Co-Fractionation / Mass Spectrometry to Identify Protein Complexes

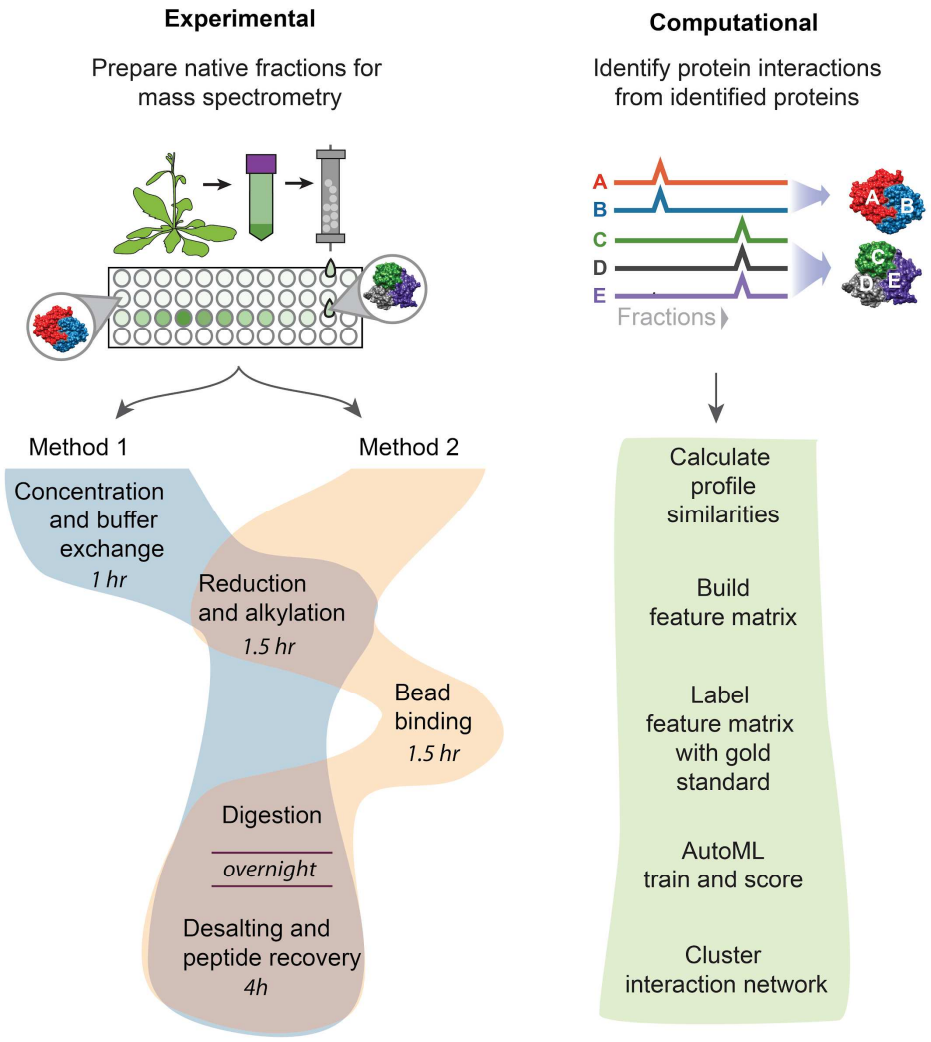
Claire D. McWhite,[1,2,3,*] Ophelia Papoulas,[1,2,3,*] Kevin Drew,[1] Vy Dang,[1] Janelle C. Leggere,[1] Wisath Sae-Lee,[1] and Edward M. Marcotte[1,4,]**
[1]Department of Molecular Biosciences and the Center for Systems and Synthetic Biology, University of Texas, Austin, TX 78712, USA
[2]These authors contributed equally to the work
[3]Technical Contact
[4]Lead Contact
*Correspondence: cmcwhite@princeton.edu, papoulas@austin.utexas.edu
**Correspondence: marcotte@icmb.utexas.edu

SUMMARY

Co-Fractionation/Mass Spectrometry (CF/MS) is a flexible and powerful method to detect physical associations between proteins. CF/MS can be applied to any tissue or organism without the need for antibodies or epitope tagging of individual proteins, distinguishing it from other methods. Here we outline experimental parameters and a new computational pipeline that together allow the successful application of this approach. These protocols are based on our experience with CF/MS of over 16 diverse organisms including plants and animals.

For complete details on the use and execution of this protocol, please refer to McWhite, Papoulas, *et al., Cell* 181(2):460-474.e14 (2020), referred to in text as McWhite, Papoulas *et al. 2020a*.

GRAPHICAL ABSTRACT



BEFORE YOU BEGIN

1. Prepare a native protein extract from your sample using any method optimized for the chosen starting material that avoids using organic solvents or other denaturing compounds. Low concentrations of non-ionic detergents (e.g. up to 1% NP40) can aid cell lysis, however higher concentrations of detergents can dissociate protein-protein interactions and interfere with the subsequent electrospray mass spectrometry.

   Note: if you include detergent and must use ultrafiltration to concentrate your lysate prior to fractionation, be mindful that micelles may also be concentrated and increase detergent concentration with consequent disruption of protein assemblies.

2. Fractionate 1-4 mg of total native protein extract into 12-100 protein fractions using any convenient method such as size exclusion chromatography, ion exchange

chromatography, isoelectric focusing, glycerol gradient separation etc. The following mass spec preparation protocols assume you have collected your fractions in a 96-well format. If you plan to use Method Two make sure that your plate will fit flush against the magnetic plate prior to collecting fractions and make any needed modifications as in the critical note for Method Two.

Pause Point: Collected fractions can be stored frozen at -80°C until you are ready to complete one of the mass spectrometry preparations outlined below.

3. Before beginning to prepare your fractions for mass spectrometry make the necessary solutions from the table below using LC/MS grade water and reagents as much as possible. If you are not using a commercially prepared solution of TCEP, prepare a 0.5 M stock in water (this will require pH adjustment for proper solubility). If you will be using Method Two additionally prepare and aliquot SpeedBeads (as described below).
4. Before beginning the computational pipeline, have a table of protein identifications for each fraction and a set of known gold standard protein complexes.
5. For maximal power in deriving protein-protein interactions, ideally have performed multiple separations using different separation techniques, as CF/MS gains substantial statistical power from reproducible co-elution behavior of interacting proteins across otherwise distinct separations.

Prepare Speedbead Slurry for Method Two

Timing: 45 minutes

6. Completely resuspend each of the 2 types of commercial beads (see Key Resources Table) provided as a 50 mg/ml slurry containing preservative. Be thorough with vortexing and pipetting to ensure homogeneity.
7. Pipette 50 µl of each commercial bead suspension into a single 1.5 ml microcentrifuge tube and mix well.
8. Collect the beads by low speed pulse centrifugation in a microfuge and remove and discard the supernatant.
9. Wash the beads by resuspension in 1 ml LC/MS-grade dH_20.
10. Repeat steps 8 and 9.
11. Collect the beads as in step 8 and resuspend in 500 µl LC/MS-grade dH_20 for the working concentration of 10 µg/µl total (5 µg/µl each bead type). This slurry can be aliquoted for ease of resuspension, and stored at 4°C (do not freeze) for at least 6 months. Resuspend beads thoroughly before use.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER

Chemicals, Peptides, and Recombinant Proteins		
1 M Tris-HCl, pH 8.0	No preference	N/A
$CaCl_2$	No preference	N/A
Tris(2-carboxyethyl)phosphine (TCEP)	No preference	N/A
Iodoacetamide	No preference	N/A
2,2,2-Trifluoroethanol (TFE)	No preference	N/A
Dimethyl Sulfoxide (DMSO)	No preference	N/A
Ethanol	No preference	N/A
Formic Acid	No preference	N/A
Acetonitrile (LC/MS grade)	No preference	N/A
Water (LC/MS grade)	No preference	N/A
Trypsin, Mass Spectrometry grade	No preference	N/A
Dithiothreitol (DTT)	No preference	N/A
Software and Algorithms		
cfmsflow	This paper, https://github.com/marcottelab/cfmsflow	
TPOT	Olson *et al.*, 2016	Version >= 0.10.0
Nextflow	Di Tommaso *et al.*, 2017	
Docker	https://www.docker.com/	
Other		
AcroPrep Advance 96-filter plate 3k MWCO	Pall	8163
QIAvac 96 or QIAvac Multiwell vacuum manifold	Qiagen	19504 or 9014579
SpeedBead Magnetic Carboxylate Modified Particles	GE Healthcare, UK	45152105050250

SpeedBead Magnetic Carboxylate Modified Particles	GE Healthcare, UK	65152105050250
Lab-in-a-plate Flow-thru plate, 5-7 µl C18	Glygen Corp.	MFNSC18.10
Magnetic Plate, 96-well separator	Thermo Fisher Scientific	A14179
Vacufuge	No preference	N/A
Deepwell 96-well plate for autosampler	Waters	186005837
Silicone plugseal gasket with slit for autosampler	Waters	186006332
1.5 ml polypropylene microcentrifuge tubes	No preference	N/A
V-bottom, 0.45ml, 96-well polypropylene microplates	No preference	N/A
Low speed (1000 x g) centrifuge with microplate adaptors	No preference	N/A
Organic-resistant plate sealing film	Eppendorf	0030127870
Plate sealing foil	Eppendorf	0030127889
Bacterial and Virus Strains		
N/A	N/A	N/A
Experimental Models: Organisms/Strains		
Arabidopsis thaliana	McWhite, Papoulas *et al.* 2020a	N/A
Brassica oleracea var. italica	McWhite, Papoulas *et al.* 2020a	N/A
Cannabis sativa	McWhite, Papoulas *et al.* 2020a	N/A
Ceratopteris richardii	McWhite, Papoulas *et al.* 2020a	N/A
Chenopodium quinoa	McWhite, Papoulas *et al.* 2020a	N/A
Chlamydomonas reinhardtii	McWhite, Papoulas *et al.* 2020a	N/A
Cocos nucifera	McWhite, Papoulas *et al.* 2020a	N/A

Glycine max	McWhite, Papoulas *et al.* 2020a	N/A
Homo sapiens	Mallam, *et al*., 2019	N/A
Mus Musculus	Mallam, *et al*., 2019; Liebeskind *et al*., 2020	N/A
Oryza sativa	McWhite, Papoulas *et al.* 2020a	N/A
Selaginella moellendorffii	McWhite, Papoulas *et al.* 2020a	N/A
Solanum lycopersicum	McWhite, Papoulas *et al.* 2020a	N/A
Triticum aestivum	McWhite, Papoulas *et al.* 2020a	N/A
Xenopus laevis	Drew, *et al.* 2020	N/A
Zea mays	McWhite, Papoulas *et al.* 2020a	N/A

MATERIALS AND EQUIPMENT

The two alternate experimental methods below each begin with separated protein fractions in a 96-deep well plate and are described briefly in McWhite, Papoulas et al. 2020a See the Key Resources Table for necessary specialty plates and equipment described in the protocols. Both methods require a 96-well plate-compatible vacuum manifold and a vacufuge (e.g. Eppendorf #022820044) with a microplate adaptor. Method Two additionally requires a magnetic slab (magnetic plate) and may require cutting of 4 support veins of plate plastic to allow the plate to sit properly on the magnet. Multichannel pipettes are helpful throughout. Solutions to make up in advance are listed in the following table with recipes below.

Solutions stored 21-25°C	needed for Method 1	needed for Method 2
Buffer A (recipe below)	✓	✓
Buffer B (recipe below)	✓	✓

Buffer C (recipe below)	✓	✓
60% Buffer B (recipe below)	✓	✓
Trypsin Digestion Buffer (recipe below)	✓	✓
70% Ethanol	N/A	✓
2% DMSO	N/A	✓
10% TFE in Trypsin Digestion Buffer	N/A	✓
0.5 M TCEP	✓	✓
Aliquoted frozen stock solutions		
1 M DTT (store -20°C)	✓	✓
550 mM Iodoacetamide (store -20°C, protect from light)	✓	✓
Trypsin 1 µg/10 µl in 10 mM acetic acid (store -80°C)	✓	✓

Reagent	**Final Concentration LC/MS water**	**Final Concentration LC/MS Acetonitrile**	**Final Concentration Formic Acid**
Buffer A	100%	0 %	0.1%
Buffer B	0%	100%	0.1%
Buffer C	95%	5%	0.1%
60% B	40%	60%	0.1%

Trypsin Digestion Buffer	**Final Concentration**
Tris-HCL pH 8.0	50 mM
$CaCl_2$	2 mM

STEP-BY-STEP METHOD DETAILS

Prepare Protein Fractions for Mass Spectrometry: Method 1, Ultrafiltration and in-solution digest.

We use two different methods for reduction/alkylation and trypsin digestion of samples in 96-well plates for mass spectrometry. If your sample contains < 300 mM salt, you can use Method 2 (see Step 5 below). Otherwise you should use Method 1 directly below.

Timing: 2 days

1. Buffer exchange and sample concentration.
    a. Using a vacuum manifold, wash preservatives from a 3k MWCO AcroPrep Advance 96 filter plate (see Key Resources Table) by flowing first 300 µl dH20, and then 300 µl Trypsin Digestion Buffer through the membrane.

        Note: If all 96 wells are not required place sealing film over unused wells. The plate can be saved, and the unused wells uncovered for use at a later date. Throughout all washes and subsequent steps do not let the wells dry out or proteins may stick irreversibly to the membrane.

    b. Stop the vacuum and shake excess liquid from the washed plate.
    c. Transfer your samples to the washed AcroPrep plate. Vacuum filter to reduce the sample volume to ~100 µl.

        Note: Sample wells will filter at unequal rates due to varied sample content. Monitor volumes and add Trypsin Digestion Buffer to samples as necessary to prevent rapidly flowing wells from drying before all wells are adequately concentrated.

    d. Add 100 µl Trypsin Digestion Buffer to all wells, pipetting up and down to mix.
    e. Continue vacuum filtration to ~ 50 µl sample volume.
    f. Transfer concentrated samples back into the original deep well plate for digestion.
2. Reduction/Alkylation
    a. To the 50 µl sample add 50 µl TFE.

        Note: Even if your sample is >50 µl do NOT add more than 50 µl TFE or the final concentration at the digestion step could inhibit trypsin activity.

    b. Add TCEP to 5 mM (from a 0.5 M stock in water), seal the plate with clear film, and incubate for 30 min at 37°C.
    c. Remove plate to room temperature (20-25°C) and add iodoacetamide to 15 mM.
    d. Seal plate and incubate for 30 min in the dark, at room temperature.

Note: Iodoacetamide is light sensitive and reactive in aqueous solution. We generally store small aliquots of 550 mM stock solution of iodoacetamide in water at -20°C. These are thawed directly prior to use in this protocol. Alternatively, a fresh stock can be made from powder directly before use.

e. Quench the unreacted iodoacetamide by the addition of DTT to 7.5 mM.

3. Digestion
    a. Add ~880 µl Trypsin Digestion Buffer to bring each sample to 1 ml volume to reduce TFE to <5%.
    b. Add 1 µg of proteomics grade Trypsin to each well.
    c. Seal the plate with clear film and digest overnight at 37°C.
    d. Stop the digest by addition of formic acid to 0.1%
4. Desalting and Peptide recovery

Note: This step uses the solid phase C18 Lab-in-a-Plate indicated in the Key Resources Table. As with the ultrafiltration plates, unused wells should be sealed with tape during plate use and can be saved for future use.

a. Condition the C18 Lab-in-a-Plate using vacuum filtration as follows:
    i. 100 µl 60% Buffer B per well. Repeat.
    ii. 100 µl Buffer A per well. Repeat 3-4 times.
b. Apply your 1 ml sample and filter through.
c. Wash by filtration of 100 µl Buffer A. Repeat 2 times.

Note: If your sample initially had very high salt add some additional Buffer A washes here as the goal is to de-salt the sample prior to mass spectrometry.

d. Place a 96-well collection plate into the appropriate vacuum manifold adaptor to collect the eluate.

CRITICAL: Strong vacuum can lead to bubbling/foaming of the eluting liquid causing cross-contamination of samples so pay attention at this step and adjust vacuum accordingly. Alternatively, samples can be eluted by using gentle centrifugation in a swinging bucket rotor equipped with microplate adaptors. During centrifugation, the C18 plate is directly seated on top of the collection plate.

e. Elute the digested peptides into the collection plate using 50 µl 60% Buffer B per well. Repeat for a total eluate volume of 100µl.
f. Evaporate the eluate in a vacufuge to dryness. This step takes approximately 3 hours.

Pause Point: If you will not be able to load samples onto a mass spectrometer within ~24 hours these samples can be stored dry, sealed with adhesive foil, -80°C.

g. For mass spectrometry, resuspend peptides in 20-35 µl Buffer C per well, and transfer them to the appropriate sample vial or plate for your autosampler system.

Note: We use Waters deep 96-well plates with a pre-slit silicone plugseal gasket as listed in the Key Resource Table. 5 µl of each sample is injected for mass spectrometry. Unused sample can be stored for a period of weeks to months in sealed plates -80°C but solvent may evaporate.

Prepare Protein Fractions for Mass Spectrometry: Method 2, Bead Binding and on-bead digest.

Method 2 for reduction/alkylation and trypsin digestion can be used if your sample contains < 300mM salt. Otherwise you must use Method 1 (see Step 1 above).

Critical: Ensure that bottom of wells can make good contact with the magnetic slab (in our case this requires cutting notches with a single-edge razor blade in 4 support ribs of the deep well plate).

Timing: 2 days

5. Reduction and Alkylation
    a. To each well containing your samples add TFE to 20% final concentration and mix by pipetting.
    b. Add TCEP to 5 mM.
    c. Seal the plate with adhesive film and incubate 45 min at 37°C.
    d. Remove the plate to room temperature (20-25°C).
    e. Add iodoacetamide to 25 mM (see note at step 2d above).
    f. Incubate for 30 min in the dark at room temperature.
    g. Quench reactivity by addition of DTT to 12 mM.
6. Bead binding of proteins
    a. To each fraction add 4 µl of bead suspension. The bead suspension is a stock slurry comprising a 1:1 mix of 5 µg/ml each of the two bead types listed in the Key Resources table.
    b. Add formic acid to 2% and acetonitrile to 50% and mix well by pipetting. These two ingredients can be made into a premix for easier dispensing.
    c. Mix gently but thoroughly by pipetting.
    d. Incubate for 30 min with gentle shaking or rocking to keep beads from settling.

        Crucial: incubation for prolonged periods of time does not help and may reduce recovery.

    e. Collect beads from this large volume by brief centrifugation (e.g. 5 min, 1000 x g) in a swinging bucket rotor equipped with deep well plate adaptors.
    f. Place the sample plate on the magnetic plate. Remove and discard the bulk of the liquid leaving behind the beads and approximately 250 µl liquid.
    g. Take the sample plate off the magnetic plate. Resuspend the beads in the remaining liquid and transfer them to a fresh 450 µl conical bottom 96-well plate to facilitate the remaining steps.

h. Place this shallow plate on the magnetic plate. Once beads have collected to the bottom, remove and discard the remaining liquid.

> **Note:** beads will vary in appearance from rust color when diffuse to darker brown when compact, and appear more aggregated or dispersed depending on protein content and other unknown parameters. None of these appearances indicates failure to bind proteins.

i. Keep the sample plate on the magnetic plate for all the following wash steps and work rapidly. Pipette carefully to avoid losing beads. The washes need not resuspend the bead pellet.
   i. Wash with 100 µl per well 70% ethanol. Repeat.
   ii. Wash with 100 µl acetonitrile. Repeat
   iii. Air dry beads briefly.

7. On-bead digestion of proteins
   a. Remove the sample plate from the magnetic and resuspend the beads in 25 µl 10% TFE/90% Trypsin Digestion Buffer.

   > **Note:** Do not be concerned if beads appear "chunky" or aggregated at this stage.

   b. Dilute enough trypsin stock solution with trypsin digestion buffer for the next step.
   c. To each sample add 25 µl of Trypsin Digestion Buffer containing 0.25 µg proteomics grade Trypsin.
   d. Seal the plate with adhesive film and incubate overnight 37°C.
8. Peptide recovery
   a. Place the sample plate on the magnet. Once the beads have collected transfer the bead-free supernatant to a fresh 96-well plate.
   b. To the transferred volume add formic acid to 1% to stop digestion.

   > **Note:** Formic acid is very volatile and corrosive to metal parts of pipettors (e.g. plungers) at concentrations > 10%. If you are pipetting undiluted formic acid we recommend using positive displacement pipettors.

   c. Remove the plate from the magnet and elute peptides from the beads with the addition of 50 µl 2% DMSO.

   > **Note**: If possible, cover samples with adhesive film and sonicate 5 min in a bath sonicator to assist elution at this point.

   d. Replace the plate on the magnet to collect the beads. Remove each sample eluate and add it to the corresponding supernatant from step 8a.
   e. Remove the plate from the magnet and elute the beads with another 50 ml 2% DMSO (no sonication necessary).
   f. Replace the bead plate on the magnet. Remove this second eluate and pool it with the first corresponding matched eluate for a total volume of 150 µl for each fraction.

g. Desalting of the eluted peptides is identical to that in Method 1 (step 4) with the only difference being the sample volumes loaded on the conditioned C18 plate (150 µl or 1 ml).

EXPECTED OUTCOMES

Prepared plates are ready for protein detection and quantification by mass spectrometry. We typically do not bother to prepare fractions corresponding to baseline UV signal in a chromatographic separation. In cases where a particular separation experiment is repeated we will also exclude those particular fractions that previously contained insufficient material to confidently identify proteins. Examples of machine settings for mass spectrometry are provided in McWhite, Papoulas *et al* 2020a Supplemental Methods.

QUANTIFICATION AND STATISTICAL ANALYSIS

Determine protein interactions from protein elution profiles

After fractionation, proteins in each fraction are identified and quantified by mass spectrometry. This information is used to create elution profiles across all fractions for each protein observed. The computational protocol identifies sets of proteins which have related elution profiles, signifying that these sets are physically associating.

To simplify the computational process of detecting protein interactions from CF/MS data, we provide cfmsflow, a Nextflow pipeline for Linux systems. This pipeline takes as inputs identified protein profiles (elution profiles) and uses known protein interactions (provided by the user) to train a model of elution profile similarity.

The pipeline is divided into 5 main steps (Figure 1). To summarize, these steps are to 1) Calculate similarities between protein elution profiles, 2) Combine similarity scores into a feature matrix, 3) Label the feature matrix with gold standard protein-protein interactions, 4) Train a model to detect and score pairwise protein interactions, 5) Cluster the resulting protein interaction network into complexes.

By default, the pipeline begins at step 1 (calculating protein profile similarities) and continues through step 5 (detect complexes), however, a subset of steps or individual steps may be run as set in the user parameter file.

BEFORE YOU BEGIN ANALYSIS

1. Install nextflow onto a Linux-based operating system (https://www.nextflow.io/docs/latest/getstarted.html), then retrieve the cfmsflow

pipeline from github: `git clone https://github.com/marcottelab/cfmsflow.git`

2. Organize protein identifications for each fractionation plate into a comma separated table with header, where the first column (named ID) contains protein identifiers, and the following columns contain protein quantification in each fraction. Values can be any abundance metric, such as Peptide Spectral Matches, peptide peak area or precursor ion intensity. Elution files from one or more separation experiments are the main data inputs to cfmsflow.

3. Obtain a file of gold standard protein complexes, such as from the CORUM database. This file should be formatted with one protein complex per line, and each protein in the complex separated by a space.

4. Check the notes and critical points below carefully before proceeding to step 1 of the pipeline.

Note: Examples of all input file formats are provided in the test_input directory of the github repository.

Note: For discrete count-based measures such as peptide-spectral matches (PSMs), correlations may be run for N repetitions with poisson noise, and scores averaged, controlled by the parameter `` `added_poisson_reps = N` ``. Intensity-based measures should not be run with added poisson noise.

Critical: The basic usage of the pipeline is:
`nextflow main.nf -params-file user_parameters.json`

Note: To resume running a pipeline after fixing an error or changing a parameter value, run the same command as above with additionally –resume to run all steps downstream of the parameter change.

Critical: Before running cfmsflow on real data, test the pipeline by running the provided example. Look for successful completion after approximately five minutes and for output to appear in the test_output/ directory.
`nextflow main.nf -params-file example_params/example_wholepipeline.json`

Timing: ~24 hours for a dataset containing 8000 proteins. Varies by computational resources

STEP-BY-STEP ANALYSIS DETAILS

1. Calculate features
    a. Correlation and distance metrics are calculated between all possible pairs of proteins in each elution file.

b. Input: File containing paths to elution files, or glob pattern matching paths to elution files.

Note: Reduce feature length by pre-filtering to only include well observed proteins, or proteins that are observed in at least n experiments.

2. Build a feature matrix
    a. Features are joined into a single table
    b. Input: Output of previous step, or file containing paths to feature files, or glob pattern matching paths to feature files.

3. Label the feature matrix with gold standard interactions
    a. A column 'label' is added to the feature matrix, with a value of 1 if a positive training interaction and -1 if a negative interaction
    b. Input: Output of previous step, or path to a feature matrix file
    c. Input: Path to gold standard interactions

Note: Training labels may be either provided by the user or generated from input gold standard complexes. Negative interactions can either be drawn from observed interactions (as in McWhite, Papoulas, *et al.* 2020a) or from faux interactions between different gold standard complexes (as in Drew 2017), as controlled by the parameter `` `negatives_from_observed = true/false` ``.

4. Train a model to score protein-protein interactions
    a. The TPOT AutoML software scans machine learning pipelines and parameters to determine optimal model parameter settings based on positive and negative labeled training interactions.
    b. A model is trained on the same training interactions using the parameter settings determined by TPOT
    c. This model is applied to the full feature matrix to give a CF-MS score to all pairs of proteins in the feature matrix.

Note: Interactions from the same gold standard complex are prevented from being split between different cross validation portions

Note: To reduce overtraining effects, lower the value of the `max_features_to_select` parameter. The model will select no more than this number of features. After modifying this parameter, rerun the nextflow command with an additional –resume flag to avoid rerunning already complete portions of the pipeline.

5. Cluster interactions
    a. Scored interactions are thresholded at a parameter-input false discovery rate threshold (default 0.1), then clustered into protein complexes with diffusion clustering (see Supplemental Methods section of McWhite, Papoulas *et al.* 2020a).

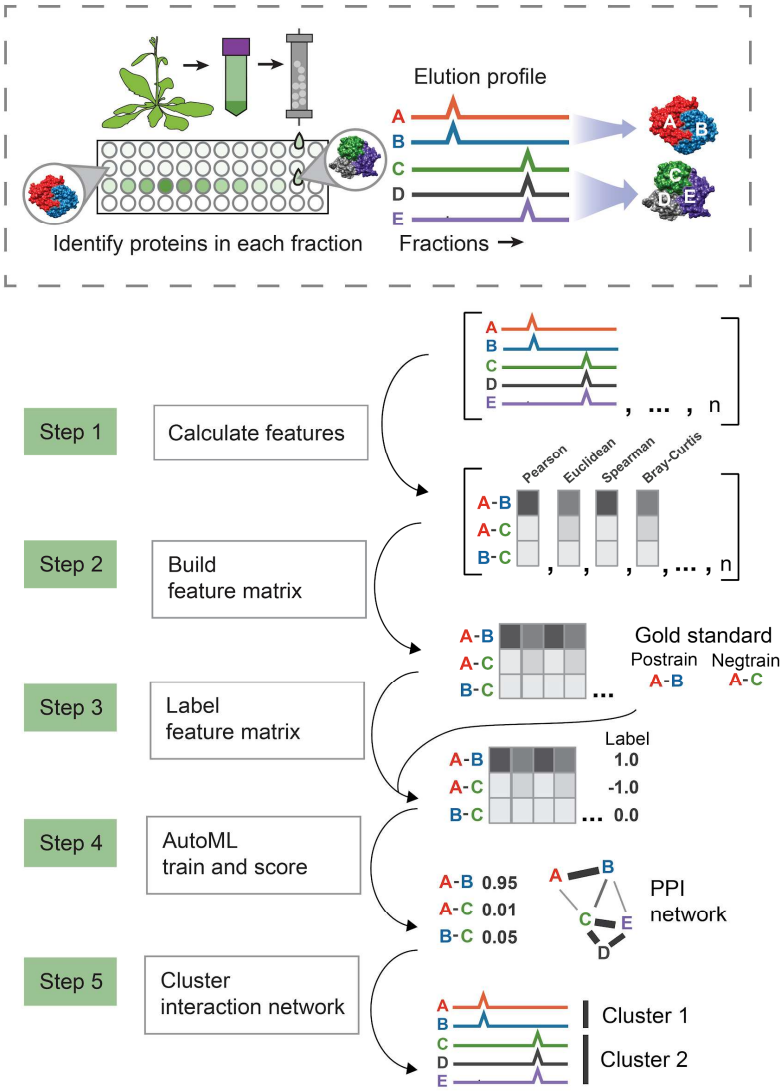


Figure 1.
Overview of computational pipeline to detect protein-protein interactions and protein complexes. In step 1, a set of similarity scores between all proteins are calculated for each fractionation. In step 2, these similarity scores are combined into one large table. In step 3, pairs of proteins that are known from prior literature to interact are labelled with a 1 (positive training label), and a set of random pairs of proteins are labeled with a -1 (negative training label). In step 4, a model is trained to distinguish these positive and negatively labeled pairs of proteins, giving a score to each pair, where a higher score indicates higher probability of interaction. In step 5, this interaction network is clustered to protein complexes.

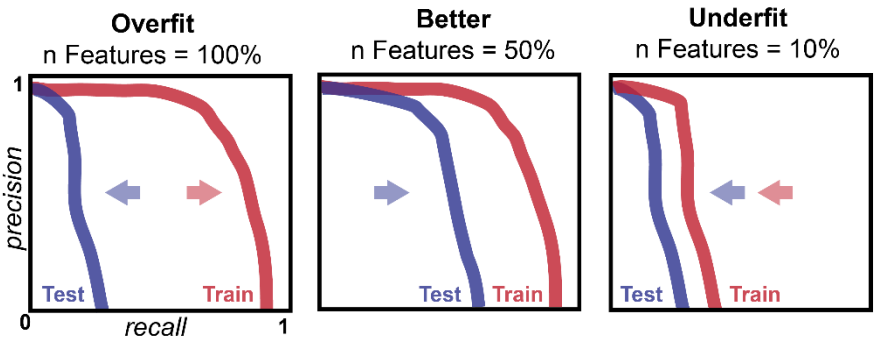


Figure 2. Using precision-recall curves to evaluate overfitting. Precision-recall curves illustrating overfit, better fit, and underfit models. A substantial difference between Test and Train precision recall curves suggests overfitting, and that the max number of features should likely be lowered to improve model performance. When a model is overfit, as features are removed, the Test precision-recall curve will shift right, while the training curve is minimally affected. Once too many features have been removed, performance in both test and training will decline.

LIMITATIONS

In our experience, starting with samples that contain too little protein is the biggest cause of failure. Fractionations beginning with less than 1 mg total protein generally produce poor results. We recommend aiming for 1-4 mg total protein at a concentration of ≥ 5 mg/ml. Failure to reach this amount can occur when beginning with too little sample. After grinding the samples on liquid nitrogen and prior to protein extraction, aim for at least a 1 mL volume of frozen powder, ideally 5 mL. Likewise, failure to extract sufficient protein can occur when the sample inherently has a very low proportion of protein per overall mass (e.g. starch-filled seeds) or a high amount of confounding non-proteinaceous substances (e.g. abundant mucous). In these cases it is important to use extraction methods specialized for the particular organism or tissue.

TROUBLESHOOTING

Problem:

Pipeline fails to run

Potential Solution:

First, confirm that the example parameter json (`example_params/example_wholepipeline.json`) runs to completion. If it does not, use provided example parameter jsons to test and troubleshoot individual steps of the pipeline. Next, confirm that the file format of your input files exactly matches the corresponding example input files, including delimiters and column names when specified. Confirm at the command line that file paths in your parameter json exist. If missing parameters are warned, add those parameters to your parameter json.

Problem:

Overfit model

Potential Solution:

When a model is overfit to the training data, it fails to generalize to data not used to train the model. Overfitting can be visually diagnosed by comparison of the precision-recall curves of training and test of interactions (Figure 2). While models will generally perform better on training labels, a large gap in performance between training and test precision-recall curves is diagnostic of an overfit model. Overfitting can often be reduced by reducing the number of features that are used to construct the model. The initial run of the pipeline will produce a `.featureimportances` file after step 4, which contains feature importances determined by Random Feature Elimination. To rerun the TPOT pipeline limiting the maximum number of features, modify the `max_features_to_select` value in your user parameters json to a smaller number than the total number of features and resume the pipeline with `nextflow main.nf -params-file user_parameters.json -resume` or create a new parameters file that begins at step 4.

Problem:

Too few or too many proteins in the output clustering file. Large complexes broken apart.

Potential Solution:

The number of interactions forwarded into clustering is controlled by the `fdr_cutoff` value in your user parameters json, which thresholds interactions with CF-MS scores above a particular false discovery rate. False discovery rate is calibrated from test positive and negative labeled interactions.

A high FDR cutoff that uses more interactions can cause an overly dense network that, in our experience, can lead to less cleanly distinguished complexes after clustering. However, an overly stringent FDR cutoff reduces recall of complexes.

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Edward Marcotte (marcotte@icmb.utexas.edu).

Data and Code Availability

The cfmsflow pipeline is available at https://github.com/marcottelab/cfmsflow. Example data is provided along with the pipeline.

ACKNOWLEDGMENTS

Research was funded by grants from the Welch Foundation (F-1515 to E.M.M.); NSF (1237975 to E.M.M.); Army Research Office (W911NF-12-1-0390); and NIH (GM123683 to C.D.M., K99 HD092613 to K.D., R35 GM122480 and R01 HD085901 to E.M.M.).

AUTHOR CONTRIBUTIONS

AUTHOR CONTRIBUTIONS

Writing - Experimental methodology, O.P.; Writing - Computational methodology and pipeline, C.D.M, Scripts, C.D.M. and K.D. Pipeline testing, V.D., W.S, J.C.L., Funding Acquisition, C.D.M., K.D., and E.M.M.; Supervision, E.M.M.

DECLARATION OF INTERESTS

The authors have no competing interest to declare

REFERENCES

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nature Biotechnology, 35(4), 316–319. doi:10.1038/nbt.3820

Drew, K., Lee, C., Cox, R. M., Dang, V., Devitt, C. C., Papoulas, O., Huizar, R. L., Marcotte, E. M., Wallingford, J. B. (2020). A systematic, label-free method for identifying RNA-associated proteins in vivo provides insights into vertebrate ciliary beating, Developmental Biology, in press. doi:10.1016/j.ydbio.2020.08.008

Liebeskind, B. J., Young, R. L., Halling, D. B., Aldrich, R. W., Marcotte, E. M. (2020) Mapping functional protein neighborhoods in the mouse brain. bioRxiv, doi:10.1101/2020.01.26.920447

Mallam, A. L., Sae-Lee, W., Schaub, J. M., Tu, F., Battenhouse, A., Jang, Y. J., Kim, J., Finkelstein, I. J., Marcotte, E. M., Drew, K. (2019) Systematic discovery of endogenous human ribonucleoprotein complexes. Cell Reports, 29(5):P1351-1368.e5 doi:10.1016/j.celrep.2019.09.060

McWhite, C. D., Papoulas, O., Drew, K., Cox, R. M., June, V., Dong, O. X., Kwon, T., Wan, C., Salmi, M. L., Roux, S. J. Jr., Browning, K. S., Chen, Z. J., Ronald, P. C., Marcotte, E. M. (2020). A pan-plant protein complex map reveals deep conservation and novel assemblies. Cell, 181(2), 460-474.e14.

Olson, R.S., and Moore, J.H. (2016). TPOT: A Tree-based Pipeline Optimization Tool for Automating Machine Learning. In Automated Machine Learning, F. Hutter, L. Kotthoff, and J. Vanschoren, eds. (Springer), pp. 66–74.