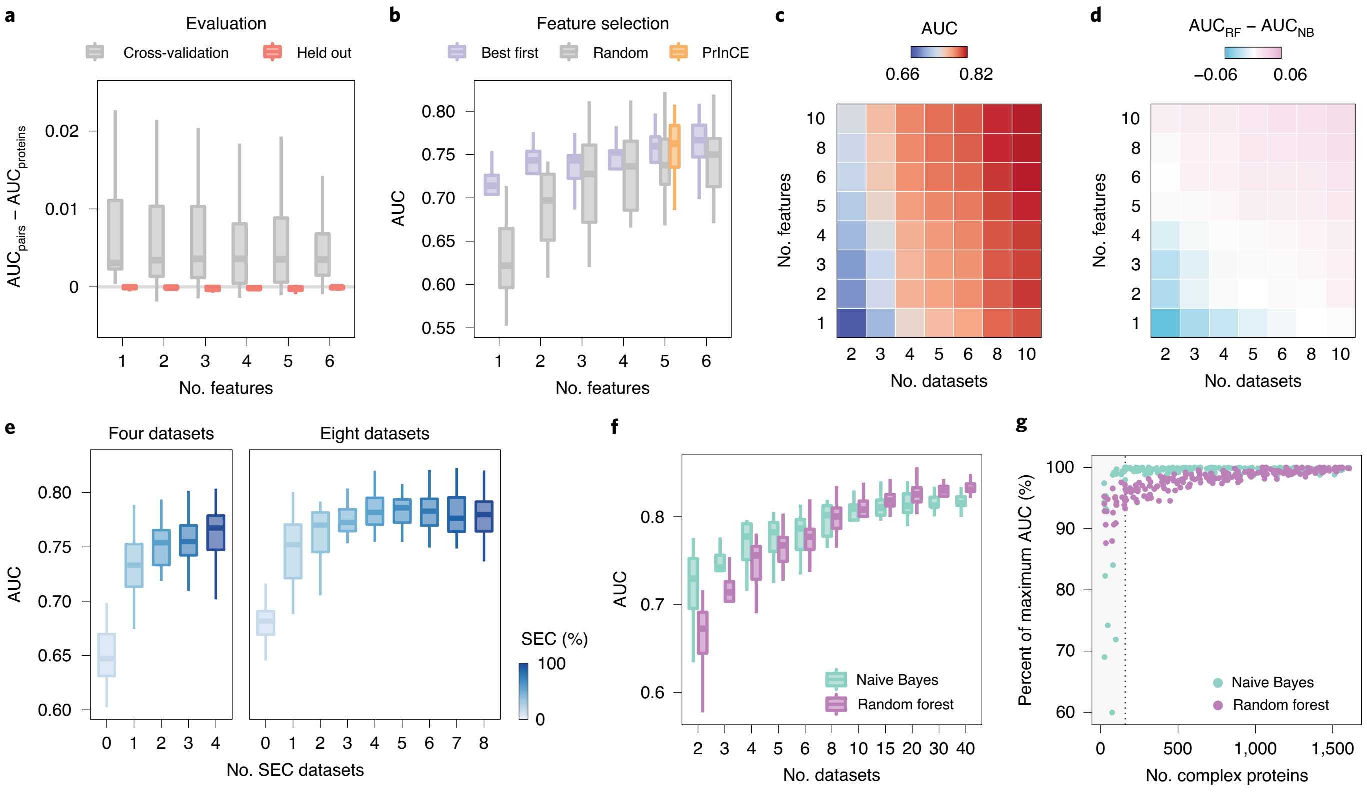


Fig. 4 | Network inference from multiple CF-MS replicates. a, Comparison of cross-validation by protein pairs or individual proteins in network inference from three CF-MS experiments, with AUCs calculated in cross-validation or an independent set of held-out protein complexes. Box plots show $n=10$ independent samples. **b**, Impact of feature selection on network inference from three CF-MS experiments using a random forest classifier, comparing one to six top-performing features, an equivalent number of random features or five features computed in PrInCE[24]. Box plots show $n=10$ independent samples. **c**, Impact of the number of top-performing features provided to a random forest classifier on network inference from two to ten CF-MS experiments. **d**, Comparison of random forest (RF) and naive Bayes (NB) classifiers in network inference from two to ten CF-MS replicates, using one to ten top-performing features. **e**, Network inference from varying ratios of SEC and IEX data. The total number of datasets being integrated is shown above the plots, and the number of SEC datasets is shown on the *x* axis. Box plots show $n=20$ independent samples. **f**, Saturation analysis of network inference from 2–40 CF-MS experiments, using a single top-performing feature. Box plots show $n=10$ independent samples. **g**, Impact of downsampling training set complexes on network inference from three CF-MS experiments. Box plots show the median (horizontal line), interquartile range (hinges) and smallest and largest values no more than 1.5 times the interquartile range (whiskers).

the feature matrix, which commonly occur when a protein is not quantified in one or more replicates. Although some classifiers can naturally handle missing values, median imputation consistently improved performance (Extended Data Fig. 8f).

Next, we examined the number of features computed for each replicate in more detail. Providing more features to the classifier increased the AUC, but the effect saturated rapidly, with only marginal increases above five features (Fig. 4c and Extended Data Fig. 8g). Moreover, the effect diminished as the number of replicates increased. With many replicates, a single top-performing feature was sufficient to achieve nearly optimal performance, and the total number of replicates had a much stronger impact on network inference. Interestingly, the optimal choice of classifier was dependent on the total amount of data available (Fig. 4d and Extended Data Fig. 9h). With many replicates, the random forest performed best and was largely insensitive to the identity of features provided as input. Conversely, in smaller collections of CF-MS data (for example, with two to three replicates), the simpler naive Bayes classifier yielded the highest AUC, but only when trained on optimal features. These observations imply that different strategies should be applied to network inference from small-scale CF-MS studies (for instance, with two to four replicates) and large compendia of CF-MS data.

Our analyses of individual datasets suggested that, of the two main approaches for cellular lysate fractionation that have been applied to human cells (that is, SEC and IEX), SEC generally achieved better resolution of known protein complexes. However, some studies have argued that multiple, orthogonal separations are necessary to minimize chance co-elution[9,11]. To test this hypothesis, we inferred networks from combinations of SEC and IEX datasets in varying ratios (Fig. 4e and Extended Data Fig. 8i). When integrating five or fewer datasets, network inference was optimized when using exclusively SEC data. However, with six or more datasets, a combination of SEC and IEX datasets was necessary to maximize the AUC. This observation supports the view that large-scale interactome mapping projects must integrate multiple fractionation approaches to achieve optimal accuracy.

One of the most promising applications of CF-MS is to map the interactomes of species not amenable to conventional techniques such as AP–MS or Y2H. However, it is currently unclear how many CF-MS experiments are needed to achieve saturating coverage of a given interactome. To address this question, we inferred networks from random samples of 2–40 human CF-MS datasets (Fig. 4f and Extended Data Fig. 10a). Performance saturated after approximately 15–20 datasets, suggesting that roughly a dozen CF-MS experiments are sufficient to produce an initial interactome map for a given organism. A related issue that may arise in less-studied organisms is that relatively few protein complexes are known, limiting the size of the training set. To evaluate the impact of training set size,

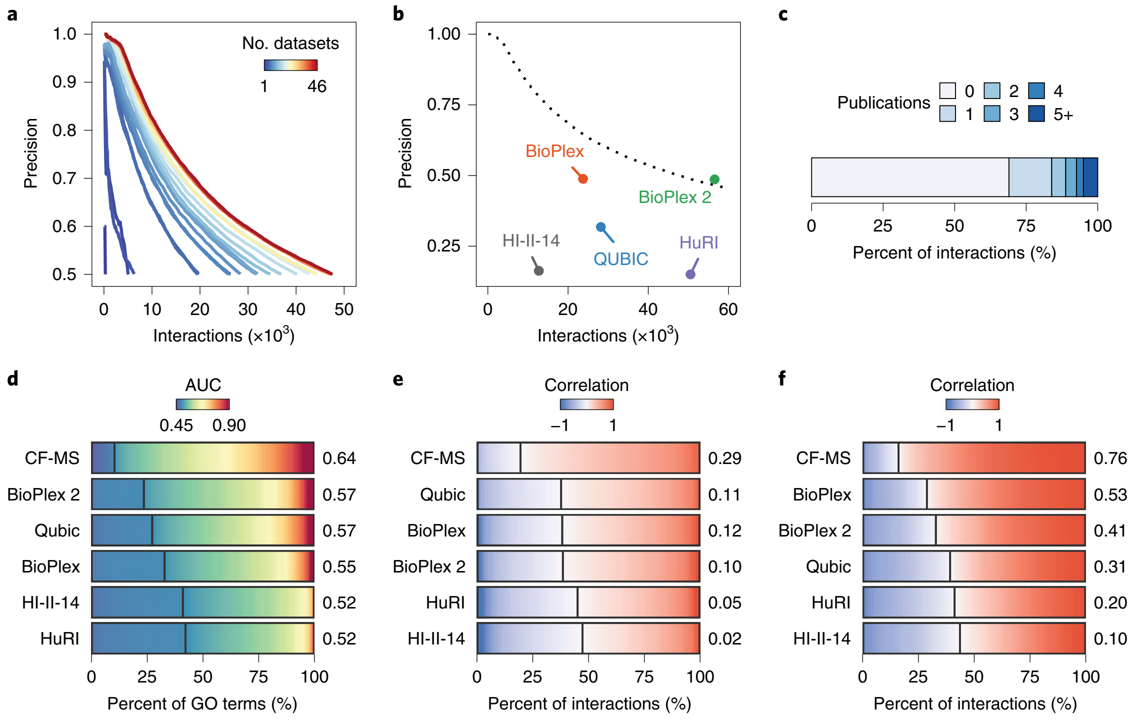


Fig. 5 | Meta-analysis defines a consensus human CF-MS interactome. a, Precision–recall curves for human interactomes inferred from random samples from 1–46 CF-MS experiments. **b**, Precision and recall of the consensus human CF-MS interactome (dotted line) and five recent high-throughput human interactome screens. **c**, Proportion of known interactions in the consensus human CF-MS interactome. **d**, Functional coherence of the consensus human CF-MS interactome and five recent high-throughput screens, as quantified by the AUC of protein-function prediction[36]. Vertical lines indicate the proportion of GO terms with AUC less than 0.5, equivalent to random chance. Text indicates median AUCs across all GO terms. **e**, Coexpression of interacting protein pairs across 294 biological conditions in the ProteomeHD resource[20]. Vertical lines show the proportion of negatively correlated pairs[24]. Text indicates median Pearson correlation coefficients across all interacting pairs. **f**, Colocalization of interacting protein pairs by subcellular proteomics[38]. Text indicates median Pearson correlation coefficients across all interacting pairs.

we performed network inference after downsampling the CORUM database. The performance of naive Bayes classifiers saturated rapidly, with the AUC reaching 98% of its maximum value with only 154 proteins in the training set (Fig. 4g and Extended Data Fig. 10b). By contrast, random forest classifiers required 904 proteins to reach the same value. This finding supports the notion that more powerful nonlinear classifiers are most useful in data-rich settings.

In sum, these analyses establish an optimal protocol for integration of data from multiple CF-MS replicates. Notably, our results expose two distinct regimes in network inference from CF-MS data and reveal different optimal workflows for low-data and data-rich scenarios.

A consensus human CF-MS interactome. Having assembled a comprehensive resource of CF-MS data and established an optimal workflow for data analysis, we next asked whether meta-analysis of all published datasets could produce a draft-quality map of the human interactome. Integration of all 46 published human CF-MS experiments, using the optimized methodology for network inference defined by our benchmarks, identified a total of 47,575 interactions at 50% precision (Fig. 5a and Supplementary Fig. 2a,b). To place this performance in context, we calculated the precision and recall of five recent large-scale human interactome screens using AP–MS or Y2H techniques[1–5]. Evaluating these screens on identical sets of true positive and true negative interactions, we found that our meta-analysis recovered more interactions than all but one of the five screens at equivalent precision and 90% as many as the exception, the BioPlex 2 network (Fig. 5b). Reassuringly, 31% of the interactions in the consensus CF-MS interactome had been previously identified by at least one small-scale or high-throughput experiment, confirming its ability to recapitulate known biology (Fig. 5c). However, the remaining 69% were not found in any protein–protein interaction database, indicating that large-scale integration of CF-MS data can expand even well-studied interactomes. Similarly, our consensus CF-MS interactome overlapped significantly with networks derived from the integration of CF-MS data with other proteome-scale datasets[9,34,35] ($P<10^{-15}$, hypergeometric test), but the majority of interactions were again new (Supplementary Fig. 2c–e). The observed degree of overlap is consistent with the fact that these previous efforts drew on largely or entirely distinct collections of CF-MS data, in addition to numerous external sources of information (Supplementary Fig. 2f).

To evaluate the biological relevance of the consensus CF-MS interactome, we asked to what extent proteins implicated in the same biological processes tended to physically interact. Remarkably, the functional coherence of the CF-MS interactome, as quantified by the AUC[36] (Methods), was substantially higher than that of human interactomes derived from AP–MS or Y2H (Fig. 5d). Similarly, we evaluated the degree to which interacting protein pairs tend to display correlated patterns of abundance in large-scale proteomic datasets or colocalize to the same subcellular compartments, observing excellent performance by both measures[20,37–39] (Fig. 5e,f and Supplementary Fig. 2g,h). In sum, these analyses provide strong support for the notion that large-scale integration of CF-MS data can produce interactome maps of quality comparable to or higher than that of systematic AP–MS or Y2H screens, and define a draft-quality map of the human interactome by CF-MS.

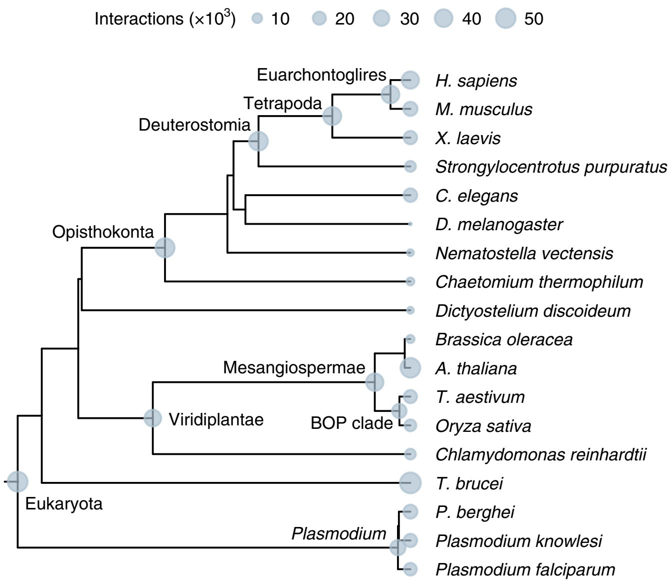


Fig. 6 | CF-MS interactomes throughout the eukaryotic evolutionary tree. Number of protein–protein interactions predicted at 50% precision for 18 species profiled by at least three published CF-MS experiments or nine selected phylogenetic clades spanning 1.8 billion years of eukaryotic evolution.

CF-MS interactomes throughout the evolutionary tree. The success of our efforts to reconstruct a draft human interactome through meta-analysis of published CF-MS experiments compelled us to ask whether this approach would extend to the other species represented in our compendium. Accordingly, we performed additional meta-analyses of CF-MS interactomes for 17 species profiled by at least three experiments. This approach identified between 1,595 (*Drosophila melanogaster*) and 57,682 (*Trypanosoma brucei*) interactions in each species (Fig. 6). For many of these species, these networks represent the first attempts to assemble systematic, unbiased interactome maps, complementing literature-curated resources derived from small-scale experiments[40].

We also asked whether we could combine information from multiple species to infer consensus interactomes for entire phylogenetic clades, such as mammals, deuterostomes or even all eukaryotes. We mapped proteins to orthogroups and applied an identical approach to infer networks for nine internal nodes in the phylogenetic tree (Fig. 6). Meta-analysis of up to 199 CF-MS experiments inferred between 24,890 (BOP clade) and 51,998 (Eukaryota) evolutionarily conserved interactions in each clade. Both the functional coherence of the networks and the total number of interactions that could be confidently inferred appeared to saturate with the addition of new species and experiments, raising the possibility of an upper bound to the performance of cross-species interactome mapping under current CF-MS experimental workflows (Supplementary Fig. 3a,b).

To probe the evolutionary conservation of the predicted interactions, we computed the overlap between the consensus CF-MS interactome predicted for humans and their parent clades, from placental mammals to eukaryotes. As expected, the conservation of human interactions decreased over evolutionary time (Supplementary Fig. 3c). However, almost two-thirds (65.6%) of human interactions were conserved in the eukaryote meta-interactome. By contrast, 18.9% of human interactions were not conserved in the Euarchontoglires network; some of these may represent candidate human-specific interactions. We observed similar trends for the mouse and *Arabidopsis* interactomes (Supplementary Fig. 3c). Thus, our 27 eukaryotic CF-MS interactomes collectively provide a hypothesis-generating resource to investigate network evolution.

Discussion

CF-MS has dramatically increased the throughput of interactome mapping, but there is little agreement within the field on the best way to carry out an experiment or analyze the resulting data. Here, we have carried out a comprehensive reanalysis of more than 200 CF-MS experiments, assembling one of the largest collections of uniformly processed proteomic data in existence. We then used this resource to systematically optimize both experimental and computational workflows for CF-MS. We provide an interactive web application to facilitate exploration of the complete dataset, available at http://cf-ms-browser.msl.ubc.ca.

A number of specific recommendations for the design of CF-MS experiments emerge from our analysis. In several cases, these recommendations deviate from what is currently standard practice in the field. One particularly striking finding is that little benefit is conferred by collecting more than 40 fractions from any given biological replicate. On the other hand, collecting additional biological replicates had a profound impact on network inference. These findings are noteworthy, given that the dominant paradigm to date has been to collect many fractions—in some cases more than 100—from a relatively small number of replicates. Our results suggest a shift away from this paradigm: that instead collecting lower-resolution profiles from many independent samples could allow higher-confidence network inference at equivalent cost. Surprisingly, our comparison of approaches to protein quantification indicated that the improved resolution afforded by SILAC labeling did not markedly improve network inference and was counterbalanced by a decrease in proteome coverage. In this respect, it is important to note that our analysis focused on interactome mapping from CF-MS under a single condition. Metabolic or chemical labeling strategies have distinct advantages for mapping rearrangements in the interactome through comparative CF-MS studies. Notably, by enabling sample multiplexing during chromatography, labeling strategies provide the only means to decouple technical variability associated with protein complex elution from biological variability in protein complex assembly. Last, our comparison of approaches to cellular lysate fractionation has the important caveat that each dataset cannot be used as its own internal control, as in the downsampling and protein-quantification analyses. We therefore cannot exclude the possibility that the observed trends are confounded by differences in the biological systems under investigation or technical differences in the mass spectrometry. Nonetheless, this comparison provides some level of evidence that SEC or N-PAGE should be preferentially considered as methods for protein complex separation. N-PAGE is particularly advantageous for the study of membrane interactions[31,41], although published N-PAGE datasets have achieved lower proteome coverage than those employing SEC, perhaps due to lower sample recovery. On the other hand, each separation method was able to best resolve at least a handful of known protein complexes, and integration of different separation methods was necessary to optimize network inference from collections of six or more CF-MS datasets. These findings indicate that the approaches to fractionation in current use yield at least partially complementary pictures of protein complex assembly.

The biological and technological heterogeneity of the resource we assembled also provided an ideal testbed for data analysis strategies. To date, published comparisons of computational tools for CF-MS data have compared entire pipelines as distributed[12–14]. Here, instead of treating these tools as black boxes, we sought to break them down into their constituent operations, in order to distinguish successful from unsuccessful strategies at each step of analysis in turn. This effort allowed us to systematically define the components of an optimal pipeline for network inference from

CF-MS data. Notably, we identified two distinct regimes of classifier performance in CF-MS data integration. In the low-data regime, simple linear classifiers achieved the best performance but were highly sensitive to the features provided as input. Conversely, in the data-rich regime, nonlinear classifiers dominated and were insensitive to feature selection. We provide an R package implementing all of the methods evaluated in this study, available from GitHub at https://github.com/fosterlab/CFTK. CFTK complements existing, 'one-size-fits-all' approaches to supervised analysis of CF-MS data, such as PrInCE[24] and EPIC[14], by providing a flexible toolkit that can be used to implement analytical workflows tailored to the data at hand. However, a limitation of CFTK is that this greater flexibility comes at the cost of a somewhat higher barrier to entry than that of existing pipelines.

We applied the optimized computational methodology that emerged from our systematic benchmarks to the corpus of reanalyzed CF-MS data and predicted consensus CF-MS interactomes for 27 species or clades. Using a precision cutoff of 50%, we generated a resource comprising more than 700,000 predicted interactions. Taken at face value, this threshold implies that up to half of all interactions in these networks represent false positives. However, we believe that this is a highly conservative estimate of network quality. Like others in the field[10,14], we treat pairs of CORUM proteins found in different complexes as true negatives. The number of true negative pairs thus grows quadratically with the number of proteins in CORUM, and consequently these outnumber true positives by a large margin. Moreover, at least some of these ostensibly true negative pairs are likely to truly interact, in pairwise interactions or protein complexes that are as-of-yet undiscovered or simply missing from the CORUM database. Estimating the absolute error rate of protein–protein interaction networks is a difficult problem that has attracted extensive discussion[42,43]. In view of these caveats, we caution that the precision is unlikely to provide an accurate estimate of the absolute error rate of either our consensus CF-MS interactome or published networks. Instead, we believe that the most relevant outcome of our analysis is instead the relative performance of the consensus human CF-MS interactome in comparison with that of other large-scale screens. When evaluated on the same terms, we found that integration of CF-MS experiments recovered more interactions at equivalent precision than all but one of these, the BioPlex 2 network, which integrated almost 6,000 AP–MS experiments. However, while our computational analyses substantiate the quality of the inferred networks at a systems level, we also caution users of this resource that any particular individual interaction should be regarded as putative until confirmed experimentally using an orthogonal technique. Such experimental validation may be challenging in biological systems not amenable to conventional approaches, such as affinity purification. Looking toward the future, integration of CF-MS with other mass spectrometric assays, such as cross-linking mass spectrometry or thermal proteome profiling, may provide a means to increase confidence in individual interactions outside of well-studied model systems[11].

Our analysis framework closely followed what has emerged as the standard workflow for CF-MS data analysis, in which supervised machine learning is applied to predict a sparse, unweighted interaction network from CF-MS data. Two deviations from this workflow are worth noting. First, previous efforts to integrate large compendia of CF-MS data[9,10,34,35] have applied graph-based clustering algorithms to the inferred networks, in order to assemble pairwise interactions into multi-protein complexes. We recently reported that these clustering algorithms are highly sensitive to small amounts of noise in the network[44]. In extreme cases, a minute perturbation to the underlying interaction network could produce a ~50% change in the complexes detected. We observed similar results across many different interaction networks and clustering algorithms, suggesting that this instability is a fundamental property of these approaches. Moreover, we observed similar instability when injecting noise directly into the underlying chromatograms. Because the process of CF-MS data acquisition is inherently noisy, we have opted not to perform such a clustering analysis here. However, our resource of predicted interactomes provides a fertile ground to develop and evaluate new, more robust clustering approaches for protein complex detection. Second, recent studies have proposed novel analytical frameworks, based either on hierarchical clustering to infer protein complexes directly from elution profiles without an intermediate network-inference step[13,45] or on the targeted investigation of known complexes, eschewing de novo inference of interactions or complexes entirely[25,46]. Because these approaches rely on many of the same basic operations evaluated here (for instance, preprocessing elution profiles and scoring their similarity), our findings should also apply to the development of these and related frameworks. More broadly, we anticipate that our resource of systematically reanalyzed CF-MS data can help spur the development of new, 'network-free' approaches to protein complex inference.

Our study defines a number of resources for the CF-MS and broader proteomic communities. First, our compendium of uniformly processed CF-MS data provides an unprecedented platform to develop and benchmark new tools for network inference. Whereas previous evaluations of computational strategies for CF-MS data have considered at most a handful of datasets[12–14,47,48], this data will empower developers to test their tools in a large number of datasets and thereby increase rigor within the field by limiting overfitting to individual datasets. More broadly, this resource can be used to test many different biological hypotheses about protein complex evolution or stoichiometry. Second, this compendium includes several comparative CF-MS studies[31,49–51], providing a platform to develop and test new computational approaches for differential analysis of CF-MS data across biological conditions. Third, because a wealth of knowledge is available about the structure of biological networks, this dataset can also provide a testbed for computational proteomics more broadly, in situations where no task-specific gold standard is available. As one example, an approach to protein inference that improves the recovery of known protein complexes is likely to also perform well in other settings. We provide complete peptide-level chromatograms for all 206 experiments in our Proteomics Identification Database (PRIDE) deposition as a resource to support the development of more accurate approaches to label-free protein quantitation[25,46] or for the identification of proteoform-specific interactions[31,52]. Finally, our efforts to infer CF-MS interactomes in 27 species or clades provide systematic protein–protein interaction maps for several understudied organisms, and a resource to understand the evolution of eukaryotic cell biology. This resource is complementary both to experimental interactome maps generated using other high-throughput techniques and efforts to computationally predict protein–protein interaction networks from sequence or structural features[53–56]. The complete resource, including all of the data from intermediate processing steps and the source code used to generate it, is available at https://fosterlab.github.io/CF-MS-analysis.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-021-01194-4.

Received: 20 November 2020; Accepted: 20 May 2021;
Published online: 01 July 2021

References

1. Rolland, T. et al. A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).
2. Luck, K. et al. A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).
3. Huttlin, E. L. et al. The BioPlex network: a systematic exploration of the human interactome. *Cell* **162**, 425–440 (2015).
4. Huttlin, E. L. et al. Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505–509 (2017).
5. Hein, M. Y. et al. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163**, 712–723 (2015).
6. Kim, Y., Jung, J. P., Pack, C.-G. & Huh, W.-K. Global analysis of protein homomerization in *Saccharomyces cerevisiae*. *Genome Res.* **29**, 135–145 (2019).
7. Werner, J. N. et al. Quantitative genome-scale analysis of protein localization in an asymmetric bacterium. *Proc. Natl Acad. Sci. USA* **106**, 7858–7863 (2009).
8. Kristensen, A. R., Gsponer, J. & Foster, L. J. A high-throughput approach for measuring temporal changes in the interactome. *Nat. Methods* **9**, 907–909 (2012).
9. Havugimana, P. C. et al. A census of human soluble protein complexes. *Cell* **150**, 1068–1081 (2012).
10. Wan, C. et al. Panorama of ancient metazoan macromolecular complexes. *Nature* **525**, 339–344 (2015).
11. McWhite, C. D. et al. A pan-plant protein complex map reveals deep conservation and novel assemblies. *Cell* **181**, 460–474 (2020).
12. Rosenberger, G. et al. SECAT: quantifying protein complex dynamics across cell states by network-centric analysis of SEC-SWATH-MS profiles. *Cell Syst.* https://doi.org/10.1016/j.cels.2020.11.006 (2020).
13. Fossati, A. et al. PCprophet: a framework for protein complex prediction and differential analysis using proteomic data. *Nat. Methods* https://doi.org/10.1038/s41592-021-01107-5 (2020).
14. Hu, L. Z. et al. EPIC: software toolkit for elution profile-based inference of protein complexes. *Nat. Methods* **16**, 737–742 (2019).
15. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).
16. Giurgiu, M. et al. CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Res.* **47**, D559–D563 (2019).
17. Skinnider, M. A. et al. An atlas of protein–protein interactions across mammalian tissues. Preprint at *bioRxiv* https://doi.org/10.1101/351247 (2018).
18. Jarzab, A. et al. Meltome atlas—thermal proteome stability across the tree of life. *Nat. Methods* **17**, 495–503 (2020).
19. Ochoa, D. et al. The functional landscape of the human phosphoproteome. *Nat. Biotechnol.* **38**, 365–373 (2020).
20. Kustatscher, G. et al. Co-regulation map of the human proteome enables identification of protein functions. *Nat. Biotechnol.* **37**, 1361–1371 (2019).
21. Mellacheruvu, D. et al. The CRAPome: a contaminant repository for affinity purification–mass spectrometry data. *Nat. Methods* **10**, 730–736 (2013).
22. Romanov, N. et al. Disentangling genetic and environmental effects on the proteotypes of individuals. *Cell* **177**, 1308–1318 (2019).
23. Skinnider, M. A., Squair, J. W. & Foster, L. J. Evaluating measures of association for single-cell transcriptomics. *Nat. Methods* **16**, 381–386 (2019).
24. Stacey, R. G., Skinnider, M. A., Scott, N. E. & Foster, L. J. A rapid and accurate approach for prediction of interactomes from co-elution data (PrInCE). *BMC Bioinformatics* **18**, 457 (2017).
25. Bludau, I. et al. Complex-centric proteome profiling by SEC-SWATH-MS for the parallel detection of hundreds of protein complexes. *Nat. Protoc.* **15**, 2341–2386 (2020).
26. Cox, J. et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).
27. Liu, H., Sadygov, R. G. & Yates, J. R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76**, 4193–4201 (2004).
28. Al Shweiki, M. R. et al. Assessment of label-free quantification in discovery proteomics and impact of technological factors and natural variability of protein abundance. *J. Proteome Res.* **16**, 1410–1424 (2017).
29. McIlwain, S. et al. Estimating relative abundances of proteins from shotgun proteomics data. *BMC Bioinformatics* **13**, 308 (2012).
30. Scott, N. E., Brown, L. M., Kristensen, A. R. & Foster, L. J. Development of a computational framework for the analysis of protein correlation profiling and spatial proteomics experiments. *J. Proteomics* **118**, 112–129 (2015).
31. Scott, N. E. et al. Interactome disassembly during apoptosis occurs independent of caspase cleavage. *Mol. Syst. Biol.* **13**, 906 (2017).
32. Pourhaghighi, R. et al. BraInMap elucidates the macromolecular connectivity landscape of mammalian brain. *Cell Syst.* **10**, 333–350 (2020).
33. Kastritis, P. L. et al. Capturing protein communities by structural proteomics in a thermophilic eukaryote. *Mol. Syst. Biol.* **13**, 936 (2017).
34. Drew, K. et al. Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Mol. Syst. Biol.* **13**, 932 (2017).
35. Drew, K., Wallingford, J. B. & Marcotte, E. M. hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. *Mol. Syst. Biol.* **17**, e10016 (2021).
36. Ballouz, S., Weber, M., Pavlidis, P. & Gillis, J. EGAD: ultra-fast functional analysis of gene networks. *Bioinformatics* **33**, 612–614 (2017).
37. Lapek, J. D. et al. Detection of dysregulated protein-association networks by high-throughput proteomics predicts cancer vulnerabilities. *Nat. Biotechnol.* **35**, 983–989 (2017).
38. Orre, L. M. et al. SubCellBarCode: proteome-wide mapping of protein localization and relocalization. *Mol. Cell* **73**, 166–182 (2019).
39. Geladaki, A. et al. Combining LOPIT with differential ultracentrifugation for high-resolution spatial proteomics. *Nat. Commun.* **10**, 331 (2019).
40. Cusick, M. E. et al. Literature-curated protein interaction datasets. *Nat. Methods* **6**, 39–46 (2009).
41. Heide, H. et al. Complexome profiling identifies TMEM126B as a component of the mitochondrial complex I assembly complex. *Cell Metab.* **16**, 538–549 (2012).
42. von Mering, C. et al. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417**, 399–403 (2002).
43. Venkatesan, K. et al. An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83–90 (2009).
44. Stacey, R. G., Skinnider, M. A. & Foster, L. J. On the robustness of graph-based clustering to random network alterations. *Mol. Cell. Proteomics* **20**, 100002 (2020).
45. McBride, Z. et al. A label-free mass spectrometry method to predict endogenous protein complex composition. *Mol. Cell. Proteomics* **18**, 1588–1606 (2019).
46. Heusel, M. et al. Complex-centric proteome profiling by SEC-SWATH-MS. *Mol. Syst. Biol.* **15**, e8438 (2019).
47. Salas, D., Stacey, R. G., Akinlaja, M. & Foster, L. J. Next-generation interactomics: considerations for the use of co-elution to measure protein interaction networks. *Mol. Cell. Proteomics* **19**, 1–10 (2020).
48. Pang, C. N. I. et al. Analytical guidelines for co-fractionation mass spectrometry obtained through global profiling of gold standard *Saccharomyces cerevisiae* protein complexes. *Mol. Cell. Proteomics* **19**, 1876–1895 (2020).
49. Gorka, M. et al. Protein Complex Identification and quantitative complexome by CN-PAGE. *Sci. Rep.* **9**, 11523 (2019).
50. Mallam, A. L. et al. Systematic discovery of endogenous human ribonucleoprotein complexes. *Cell Rep.* **29**, 1351–1368 (2019).
51. Drew, K. et al. A systematic, label-free method for identifying RNA-associated proteins in vivo provides insights into vertebrate ciliary beating machinery. *Dev. Biol.* **467**, 108–117 (2020).
52. Bludau, I. et al. Systematic detection of functional proteoform groups from bottom–up proteomic datasets. Preprint at *bioRxiv* https://doi.org/10.1101/2020.12.22.423928 (2020).
53. Garzón, J. I. et al. A computational interactome and functional annotation for the human proteome. *eLife* **5**, e18715 (2016).
54. Meyer, M. J. et al. Interactome INSIDER: a structural interactome browser for genomic studies. *Nat. Methods* **15**, 107–114 (2018).
55. Cunningham, J. M., Koytiger, G., Sorger, P. K. & AlQuraishi, M. Biophysical prediction of protein–peptide interactions and signaling networks using machine learning. *Nat. Methods* **17**, 175–183 (2020).
56. Hopf, T. A. et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* **3**, e03430 (2014).
57. Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D. & von Mering, C. Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* **15**, 3163–3168 (2015).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

MaxQuant searches. Through an extensive review of the literature, we identified a total of 206 published CF-MS experiments for which raw mass spectrometric data were available from proteomic repositories as of May 2020. Raw data were downloaded from the MassIVE or PRIDE repositories, and experimental designs were manually curated to group files into experiments and fractions. Instrument time for each file analyzed was calculated from Thermo RAW files using RawTools (version 2.0.2)[58] and manually retrieved from the corresponding publications for other formats. The complete list of files analyzed in this study is provided in Supplementary Table 1b.

MaxQuant (version 1.6.5.0) was used to search each experiment against the UniProt complete proteome for the corresponding species, including unreviewed accessions and isoforms, after removal of proteins less than ten amino acids long and supplementation with a list of common contaminants provided by MaxQuant. For a subset of experiments from the accession PXD009039 profiling mouse erythrocytes infected with *Plasmodium berghei*, the proteome searched was a concatenation of the *P. berghei* and mouse proteomes, with mouse proteins subsequently discarded for downstream analysis. Search parameters varied across experiments, but, in general, carbamidomethylation of cysteine was set as a fixed modification, while protein N-terminal acetylation and methionine oxidation were set as variable modifications, and trypsin cleavage was used with up to two missed cleavages. For some datasets, more specific settings were used, including replacing carbamidomethylation with *N*-ethylmaleimide on cysteines as a fixed modification, LysC cleavage or modifying the multiplicity for SILAC and dimethyl labeling experiments. Code used to download the raw data, create 'mqpar.xml' files and carry out MaxQuant searches is available from GitHub at https://github.com/skinnider/CF-MS-searches. MaxQuant outputs are available from PRIDE under the accession PXD022048.

Quality control. MaxQuant outputs ('proteinGroups.txt' files) were preprocessed by removing potential contaminants, reverse hits and proteins identified only by peptides carrying one or more modified amino acids[59]. Total numbers of tandem mass spectra and sequenced peptides were obtained from MaxQuant 'summary.txt' files. Coverage of protein complexes was assessed with respect to the core set of protein complexes from CORUM version 3.0 (file 'coreComplexes.txt') with redundant entries removed[16]. Analysis of GO terms enriched among complex proteins detected versus not detected by CF-MS was performed using the conditional hypergeometric test[60] implemented in the 'GOstats' R package[61]. Mouse and human whole-organism protein-abundance estimates in parts per million were obtained from PaxDb (version 4.1)[57]. To evaluate coverage of low-abundance proteins, we divided human proteins into three bins based on protein-abundance estimates from PaxDb. We then removed proteins not detected in any human CF-MS dataset in order to mitigate the influence of proteins not compatible with CF-MS (for instance, proteins that do not participate in any macromolecular complexes). Then, for each dataset, we computed the mean proportion of fractions in which proteins from each bin were quantified to obtain a measure of coverage for lowly, moderately and highly abundant proteins.

Analysis of individual CF-MS experiments. A total of 165 analytical pipelines were evaluated for their ability to recover known protein complexes within individual experiments from the subset of 67 human and mouse CF-MS experiments, or protein pairs annotated with the same GO term across the complete set of 206 experiments. Each pipeline consisted of a measure of association used to quantify the similarity of each chromatogram pair, a strategy for handling missing values and, optionally, a transformation or normalization of the chromatograms. The 24 measures of association evaluated here included 17 previously evaluated in the context of coexpression network inference from single-cell transcriptomics[23], as well as seven measures specifically proposed for the analysis of CF-MS data or '-omic' data more generally. These included the Bayesian correlation[62], Bray–Curtis distance and weighted cross-correlation, all of which have been employed to analyze CF-MS data in previous studies; unsupervised machine learning methods including treeClust and GENIE3 (ref. [63]), which have been proposed for protein co-regulation analysis;[20] the profile correlation, which has been applied to analyze AP–MS data[5]; and the distance correlation[64], which has been proposed for the analysis of high-throughput datasets more generally[65,66]. The 17 previously evaluated metrics were implemented as in the 'dismay' R package[23]. The Bayesian correlation was calculated using the R script accompanying the original publication[62]. The Bray–Curtis distance was calculated using the 'vegan' R package[67]. The weighted cross-correlation was calculated using the 'wcc' function from the 'wccsom' R package as previously described[9]. The distance correlation was calculated using the 'Pigengene' R package[68]. Missing values were either treated as missing (NAs), treated as zeros, or imputed with random, near-zero noise using the 'clean_profile' function from the 'PrInCE' R package[24]. Chromatograms were optionally log transformed or quantile normalized using the 'normalize.quantiles' function from the 'preprocessCore' R package. A subset of combinations were discarded involving metrics that were unable to handle missing values, such as GENIE3, or for which zeros and missing values were interpreted identically, such as the Jaccard index.

To evaluate the ability of each approach to recover known protein complexes, we implemented a framework based on ROC analysis[22]. We focused our evaluation at the level of pairwise interactions, rather than at the level of protein complexes, because we have found that the graph-based clustering approaches used to infer protein complexes from pairwise interactions are highly sensitive to noise at both the protein chromatogram and interaction network levels[44]. By comparison, inference of pairwise interactions was substantially more robust. We computed the area under the ROC curve (AUC), which reflects the probability that any arbitrary true interaction will be ranked higher than any arbitrary non-interacting pair. An AUC of 0.5 therefore reflects random guesses, with true interactions and non-interacting pairs ranked equally, while an AUC of 1.0 reflects perfect discrimination of interacting and non-interacting pairs. We used protein complexes from the CORUM database[16] to label true positive and true negative pairs. Protein pairs in the same complex were labeled as true positives, and protein pairs in different complexes were labeled as true negatives. Protein groups mapping to more than one gene symbol and proteins not belonging to any complex were discarded for this analysis. Notably, while some protein pairs in different complexes in the CORUM database may in fact represent truly interacting pairs, we reasoned that this definition of true negatives would be consistent with those employed by most studies in the field of CF-MS to date and would provide the most representative and unbiased collection of true negatives given our inherently incomplete understanding of non-interacting protein pairs across the human proteome. Alternatively, membership in the same GO slim term was used to label positive pairs. GO annotation files for UniProt proteomes were obtained from the GOA FTP server (http://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/). The three root nodes (biological process, molecular function and cellular compartment) were removed, and the remaining GO terms were filtered to exclude terms annotated to fewer than ten or more than 100 proteins in order to mitigate the influence of very broad or specific terms. Protein groups with discordant mappings (for example, with one protein annotated with the GO term of interest but the other protein not annotated with the GO term of interest) were discarded. AUCs for each individual GO slim term were then summarized to the median AUC over all GO slim terms. iBAQ protein quantitation was used for all ROC analyses. To produce the heatmap shown in Extended Data Fig. 2b, the mean AUC over all missing-value strategies and chromatogram transformations was calculated for each measure of association in each dataset. The mean AUCs were then ranked within each dataset to produce a dataset-specific ranking of each measure of association.

We used several different approaches to perform statistical comparisons of the 24 measures of association. First, we performed univariate statistical analyses to test for differences in overall performance between each pair of metrics. We used the nonparametric Brunner–Munzel test[69], as implemented in the 'lawstat' R package, to test for the stochastic equality of the AUCs obtained using each measure of association. We carried out similar univariate analyses using the Brunner–Munzel test to compare approaches to missing-value handling and chromatogram normalization. Second, we observed that some measures of association achieved good performance only after specific preprocessing strategies were applied. We therefore performed a second univariate analysis, retaining only AUCs derived from the optimal preprocessing pipeline for each measure of association (that is, the combination of missing-value handling and chromatogram normalization that yielded the highest median AUC). Third, to identify statistically significant interactions between preprocessing strategies and measures of association, we performed a multivariable statistical analysis, fitting a linear model to the protein AUC, including terms for the measure of association, missing-value handling, chromatogram normalization and interactions between them. The four measures of co-occurrence studied here (that is, the Jaccard index, Hamming distance, Dice coefficient and co-dependency index) are invariant to preprocessing decisions and were omitted from this analysis. Finally, to compare approaches to label-free protein quantification, we used the paired Brunner–Munzel test[70], as implemented in the 'nparcomp' R package, to compare the AUCs derived from each dataset after quantifying protein abundance in each fraction with one of three different algorithms.

To compare peak-centric and whole-chromatogram similarities, we used the R implementation of PrInCE[71] to deconvolve chromatograms into a mixture of Gaussians and then computed the co-apex score (defined as the Euclidean distance between the closest (μ, σ) pairs, where μ and σ are parameters of Gaussians fitted to any two chromatograms[24]) as a representative peak-centric metric. The performance of approaches to chromatogram deconvolution is highly sensitive to small anomalies in the order of the fractions (for example, missing fractions, sample mix-ups), and the procedure implemented in PrInCE is incompatible with sequential fractionation approaches. To put peak-centric approaches on the best possible footing and avoid spuriously underestimating their performance, we therefore limited our analysis to a subset of 20 CF-MS datasets generated in our own laboratory, for which we can be entirely confident about the application of the deconvolution procedure. We then repeated the ROC analysis as described above using the co-apex score to rank co-eluting protein pairs. Because not all chromatograms are amenable to deconvolution, we recalculated the AUC of all 24 other measures of association for only the proteins that could be fitted with a mixture of Gaussians ($r^2 \geq 0.5$) to compare all methods on the same set of proteins.

To evaluate the impact of other preprocessing decisions on CF-MS data analysis, we performed identical ROC analyses after filtering proteins quantified in less than 1–20 fractions and after filtering protein pairs co-occurring in less than one to ten fractions. These analyses were carried out with missing values treated as zeros and without transforming or normalizing chromatograms, using either the Pearson correlation or the mutual information as the measure of association.

Comparison of experimental designs. To evaluate the impact of the number of fractions collected from a given replicate, we downsampled published datasets to between five and 50 fractions and then repeated the ROC analysis described above. Only experiments that originally profiled at least 50 fractions were included. This analysis was performed with either the Pearson correlation or the mutual information as the measure of association, missing values treated as zeros and no transformation of the chromatograms. We additionally performed downsampling of adjacent fractions by passing a sliding window of fixed width along the chromatograms and retaining the maximum correlation for each protein pair over all windows, observing similar results.

To quantify the impact of biological replication, an analogous analysis was performed for a set of nine CF-MS experiments with at least three biological replicates, in which fractions were sampled at random from between one and five replicates. Protein groups were mapped to gene symbols to enable matching across replicates. For gene symbols that mapped to more than one protein group, only the chromatogram with the fewest missing values was retained. Duplicated chromatograms (arising from the reverse case, in which a single protein group mapped to more than one gene symbol) were discarded at random. For each parameter combination (that is, for a given number of fractions and number of biological replicates), separate ROC analyses were performed for ten random samples of fractions, and the mean AUC over all ten samples was calculated. Because individual experiments displayed very different intrinsic powers to resolve known protein complexes, AUCs were rescaled to the range [0, 1] for each experiment separately to enable comparison.

To compare label-free and isotopic labeling approaches to protein quantification, iBAQ intensities were extracted from medium and/or heavy channels for a total of 20 SILAC experiments. These were then compared to the medium (heavy) over light ratios, using SILAC internal standards, for their ability to resolve known protein complexes or proteins annotated with the same GO term as described above. A smaller number of dimethyl labeling datasets ($n = 3$) were discarded from this analysis on the grounds that confident conclusions could not be drawn about this technique from such a small number of replicates, each with relatively few proteins quantified. Statistical comparisons of iBAQ and SILAC versions of the same dataset were performed using a paired t-test.

For the comparison of protein complex fractionation approaches, experiments that combined N-PAGE or SEC with cross-linking, as well as two experiments employing sucrose gradients, were excluded. In addition to computing the AUC over all intracomplex and intercomplex interactions in CORUM, we also performed separate ROC analyses for each individual complex in turn, analogously to ROC analyses performed for proteins annotated with the same GO term described above. We limited our comparison to the set of complexes that were detected in at least one experiment using each separation method to avoid biasing the analysis toward proteins incompatible with specific approaches. We compared the resulting AUCs using the Brunner–Munzel test as described above. Additionally, we extended our multivariable statistical analysis by adding a term for the fractionation method to the linear models described above. Separately, we sought to identify complexes for which at least three subunits were detected exclusively by one method or which were resolved significantly better by one method. To address the latter question, we performed a one-tailed t-test comparing the AUCs achieved by each fractionation method to those of all other methods (a 'one-versus-rest' comparison) and then corrected for multiple-hypothesis testing using the false discovery rate.

Integration of multiple CF-MS replicates. To combine information from multiple CF-MS replicates in network inference, we employed a supervised machine learning workflow as previously implemented in a number of previous studies and software tools[9–11,14,17,32,33,72–76]. Briefly, within each replicate, a series of features indicative of protein complex co-membership were calculated for each protein pair. The calculated features were then merged to produce a single feature matrix, which was provided to a machine learning classifier as input alongside a set of known protein complexes. As in the ROC analysis, proteins belonging to the same complex were labeled as true positives, whereas proteins in different complexes were labeled as true negatives. The classifier was trained using fivefold cross-validation both to minimize overfitting and to allow it to make predictions for protein pairs within the training set of known protein complexes. Protein pairs were ranked in descending order by their mean classifier score across all five folds, and the AUC was calculated as described above.

We considered several variations on this general approach. First, we varied both the number and identities of the features provided to the classifier as input. Calculating between one and ten features per replicate, we drew features either at random from the set of 165 analytical pipelines or in descending order from a list of pipelines arranged by their combined AUCs in the protein complex and GO term analyses ('best first'). For best-first features, only the single best combination of missing-value handling and chromatogram transformation was retained for any given measure of association to avoid repeatedly drawing slightly different versions of the same feature. Alternatively, a set of five features computed in the 'PrInCE' R package was used as input (a single PrInCE feature, the co-apex score, derived from fitting a mixture of Gaussians to each chromatogram, was omitted). We also attempted to perform a comparison to the EPIC toolkit[14] but were ultimately unable to perform a fair comparison because the network returned to users by this tool is obtained by both training and testing the classifier on the input set of 'gold standard' complexes (that is, the network is inferred without cross-validation). We also varied the total number of CF-MS datasets sampled from a minimum of two to a maximum of 40. For each combination of parameters, network inference was repeated ten times, each with different samples of CF-MS replicates. Proteins quantified in less than four fractions were discarded from each replicate. Protein groups were mapped to gene symbols as described above.

We also considered the structure of the cross-validation procedure itself. A set of known protein complexes can be separated on the basis of complex subunits (proteins) or pairwise interactions between complex subunits (protein pairs) as illustrated schematically in Extended Data Fig. 8a. To compare these two cross-validation approaches, the set of CORUM protein complexes was converted to an adjacency matrix, and the proteins in the row and column names of the matrix were split into five folds. Alternatively, the adjacency matrix was converted to a pairwise data frame, which was then split into five folds. We then computed the difference in the apparent AUC obtained from cross-validation with complexes split by protein pairs or by proteins. The difference between the two AUCs was also recalculated in an independent held-out set of complexes comprising 30% of the CORUM database, which was withheld at the beginning of the experiment.

Other adaptations to the machine learning workflow that were considered included the impact of merging matrices from multiple CF-MS replicates before feature calculation, leading to a single feature matrix for all replicates. Additionally, because the naive Bayes classifier can naturally handle missing values by simply excluding them from the likelihood calculation, we compared the impact of imputing missing features with their median value versus leaving the missing values in place. Because neither of these two adaptations improved performance, they were not considered further. Finally, we compared two different classifiers: a naive Bayes classifier, which exemplifies a broader category of relatively simple, linear classifiers, previously found to perform well with CF-MS data[17,24,75,77,78]; and a random forest classifier, which exemplifies a family of more complex, nonlinear classifiers based on decision trees that have also found wide use in CF-MS analysis[9,11,32,33,73,74]. The implementations of these classifiers from the R packages 'naivebayes' and 'randomForest' were used, respectively.

To identify pairs of features with synergistic or antagonistic interactions, we performed network inference using all pairwise combinations of the 24 measures of association. For each measure of association, we considered only the combination of missing-value handling strategy and chromatogram transformation that yielded the optimal results in the single-dataset setting (that is, we took the row maximum from Extended Data Fig. 4 after summing the AUCs obtained for resolving protein complexes and pairs of proteins annotated with the same GO term). We used either a random forest or a naive Bayes classifier to integrate combinations of three or six CF-MS datasets. For each pair of features, we performed network inference from ten random samples of CF-MS datasets of the given size, yielding a total of 11,040 networks. We then fit a linear mixed model with a random effect for the specific combination of CF-MS datasets to the AUC of each network, as we found that the identities of the datasets chosen for integration had a marked effect on the AUC, and tested the statistical significance of the interaction between the two features using the 'lmerTest' R package to optimize the restricted maximum likelihood and obtain P values from the Satterthwaite approximation for degrees of freedom. To identify reproducible interactions, we tallied the number of times a synergistic or antagonistic interaction was detected in the four evaluation scenarios (two classifiers, three or six datasets); we did not identify any discordant interactions. To compare networks inferred from combinations of human SEC and IEX datasets, we trained a random forest classifier on the six top-performing features from each dataset, drawing a total of 20 random samples for each total number of datasets and proportion of SEC datasets.

Downsampling analysis of training set complexes. To estimate the minimum number of protein complexes required for robust network inference in species in which few protein complexes may be known, we repeated the analyses of multiple CF-MS replicates after downsampling the CORUM database to include between 5% and 100% of complex proteins. For this analysis, we considered only combinations of two to four replicates, computing the six 'best-first' features in each replicate. As described above, network inference was repeated ten times for each combination of parameters, drawing a new sample of CF-MS replicates each time, and these samples were held constant throughout the downsampling of CORUM. Because the range of AUC values that could be obtained was found to vary strongly with the specific replicates selected, AUCs were rescaled to the range [0, 1] for each combination of replicates to enable comparison of saturation curves across the ten samples.

Inference and validation of a consensus human interactome by CF-MS. To infer a consensus human interactome by CF-MS, we combined information from all

46 CF-MS datasets using tenfold cross-validation and testing a slightly expanded set of classifiers, including logistic regression[72] and support-vector machines[10,14]. A single 'best-first' feature was calculated from each replicate. Networks were also inferred from samples of between two and 45 CF-MS replicates, as well as from each individual replicate in turn.

To evaluate the potential of large-scale CF-MS analysis for interactome mapping, the consensus CF-MS interactome derived from integration of all 46 human datasets was compared to five recently published systematic screens of the human interactome: three achieved using AP–MS[3–5] and two achieved using Y2H[1,2]. Networks were first compared on the basis of their precision, as calculated based on the CORUM database, and the total number of interactions identified. Next, we computed the functional coherence of each network, defined as the degree to which the function of any given protein can be predicted from those of its interacting partners, based on the principle of 'guilt by association' (refs. [36,79]). Briefly, each protein in the network is annotated with its known functions (here, GO terms), and a subset of these labels are then withheld. A simple neighbor-voting algorithm[80] is then used to predict functions for the withheld proteins by assigning a score for each GO term that represents the proportion of the protein's interacting partners annotated with the same term. This process is repeated in threefold cross-validation, and the mean AUC over cross-validation folds is computed for each GO term. A high AUC is characteristic of networks in which proteins that share biological functions tend to be physically connected. Functional coherence analysis was carried out using the 'EGAD' R package[36], filtering GO terms annotated to less than ten or more than 100 proteins as described above. We additionally compared networks based on the tendency for interacting proteins to display correlated patterns of abundance in two large-scale proteomic datasets[20,37] and to colocalize to the same subcellular compartments in two subcellular proteomic datasets[38,39]. Both protein coexpression and colocalization were quantified using the Pearson correlation.

Inference of CF-MS interactomes for 27 species or clades. Finally, we carried out a similar analysis to integrate CF-MS experiments from other non-human species or across multiple species simultaneously for broad phylogenetic groups such as mammals or tetrapods. For these analyses, protein groups were mapped to eggNOG orthogroups in the 'euk' database using the eggnog-mapper tool (version 1.0.3)[81]. Proteins that mapped to more than one orthogroup were discarded, and profiles were constructed for each orthogroup by summing protein groups mapping to that orthogroup. As in the consensus human interactome, random forest classifiers were trained with a single 'best-first' feature per replicate. Functional coherence was evaluated by mapping UniProt accessions in GOA files to eggNOG orthogroups. The phylogenetic tree was obtained from TimeTree[82].

Visualization. Throughout the text, box plots show the median (horizontal line), interquartile range (hinges) and smallest and largest values no more than 1.5 times the interquartile range (whiskers).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

A list of all raw mass spectrometry files analyzed in this study and their accession numbers in PRIDE or MassIVE repositories is provided in Supplementary Table 1. All data generated in this study are available at multiple levels of analysis from the following sources: protein chromatograms and protein–protein interaction networks for up to ten proteins can be visualized and downloaded via an interactive web application at http://cf-ms-browser.msl.ubc.ca; processed chromatograms and MaxQuant proteinGroups.txt files are available via Zenodo at https://doi.org/10.5281/zenodo.4499320; complete MaxQuant outputs for all 206 experiments were deposited to the PRIDE repository[83] with the dataset identifier PXD022048; predicted interactomes for 27 species and clades, including the consensus human CF-MS interactome, are available via Zenodo at https://doi.org/10.5281/zenodo.4245282. An overview of all publicly available resources generated in this study is provided at the supporting website (https://fosterlab.github.io/CF-MS-analysis).

Code availability

Source code used to download and reanalyze publicly available CF-MS data using MaxQuant is available at https://github.com/skinnider/CF-MS-searches (https://doi.org/10.5281/zenodo.4774750). Source code used to carry out analyses presented in the paper, with relevant intermediate data files, is available at https://github.com/skinnider/CF-MS-analysis (https://doi.org/10.5281/zenodo.4774754). Source code for the CF-MS browser web application is available at https://github.com/skinnider/CF-MS-browser (https://doi.org/10.5281/zenodo.4774752). The CFTK R package is available at https://github.com/fosterlab/CFTK (https://doi.org/10.5281/zenodo.4774771).

References

58. Kovalchik, K. A. et al. RawTools: rapid and dynamic interrogation of Orbitrap data files for mass spectrometer system management. *J. Proteome Res.* **18**, 700–708 (2019).
59. Bogdanow, B., Zauber, H. & Selbach, M. Systematic errors in peptide and protein identification and quantification by modified peptides. *Mol. Cell. Proteomics* **15**, 2791–2801 (2016).
60. Alexa, A., Rahnenführer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600–1607 (2006).
61. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257–258 (2007).
62. Sánchez-Taltavull, D., Ramachandran, P., Lau, N. & Perkins, T. J. Bayesian correlation analysis for sequence count data. *PLoS ONE* **11**, e0163595 (2016).
63. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* **5**, e12776 (2010).
64. Székely, G. J., Rizzo, M. L. & Bakirov, N. K. Measuring and testing dependence by correlation of distances. *Ann. Stat.* **35**, 2769–2794 (2007).
65. Simon, N. & Tibshirani, R. Comment on "Detecting novel associations in large data sets" by Reshef et al., Science Dec. 16, 2011. Preprint at https://arxiv.org/abs/1401.7645 (2014).
66. Kinney, J. B. & Atwal, G. S. Equitability, mutual information, and the maximal information coefficient. *Proc. Natl Acad. Sci. USA* **111**, 3354–3359 (2014).
67. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).
68. Foroushani, A. et al. Large-scale gene network analysis reveals the significance of extracellular matrix pathway and homeobox genes in acute myeloid leukemia: an introduction to the Pigengene package and its applications. *BMC Med. Genomics* **10**, 16 (2017).
69. Brunner, E. & Munzel, U. The nonparametric Behrens–Fisher problem: asymptotic theory and a small-sample approximation. *Biomed. J.* **42**, 17–25 (2000).
70. Munzel, U. & Brunner, E. An exact paired rank test. *Biomed. J.* **44**, 584–593 (2002).
71. Skinnider, M. A., Cai, C., Stacey, R. G. & Foster, L. J. PrInCE: an R/Bioconductor package for protein–protein interaction network inference from co-fractionation mass spectrometry data. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btab022 (2021).
72. Larance, M. et al. Global membrane protein interactome analysis using in vivo crosslinking and mass spectrometry-based protein correlation profiling. *Mol. Cell. Proteomics* **15**, 2476–2490 (2016).
73. Crozier, T. W. M., Tinti, M., Larance, M., Lamond, A. I. & Ferguson, M. A. J. Prediction of protein complexes in *Trypanosoma brucei* by protein correlation profiling mass spectrometry and machine learning. *Mol. Cell. Proteomics* **16**, 2254–2267 (2017).
74. Hillier, C. et al. Landscape of the *Plasmodium* interactome reveals both conserved and species-specific functionality. *Cell Rep.* **28**, 1635–1647 (2019).
75. Kerr, C. H. et al. Dynamic rewiring of the human interactome by interferon signaling. *Genome Biol.* **21**, 140 (2020).
76. Liebeskind, B. J., Aldrich, R. W. & Marcotte, E. M. Ancestral reconstruction of protein interaction networks. *PLoS Comput. Biol.* **15**, e1007396 (2019).
77. Skinnider, M. A., Stacey, R. G. & Foster, L. J. Genomic data integration systematically biases interactome mapping. *PLoS Comput. Biol.* **14**, e1006474 (2018).
78. Carlson, M. L. et al. Profiling the *Escherichia coli* membrane protein interactome captured in Peptidisc libraries. *eLife* **8**, e46615 (2019).
79. Oliver, S. Guilt-by-association goes global. *Nature* **403**, 601–603 (2000).
80. Schwikowski, B., Uetz, P. & Fields, S. A network of protein–protein interactions in yeast. *Nat. Biotechnol.* **18**, 1257–1261 (2000).
81. Huerta-Cepas, J. et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
82. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
83. 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).

Acknowledgements

This work was supported by funding from Genome Canada and Genome British Columbia (project 214PRO). M.A.S. acknowledges support from a CIHR Vanier Canada Graduate Scholarship, an Izaak Walton Killam Memorial Pre-Doctoral Fellowship, a University of British Columbia (UBC) Four Year Fellowship and a Vancouver Coastal Health–CIHR–UBC MD/PhD Studentship. This work was enabled in part by the support provided by WestGrid and Compute Canada and through computational resources and services provided by Advanced Research Computing at the

UBC. We thank Microsoft for the donation of cloud computing resources that enabled part of this work, T. Clark and J. Moon for advice on MaxQuant searches and D. Vavilov for assistance with the web server.

Author contributions

M.A.S. and L.J.F. designed experiments. M.A.S. performed experiments. M.A.S. and L.J.F. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

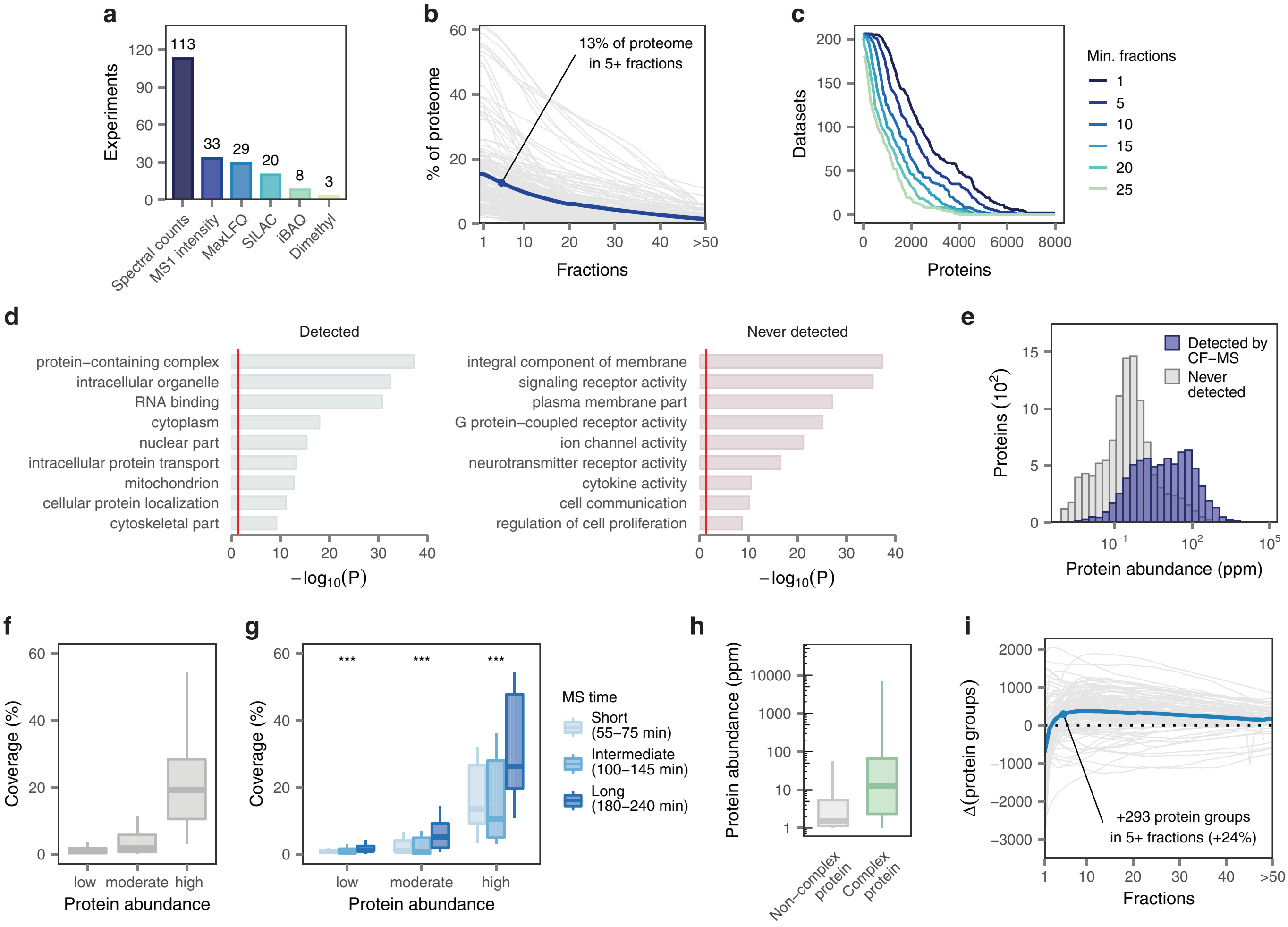
Extended data is available for this paper at https://doi.org/10.1038/s41592-021-01194-4.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-021-01194-4.

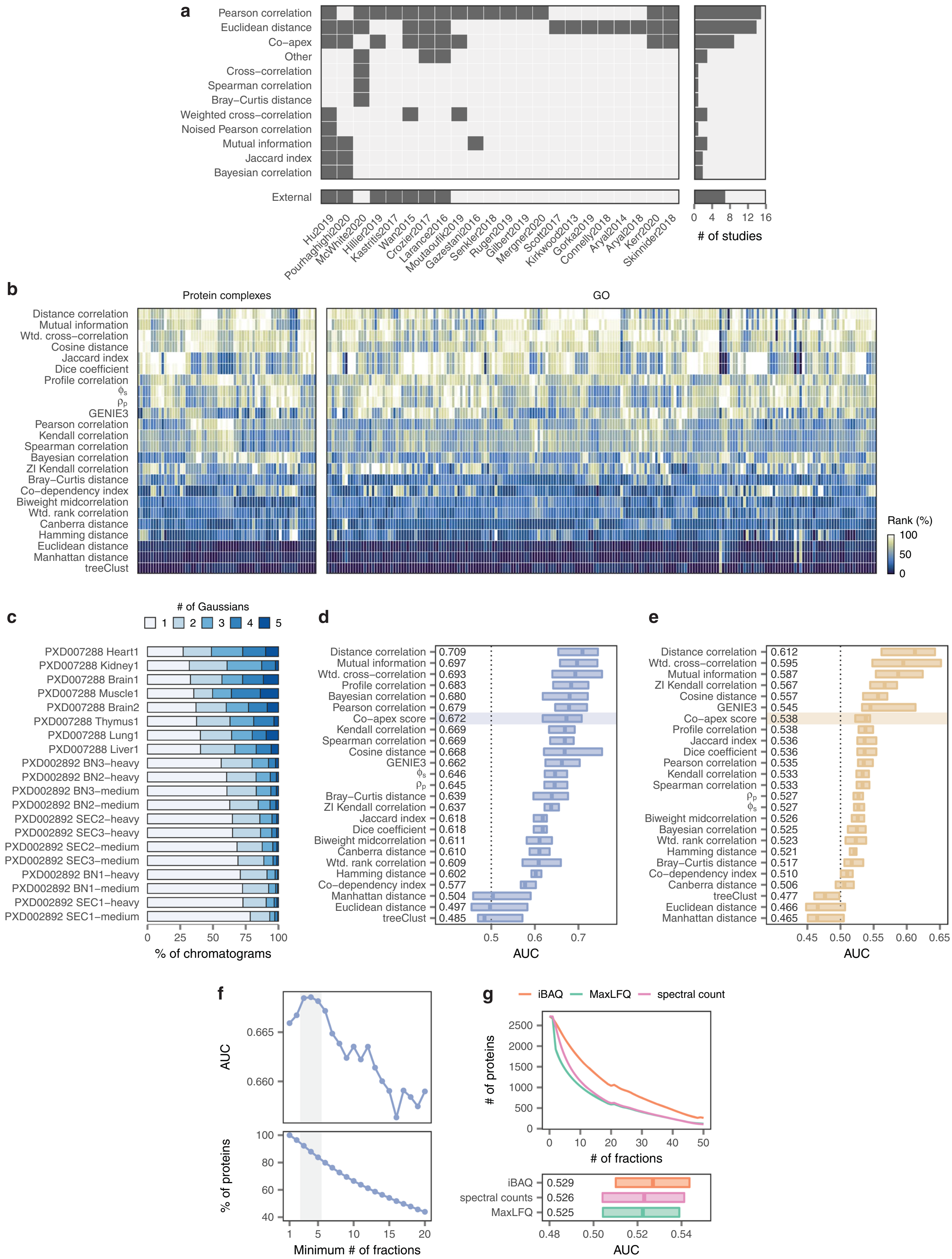
Correspondence and requests for materials should be addressed to L.J.F.

Peer review information *Nature Methods* thanks Fridtjof Lund-Johansen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Arunima Singh was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

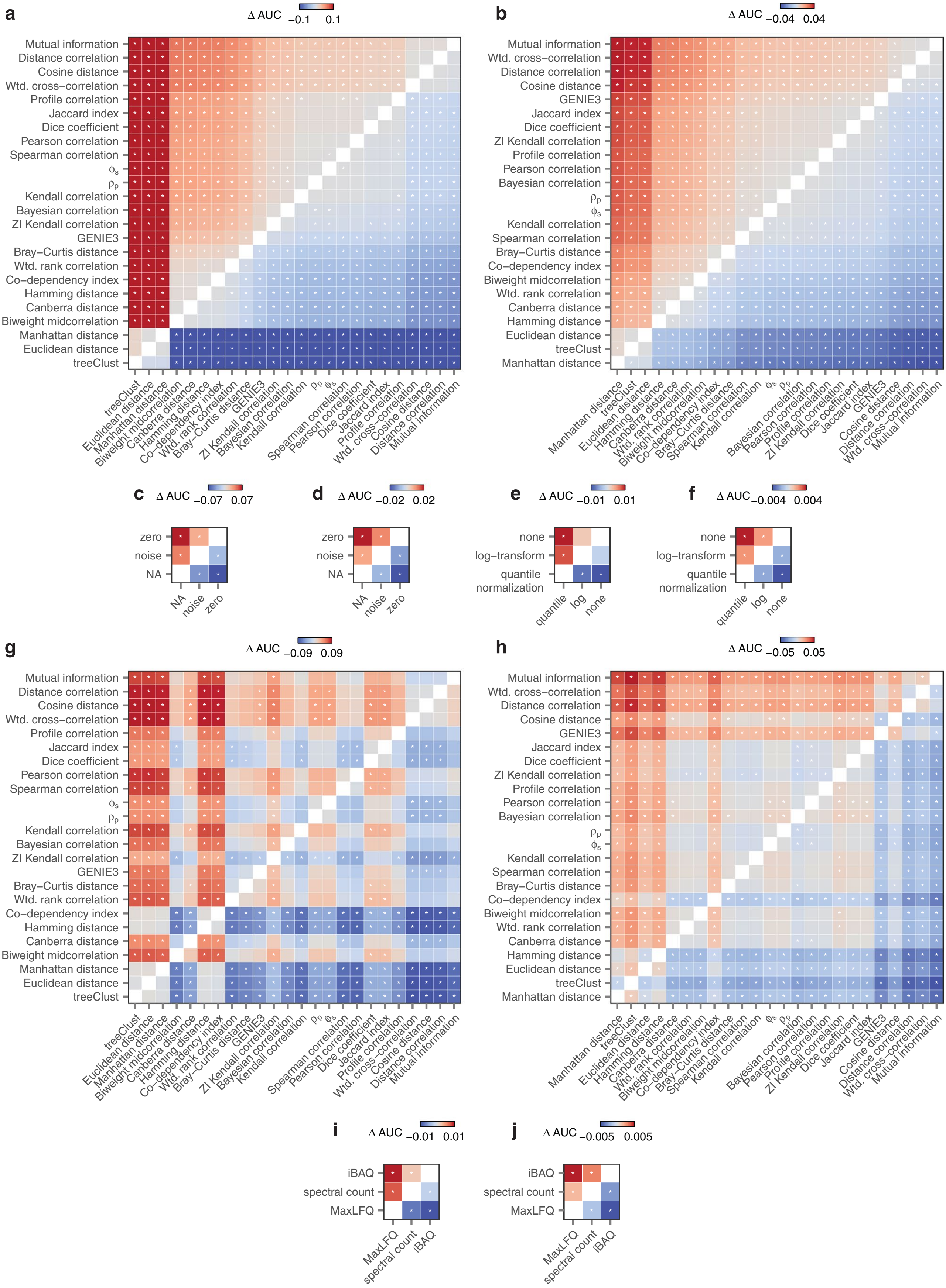


Extended Data Fig. 1 | A uniformly processed resource of CF-MS data. a, Approaches to protein quantification employed by published CF-MS experiments. SILAC, stable isotope labelling by amino acids in cell culture; iBAQ, intensity-based absolute quantification. **b**, Proportion of the organismal proteome quantified in each CF-MS experiment (grey lines, individual datasets; blue line, mean across all datasets). **c**, Cumulative distribution of the number of proteins quantified per dataset in between one and 25 fractions. **d**, GO term enrichment among CORUM proteins detected in at least one CF-MS fraction, left, or never detected, right. **e**, PaxDb consensus protein abundance of mouse proteins detected or never detected by CF-MS. **f**, Coverage of high, moderate, and low abundance proteins (expressed as a mean proportion of fractions in which these proteins were detected) in published human CF-MS experiments ($n = 46$). **g**, As in **f**, but with CF-MS experiments divided into three groups based on the length of the liquid chromatography gradient. ***, $p < 0.001$, two-sided Spearman rank correlation. **h**, PaxDb consensus protein abundance of human proteins in the CORUM database and non-CORUM proteins. **i**, Difference in the number of protein groups quantified in each CF-MS experiment, compared to the processed chromatogram data accompanying the original publications (grey lines, individual datasets; blue line, mean across all datasets).



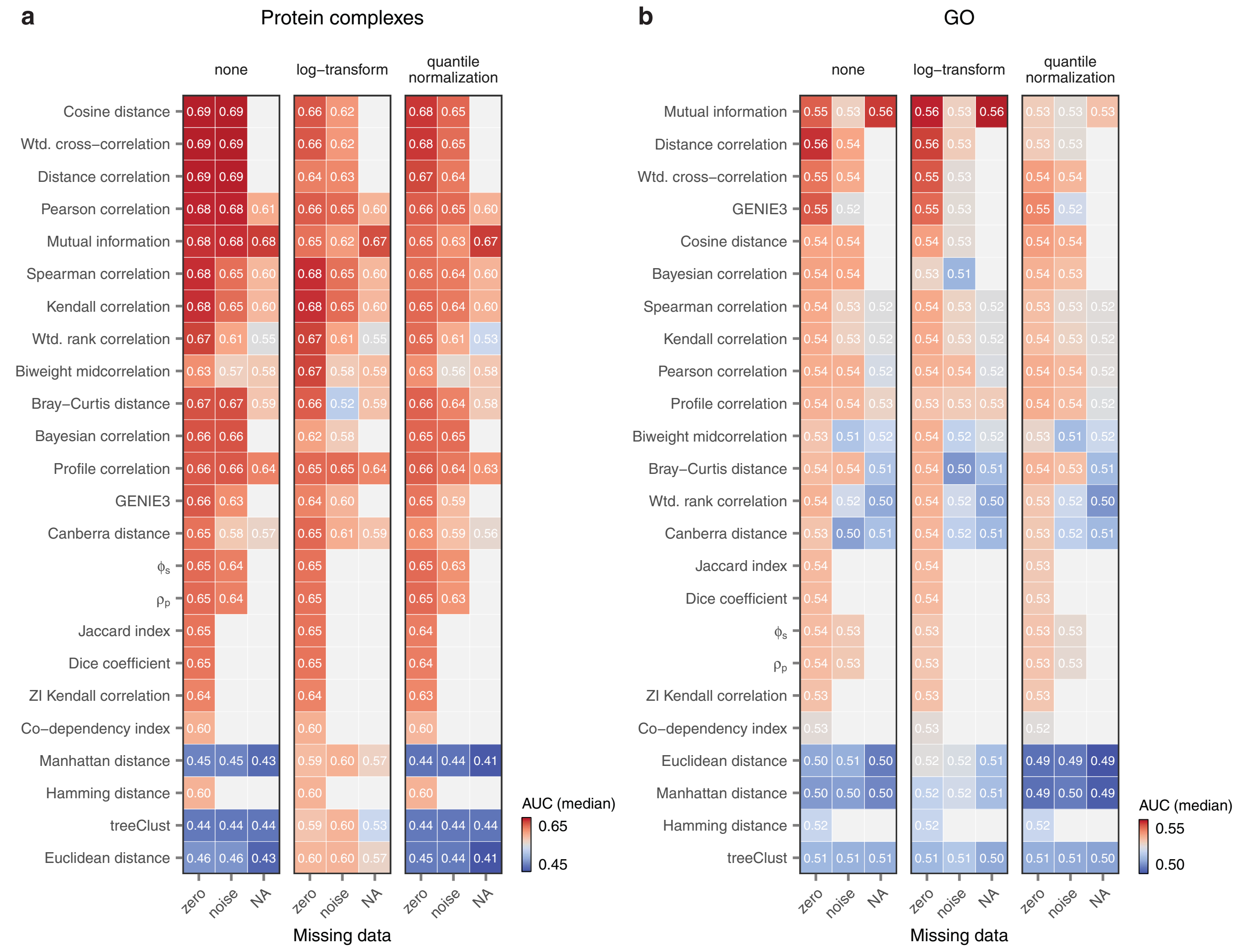
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Benchmarking computational analysis of individual CF-MS datasets. a, Measures of association used to quantify the similarity of two protein chromatograms in published CF-MS studies. Bottom row indicates the incorporation of external genomic datasets[77]. **b**, Ranks of each measure of association in identifying protein pairs in the same protein complex, left, or annotated to the same GO term, right, across individual CF-MS datasets. **c**, Number of peaks detected in 20 CF-MS datasets by fitting a mixture of Gaussians to each protein chromatogram. **d**, Recovery of known protein complexes in the 20 CF-MS datasets from **c**, scoring only chromatograms that could be fit with a mixture of Gaussians ($r^2 \geq 0.5$) and comparing the 24 different measures of association shown in Fig. 2 with the co-apex score. Inset text shows the median AUC for each measure of association. **e**, As in **d**, but for proteins annotated to the same GO term. **f**, Recovery of known protein complexes, top, and proportion of originally quantified proteins, bottom, when filtering profiles not detected in some minimum number of fractions, using mutual information as a measure of profile similarity. **g**, Mean number of protein groups identified, top, and recovery of proteins annotated to the same GO term, bottom, for three approaches to label-free quantification implemented in MaxQuant.

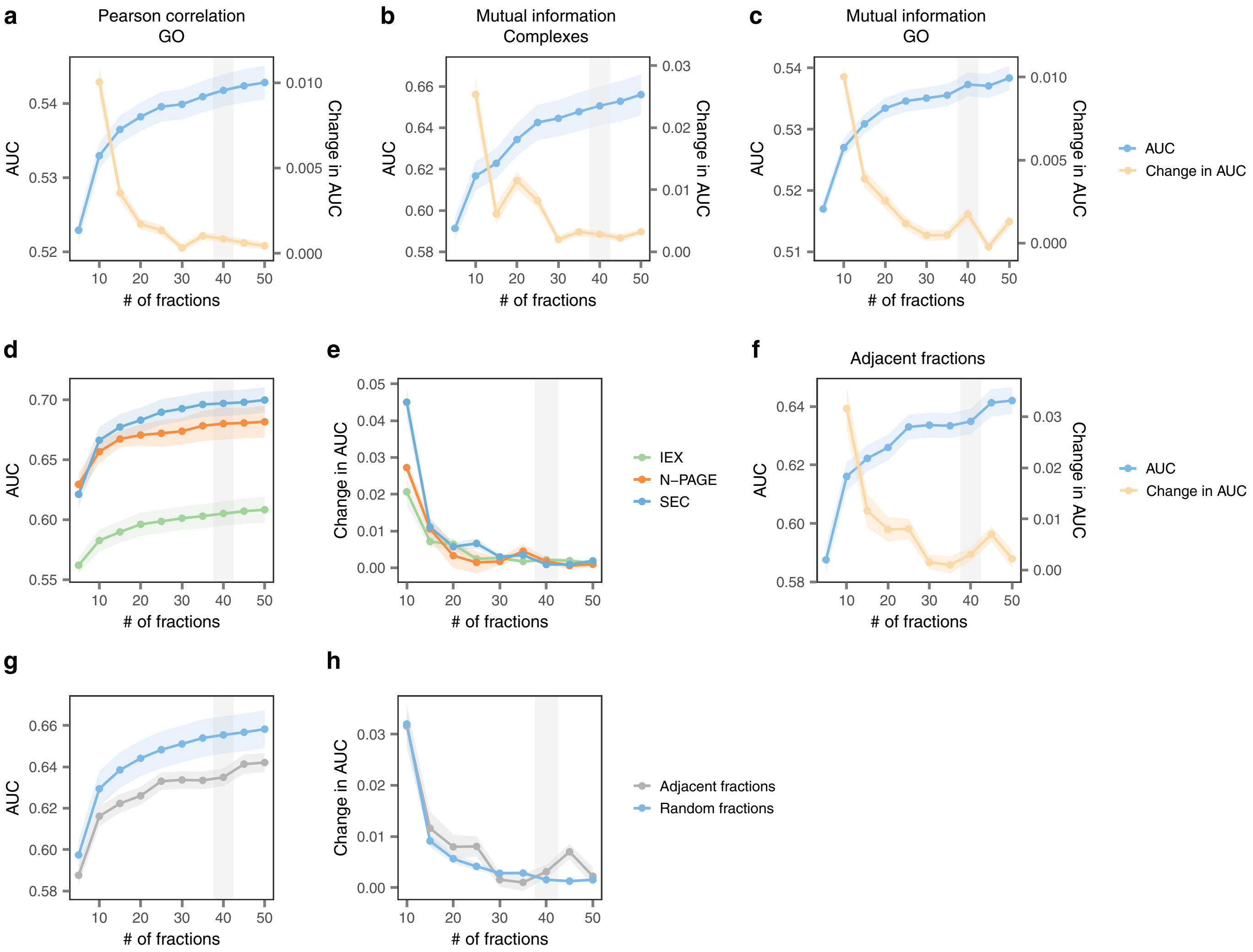


Extended Data Fig. 3 | See next page for caption.

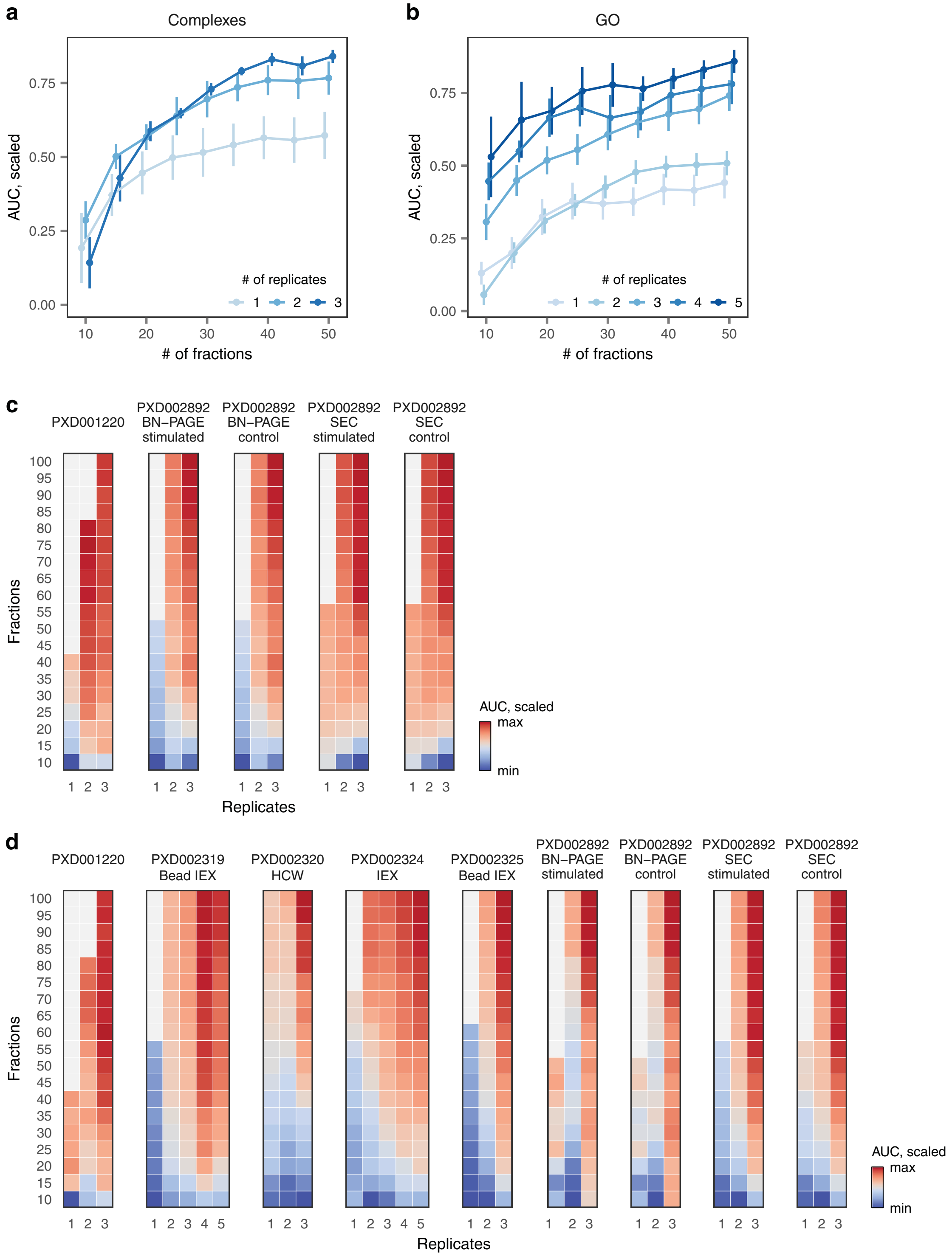
Extended Data Fig. 3 | Univariate statistical analysis of computational approaches to individual CF-MS datasets. a, Difference in the median protein complex AUC between each pair of measures of association. Asterisks indicate pairs of measures of association with a p-value less than 0.05 in a two-sided Brunner-Munzel test. The difference in median AUCs is capped at [−0.1, +0.1] to improve visualization. **b**, As in **a**, but for GO terms. **c**, Difference in the median protein complex AUC between each pair of missing value handling strategies. Asterisks indicate pairs of measures of association with a p-value less than 0.05 in a two-sided Brunner-Munzel test. **d**, As in **c**, but for GO terms. **e**, Difference in the median protein complex AUC between each pair of chromatogram normalization approaches. Asterisks indicate pairs of measures of association with a p-value less than 0.05 in a two-sided Brunner-Munzel test. **f**, As in **e**, but for GO terms. **g**, Difference in the median protein complex AUC between each pair of measures of association, considering only the single best combination of missing value handling and chromatogram normalization for each measure of association. Asterisks indicate pairs of measures of association with a p-value less than 0.05 in a two-sided Brunner-Munzel test. **h**, As in **g**, but for GO terms. **i**, Median difference in the protein complex AUC between matched datasets with label-free protein quantification performed by one of three algorithms within MaxQuant. Asterisks indicate pairs of measures of association with a p-value less than 0.05 in a two-sided paired Brunner-Munzel test. **j**, As in **i**, but for GO terms.



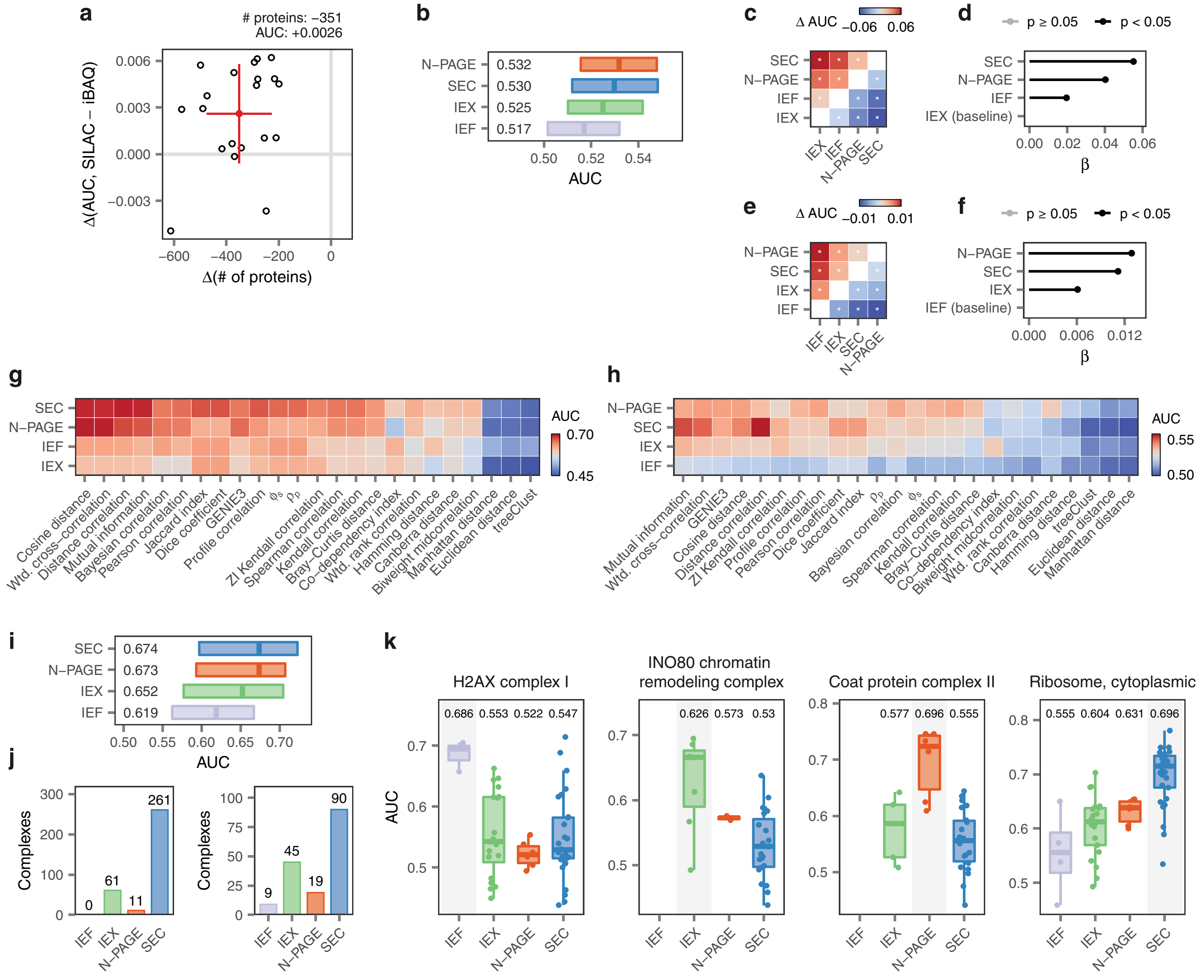
Extended Data Fig. 4 | Analysis pipelines for individual CF-MS datasets. a, Recovery of known protein complexes with 163 valid combinations of measures of association, missing value handling, and normalization strategies. **b**, As in **a**, but for proteins annotated to the same GO term.



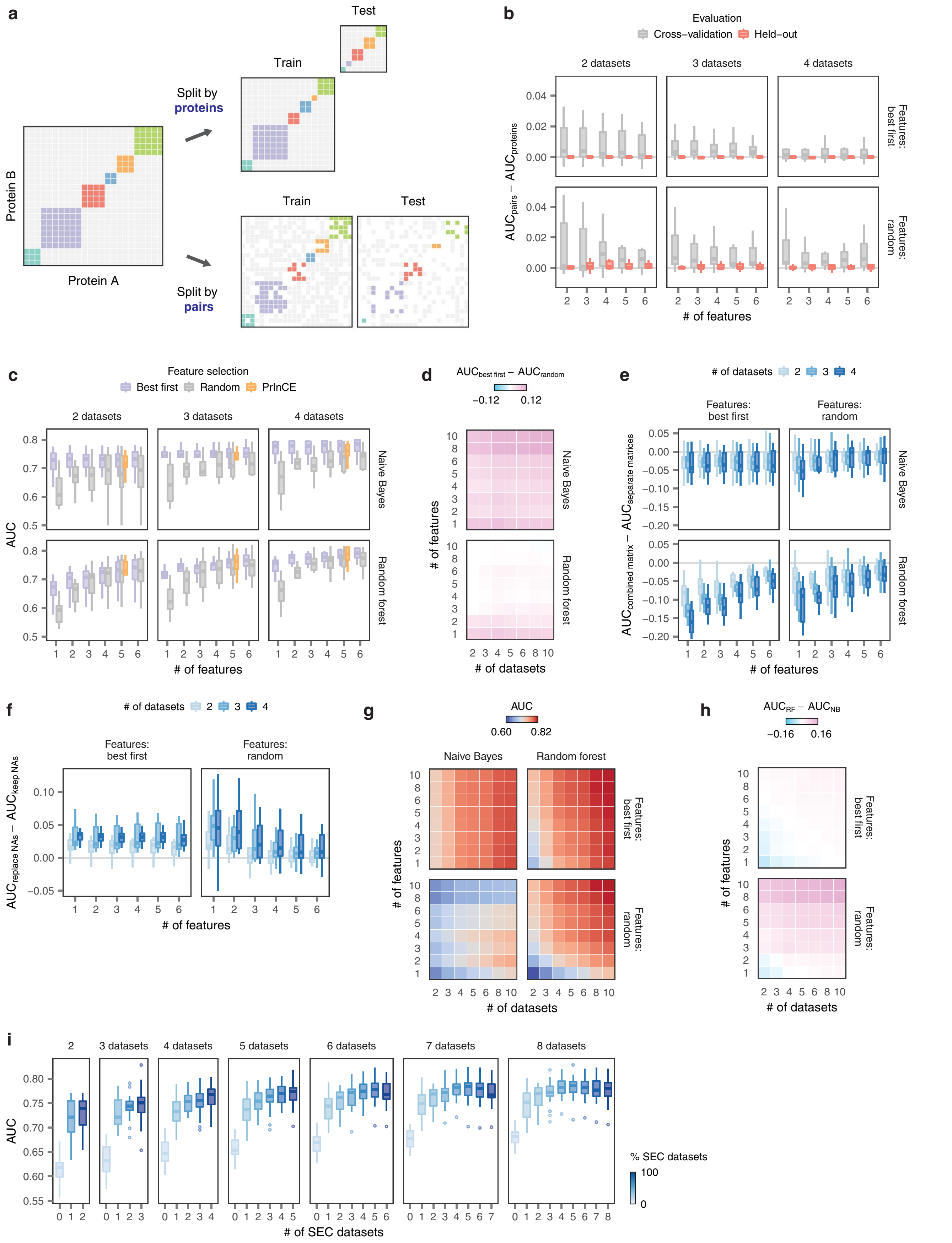
Extended Data Fig. 5 | Downsampling analysis of published CF-MS experiments. a, Recovery of proteins annotated to the same GO term after downsampling CF-MS chromatograms to a fixed number of fractions. **b–c**, Recovery of protein complexes, **b**, and GO terms, **c**, in downsampled CF-MS chromatograms, using mutual information as the measure of profile similarity. **d–e**, Recovery of protein complexes (AUC, **d**, and change in AUC, **e**), in downsampled CF-MS chromatograms, with experiments divided based on the separation method used (IEX, ion exchange chromatography; N-PAGE, native polyacrylamide gel electrophoresis; SEC, size exclusion chromatography). **f**, Recovery of protein complexes when downsampling windows of adjacent fractions of fixed length, rather than downsampling fractions randomly from the chromatogram matrix. **g–h**, Comparison of protein complex recovery (AUC, **g**, and change in AUC, **h**), in downsampled CF-MS chromatograms when drawing samples of random fractions or adjacent fractions from the chromatogram matrix. **a–h**, Shaded area shows the standard error.



Extended Data Fig. 6 | Downsampling analysis of published CF-MS experiments incorporating multiple biological replicates. a–b, Recovery of known protein complexes, **a**, and proteins annotated to the same GO term, **b**, in downsampled CF-MS chromatograms with fractions sampled from one to five replicates, as shown in Fig. 3b but visualized here as a line graph instead. Error bars show the standard error of the mean. **c–d**, Recovery of known protein complexes, **c**, and proteins annotated to the same GO term, **d**, in downsampled CF-MS chromatograms with fractions sampled from one to five replicates, within individual CF-MS datasets.

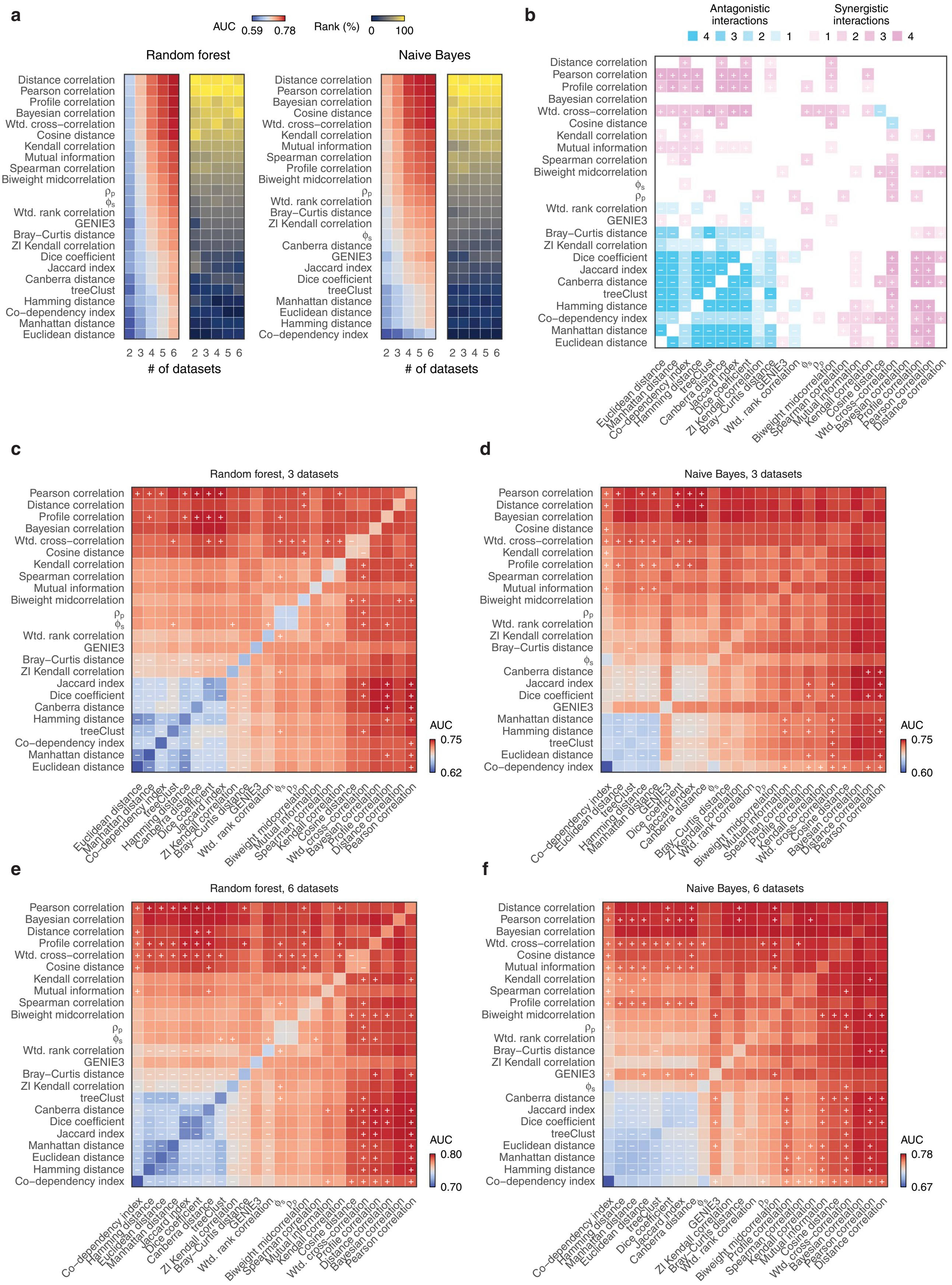


Extended Data Fig. 7 | Protein quantification and chromatographic separation. a, Comparison of GO term recovery and proteome coverage between SILAC ratios and iBAQ intensities from individual isotopologue channels in 20 SILAC datasets. Caption and error bars show the mean and standard deviation of the differences in the number of protein groups quantified and the AUC between SILAC ratios and iBAQ intensities. **b**, Recovery of proteins annotated to the same GO term in CF-MS experiments grouped by fractionation method. **c**, Difference in the median protein complex AUC between each pair of fractionation methods. Asterisks indicate pairs of measures of association with a p-value less than 0.05 in a two-sided Brunner-Munzel test. **d**, Regression coefficients for fractionation methods in multivariable statistical analysis, estimated by a linear model fit to the protein complex AUC and including terms for measures of association, missing value handling strategies, approaches to chromatogram normalization, and interactions between them. **e**, As in **c**, but for GO terms. **f**, As in **d**, but for GO terms. **g**, Recovery of known protein complexes in published CF-MS experiments grouped by fraction method, with each measure of association shown separately. **h**, As in **g**, but for proteins annotated to the same GO term. **i**, Recovery of individual protein complexes in published CF-MS experiments grouped by fractionation method. **j**, Number of protein complexes with at least three subunits detected exclusively by one separation method, left, and resolved significantly better by one of the four separation methods, right, across 67 human and mouse CF-MS datasets. **k**, Examples of protein complexes resolved best by each of the four separation methods. Inset text shows the median AUC.



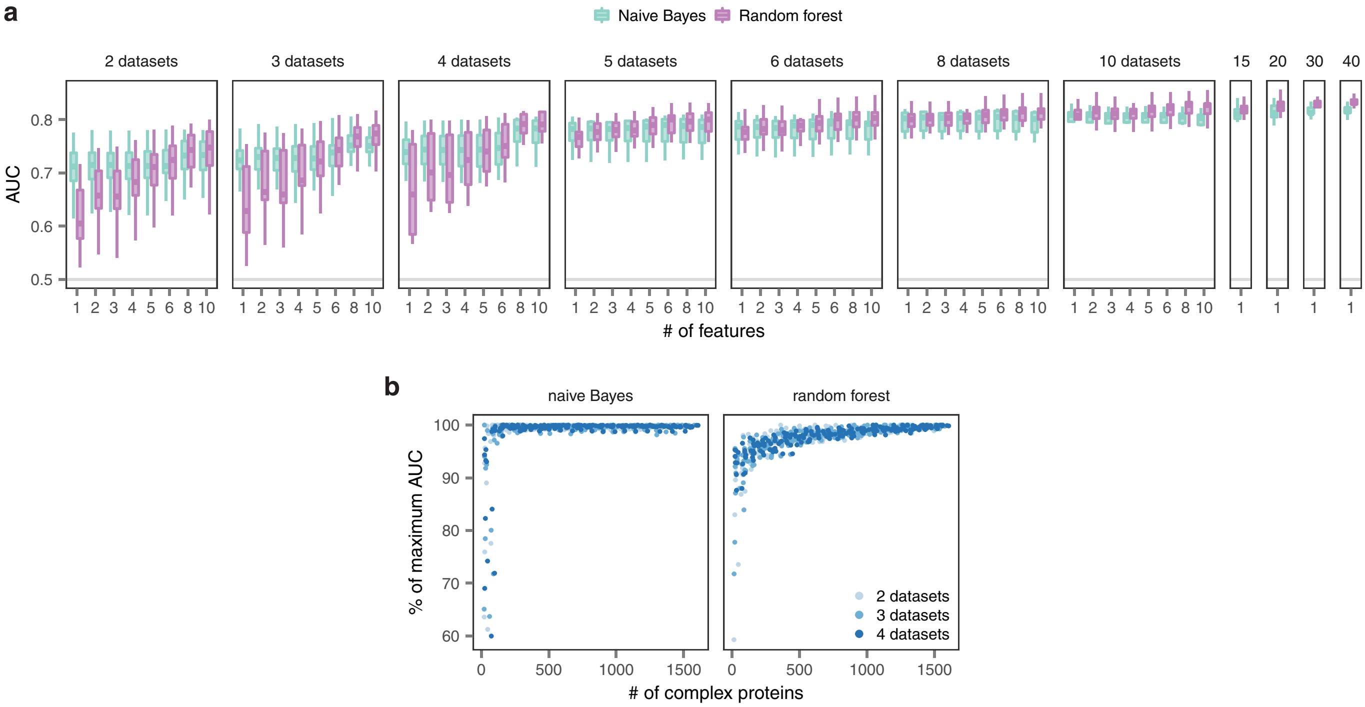
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Machine learning workflows for the integration of multiple CF-MS replicates. a, Schematic overview of cross-validation approaches for CF-MS data. **b**, Comparison of cross-validation by protein pairs or individual proteins in network inference from two to four CF-MS experiments using a naive Bayes classifier, with AUCs calculated in cross-validation or in an independent set of held-out protein complexes. **c**, Impact of feature selection on network inference from two to four CF-MS experiments, comparing between one and six top-performing features, an equivalent number of random features, or five features computed in PrInCE. **d**, Comparison of top-performing or random features in network inference from two to ten CF-MS experiments, using between one and ten top-performing features. **e**, Comparison of network inference with features calculated from concatenated matrices of two to four CF-MS experiments, or with features calculated from individual experiments. **f**, Comparison of network inference from two to four CF-MS experiments using a naive Bayes classifier before and after median imputation of missing values. **g**, Impact of the number of top-performing or random features provided as input on network inference from two to ten CF-MS experiments. **h**, Comparison of random forest and naive Bayes classifiers in network inference from two to ten CF-MS replicates, using between one and ten features. **i**, Network inference from human CF-MS data when integrating varying proportions of SEC and IEX experiments. The total number of CF-MS datasets is shown above the plots, and the number of SEC datasets is shown on the x-axis.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Synergistic and antagonistic feature combinations in network inference from CF-MS data. a, Performance (AUC, left, and rank, right) of naive Bayes and random forest classifiers trained on 24 measures of association in network inference from combinations of between two and six CF-MS datasets. Each cell reflects the mean AUC from 10 random combinations of datasets. **b**, Summary of synergistic and antagonistic interactions between features in CF-MS network inference, as shown in detail in panels **c–f**. Fill reflects the number of times a synergistic (magenta) or antagonistic (cyan) interaction was detected between two features. Network inference was performed using all possible combinations of 24 measures of association from ten random combinations of three or six CF-MS datasets, using either a random forest or naive Bayes classifier. Rows and columns are arranged by the mean performance of individual features across all combinations shown in **a** (both classifiers, two to six datasets). **c**, Performance (AUC) of networks inferred from combinations of three CF-MS datasets using a random forest classifier. Rows and columns are arranged by the mean performance of individual features in the same scenario. Text highlights significantly synergistic (+) and antagonistic (–) interactions. Each cell shows the mean AUC from 10 random combinations of datasets. **d**, As in **c**, but using a naive Bayes classifier. **e**, As in **c**, but for networks inferred from combinations of six CF-MS datasets. **f**, As in **c**, but for networks inferred from combinations of six CF-MS datasets, using a naive Bayes classifier.



Extended Data Fig. 10 | Saturation analysis of network inference from CF-MS data. a, Saturation analysis of network inference from two to 40 CF-MS experiments, using variable numbers of top-performing features. Boxplots show *n* = 10 independent samples. **b**, Impact of downsampling training set complexes on network inference from two to four CF-MS replicates.

nature research

Corresponding author(s): Leonard Foster

Last updated by author(s): 2021-Apr-26

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The data analyzed in this study was obtained from publicly available proteomics repositories (PRIDE and MASSIVE) and re-analyzed with MaxQuant (version 1.6.5.0) and RawTools (version 2.0.2). A complete list of all re-analyzed files and experiments is provided in Supplementary Table 1. Source code and data used to generate MaxQuant parameter files and run MaxQuant searches is available at http://github.com/skinnider/CF-MS-searches.
Data analysis	Data analysis was performed in R (version 3.6.3). Custom code is available from the following repositories: source code used to download and re-analyze publicly available CF–MS data using MaxQuant is available at https://github.com/skinnider/CF-MS-searches. Source code used to carry out the analyses presented in the paper, with relevant intermediate data files, is available from https://github.com/skinnider/CF-MS-analysis. Source code for the CF–MS browser web application is available from https://github.com/skinnider/CF-MS-browser. The CFTK R package is available from https://github.com/fosterlab/CFTK.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

A list of all raw mass spectrometry files analyzed in this study, and their accession numbers in the PRIDE or MASSIVE repositories, is provided in Supplementary

Table 1. All of the data generated in this manuscript is available, at multiple levels of analysis, from the following sources:
- Protein chromatograms and protein-protein interaction networks for up to ten proteins can be visualized and downloaded via an interactive web application at http://cf-ms-browser.msl.ubc.ca.
- Processed chromatograms and MaxQuant proteinGroups.txt files are available via Zenodo at http://doi.org/10.5281/zenodo.4499320.
- Complete MaxQuant outputs for all 206 experiments have been deposited to the PRIDE repository83 with the dataset identifier PXD022048.
- Predicted interactomes for 27 species and clades, including the consensus human CF-MS interactome, are available via Zenodo at https://doi.org/10.5281/zenodo.4245282.
An overview of all of the publicly available resources generated in this study is provided at the supporting website (https://fosterlab.github.io/CF-MS-analysis).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size was determined by reviewing the literature to assemble a list of all published studies with data available on public proteomics repositories.
Data exclusions	Published CF-MS data collected using data-independent acquisition (DIA) was excluded as it could not be re-analyzed using a consistent workflow with the vast majority of published CF-MS data, at the time the analyses were carried out.
Replication	Replication was performed in the sense that the conclusions are based on a meta-analysis of essentially all published experiments with publicly available data at the time of writing.
Randomization	Randomization was not relevant to the study because it focused on the re-analysis of published data.
Blinding	Blinding was not relevant to the study because it focused on the re-analysis of published data.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging