

# **MS-based Interactomics:** Computational resources and tools for studying the physical interactome

**ASMS Bioinformatics MS Interest Group**

Wednesday Evening Workshop

Isabell Bludau & Bill Noble

# What is 'interactomics' and why do we discuss it?

- Many MS-omics studies focus on **cataloging and quantifying individual molecules** of a particular type
  - e.g. quantitative protein or metabolite matrix
- Most biological **molecules don't operate in isolation** but they interact with each other
  - protein complexes
  - activity regulation via metabolite/drug binding
- **'Interactome'** = **comprehensive** set of **molecular interactions** in biological system
  - here we focus only on **physical** (not functional) interactions

# Current MS-based techniques for large-scale interactomics

Protein-protein interaction (PPI) networks:

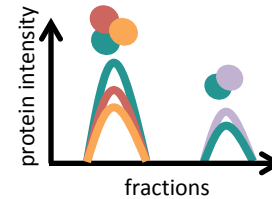
- **Affinity-purification MS (AP-MS)**
- Proximity-dependent labeling: APEX, BioID



⇒ Interaction network

Protein-protein complexes:

- **Protein co-fractionation MS (CoFrac-MS)**



⇒ Protein complexes

Structural information on PPIs:

- Cross-linking MS

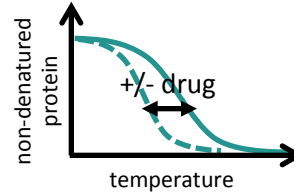


⇒ Structure: interacting protein residues

# Current MS-based techniques for large-scale interactomics

Protein-metabolite/drug interactions:

- **Thermal proteome profiling (TPP)** / Cellular Thermal Shift Assay (CETSA)
- Limited proteolysis-coupled MS (LipMS)



⇒ Protein-ligand interactions

Protein-RNA interactions:

- Protein-RNA crosslinking



⇒ Protein-RNA interactions at residue resolution



# Available resources and databases

## Databases based on:

- High-throughput methods
- Low-throughput assays
- Computational predictions

## Goals:

- Expand current knowledge
- Use knowledge from existing databases
- Look at interactome changes and dynamics

## Databases:

### Protein-protein interactions (PPIs)

- STRING (<https://string-db.org/>)
- BioPlex (<http://bioplex.hms.harvard.edu/>)
- PrePPI (<https://honiglab.c2b2.columbia.edu/PrePPI/>)

### Protein complexes

- CORUM (<https://mips.helmholtz-muenchen.de/corum/>)
- Complex Portal (<https://www.ebi.ac.uk/complexportal/home>)

### Protein-metabolite/drug interactions

- STITCH (<http://stitch.embl.de/>)

### General interactions:

- IntAct (<https://www.ebi.ac.uk/intact/>)

# Computational challenges and what we would like to discuss today

- **Protein-protein interaction networks: affinity-purification MS (AP-MS)**

- Many reciprocal pull-downs to map the full PPI network of a cell
- 100s - 1000s of MS measurements > prevent error accumulation
- Confidently distinguish true from false interactions

*Eduard  
Huttlin*

- **Protein-protein complexes: Co-fractionation MS (CoFrac-MS)**

- Comprehensive interactome map from a single experiment (< 100 MS measurements)
- Distinguish true interactions from random co-elutions

*Isabell  
Bludau*

- **Protein-drug interactions: thermal proteome profiling (TTP)**

- Many metabolites are tested against thousands of proteins
- How to estimate significance?

*Dominic  
Helm*

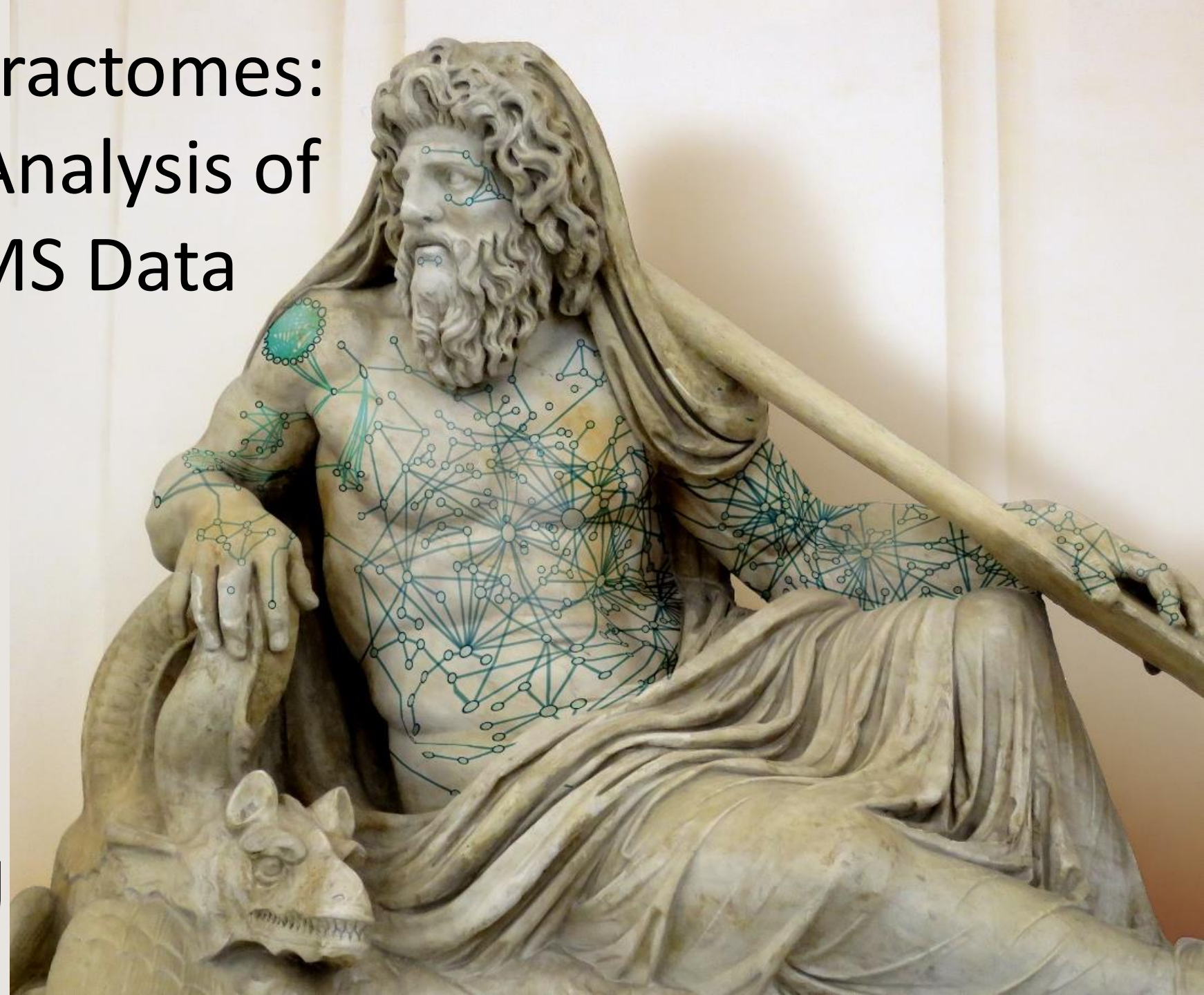
Edward Huttlin

Harvard Medical School

# From IP's to Interactomes: Computational Analysis of Large-scale AP-MS Data

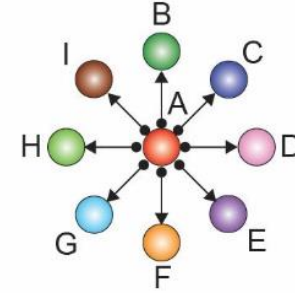
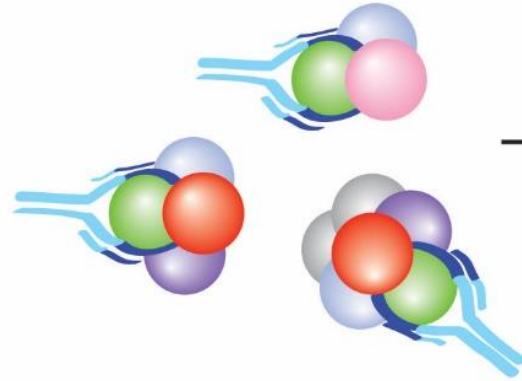
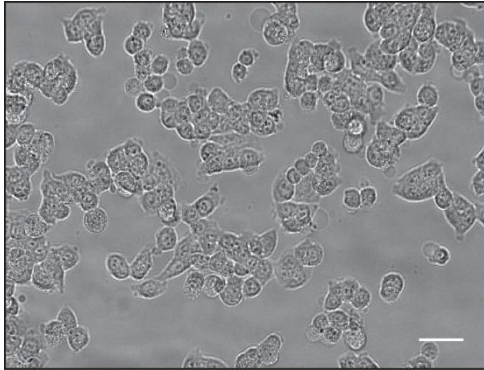
Ed Huttlin

Harvard Medical School  
Department of Cell Biology  
Boston, Massachusetts

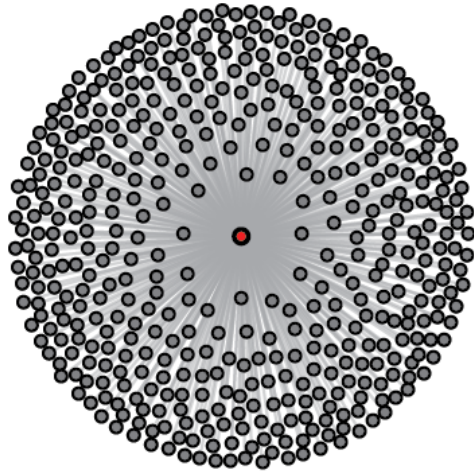


# The Achilles Heel of AP-MS: Which Interactions are Real?

## Affinity-Purification MS

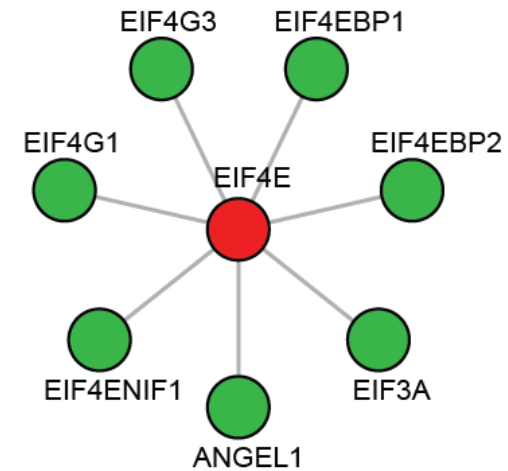


Unfiltered AP-MS Data



468 Proteins Identified

Filtered AP-MS Data

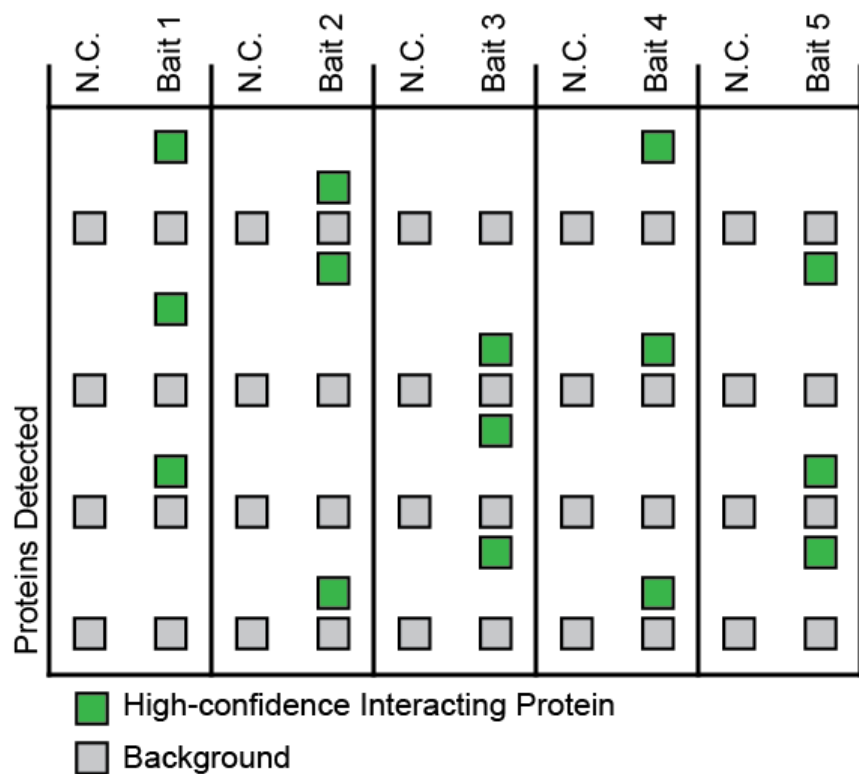


7 High-Confidence Interacting Proteins



# Computational Strategies for Identifying Interacting Proteins

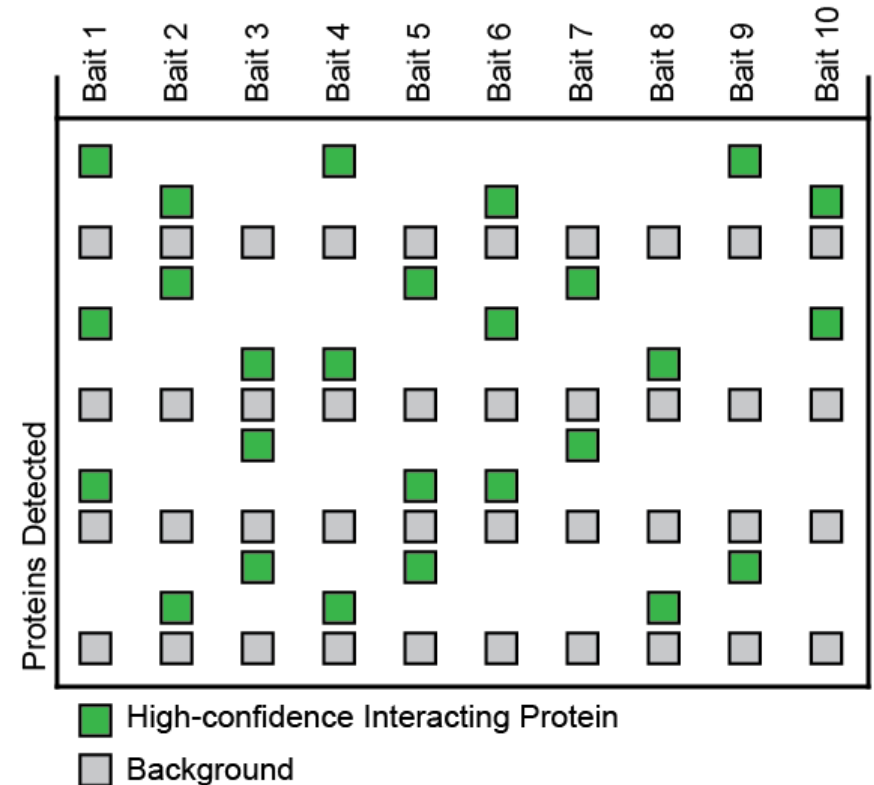
## COMPARISON WITH NEGATIVE CONTROLS



*Negative control IP's are used to Define background. Proteins present At higher abundance than background Are assumed to be interactors.*

**Examples:** QUBIC, SAINT, many others

## COMPARISON ACROSS UNRELATED IP's



*Background is assumed to be constant across IP's. Interacting proteins are found by seeking proteins Whose abundance is increased above their Average across many IP's of unrelated baits.*

**Examples:** CompPASS, SAINT, MiST, HGScore, Socio-affinity Index

# CompPASS Scoring Algorithm

Stats Table

	Bait 1	Bait 2	Bait 3	Bait 4	Bait k
Interactor 1	$X_{1,1}$	$X_{2,1}$	$X_{3,1}$	$X_{4,1}$	$X_{k,1}$
Interactor 2	$X_{1,2}$	$X_{2,2}$	$X_{3,2}$	$X_{4,2}$	$X_{k,2}$
Interactor 3	$X_{1,3}$	$X_{2,3}$	$X_{3,3}$	$X_{4,3}$	$X_{k,3}$
Interactor 4	$X_{1,4}$	$X_{2,4}$	$X_{3,4}$	$X_{4,4}$	$X_{k,4}$
Interactor m	$X_{1,m}$	$X_{2,m}$	$X_{3,m}$	$X_{4,m}$	$X_{k,m}$

$X_{i,j}$  = Spectral counts for interactor  $j$  with bait  $i$

Is the protein more abundant than usual in this IP?

**Z-Score**

$$Z = \frac{X_{i,j} - \bar{X}}{S}$$

**WD<sup>N</sup>-Score**

$$WD^N = \sqrt{\left[\frac{N}{n} \frac{S}{\bar{X}}\right]^p X_{i,j}}$$

How frequently is this protein detected? Is it detected reproducibly?

$\bar{X}$  = Average spectral counts for protein across baits  
 $X_{i,j}$  = Spectral counts for interactor  $j$  with bait  $i$   
 $S$  = Standard Deviation spectral counts across baits

$N$  = Number of runs  
 $n$  = Number of runs in which the protein is found  
 $p$  = Number of replicates observed (1 or 2)




Mat Sowa



Wade Harper

# An R Implementation of CompPASS



## CompPASS

Comparative Proteomic Analysis Software Suite

Experiments

Browse... SampleData\_Set1.tsv

Upload complete

External Stats

Browse... No file selected

Normalization Factor

0.98

### Input Instructions

Download Sample Input

The input file must be a tab-delimited text file no larger than 5Mb with the following five columns:

1. Experiment ID
2. Replicate
3. Bait
4. Prey
5. Spectral Count

Download

Show 25 entries

Search: EIF4E

Experiment.ID	Bait	Prey	AvePSM	Z	WD	Entropy
14613	EIF4E	EIF4G3	95.0	6.92955050	7.83625947	0.9992880
14613	EIF4E	ANGEL1	10.5	6.92964646	5.23832034	0.8756634
14613	EIF4E	EIF4ENIF1	30.0	6.92579623	4.28384182	0.9805264
14613	EIF4E	EIF4EBP2	21.0	6.89832377	3.39071344	0.9859094
14613	EIF4E	EIF4G1	134.0	6.92246807	2.18340357	0.9910094
14613	EIF4E	CBR3	1.0	6.92964646	1.61658075	1.0000000
14613	EIF4E	MAPRE3	1.0	6.92964646	1.61658075	1.0000000
14613	EIF4E	EIF4E	56.0	151.52528743	1.57844143	0.9888990
14613	EIF4E	PABPC4	3.5	0.69158986	1.35581151	0.9886994
14613	EIF4E	EIF4EBP1	30.5	6.92011991	1.33350970	0.9450667
14613	EIF4E	EIF3A	3.5	5.61111102	1.07350240	0.8960382
14613	EIF4E	EIF3F	1.5	6.26471601	0.43204938	0.9544340
14613	EIF4E	EIF3B	5.0	6.13449264	0.40000000	1.0000000
14612	SFN	EIF4EBP1	1.0	0.07426017	0.24146059	1.0000000



David Nusinow

<https://github.com/dnusinow/cRomppass>

<http://bioplex.hms.harvard.edu/comppass/>



# From IP's to Interactomes: Meeting the Challenges of Scale

BioDex 1.0    BioDex 2.0    BioDex 3.0    100

Parameter:	AP-MS Baits	LC-MS Runs	MS <sup>2</sup> Spectra	Peptide-Spectral Matches	Protein Identifications
Median Statistics Per Run:	1	2	14,090	3,923	805
<b>293T:</b>	<b>10,128</b>	<b>20,256</b>	<b>285,407,040</b>	<b>79,464,288</b>	<b>16,306,080</b>
<b>HCT116:</b>	<b>5,522</b>	<b>11,044</b>	<b>155,609,960</b>	<b>43,325,612</b>	<b>8,890,420</b>
<b>Total:</b>	<b>15,650</b>	<b>31,300</b>	<b>441,017,000</b>	<b>122,789,900</b>	<b>25,196,500</b>

## CHALLENGES OF SCALE

**1.**

How to do all those searches?

**2.**

How to ensure quality across thousands of runs?

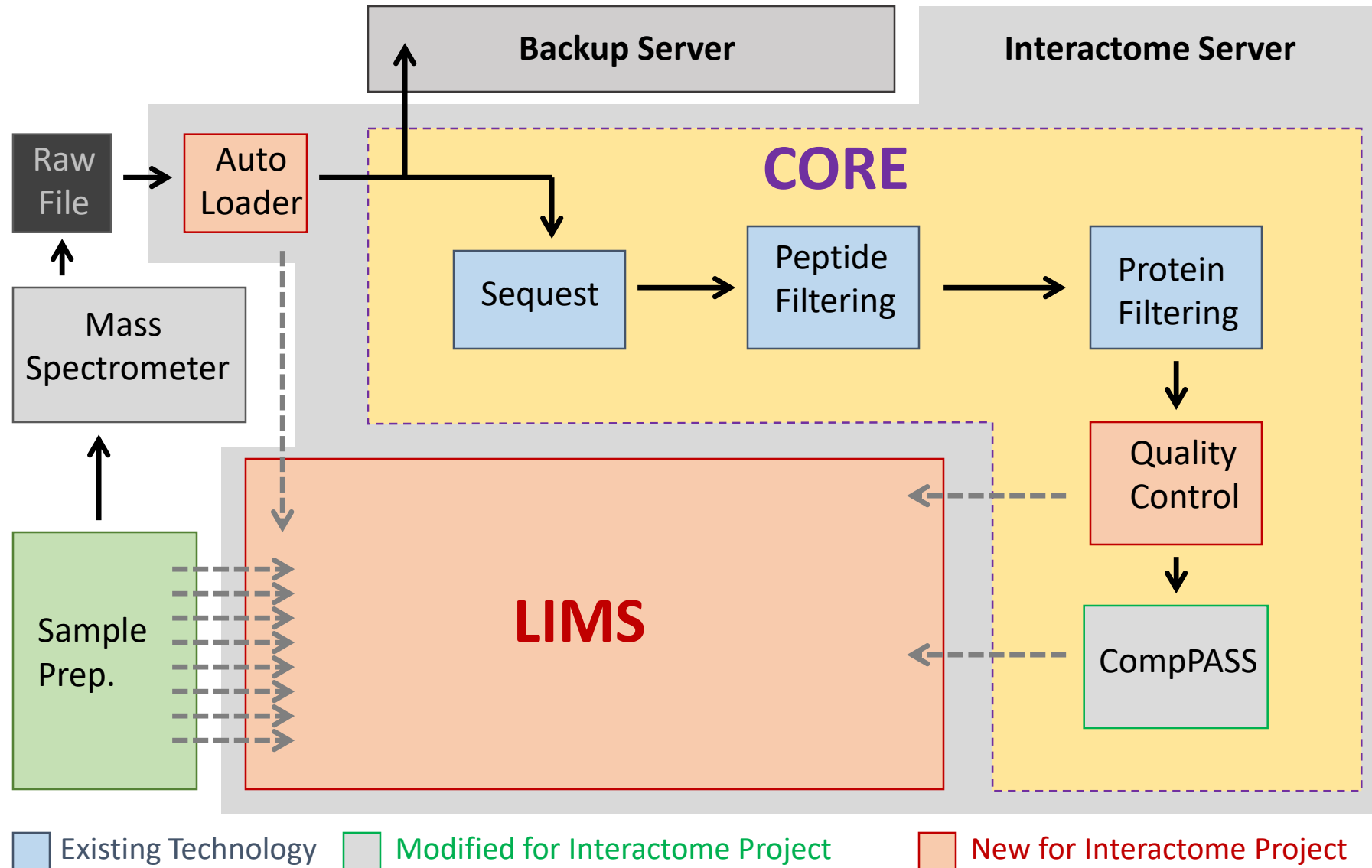
**3.**

How to minimize errors due to false positive ID's?

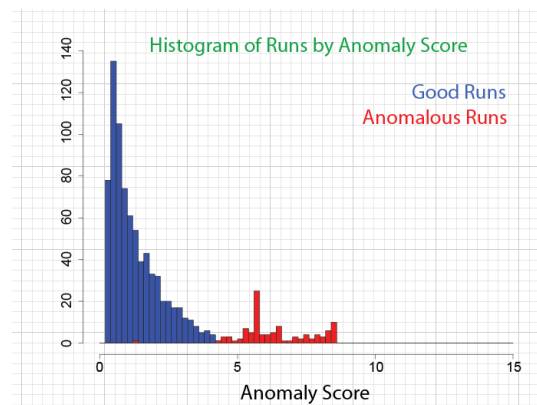
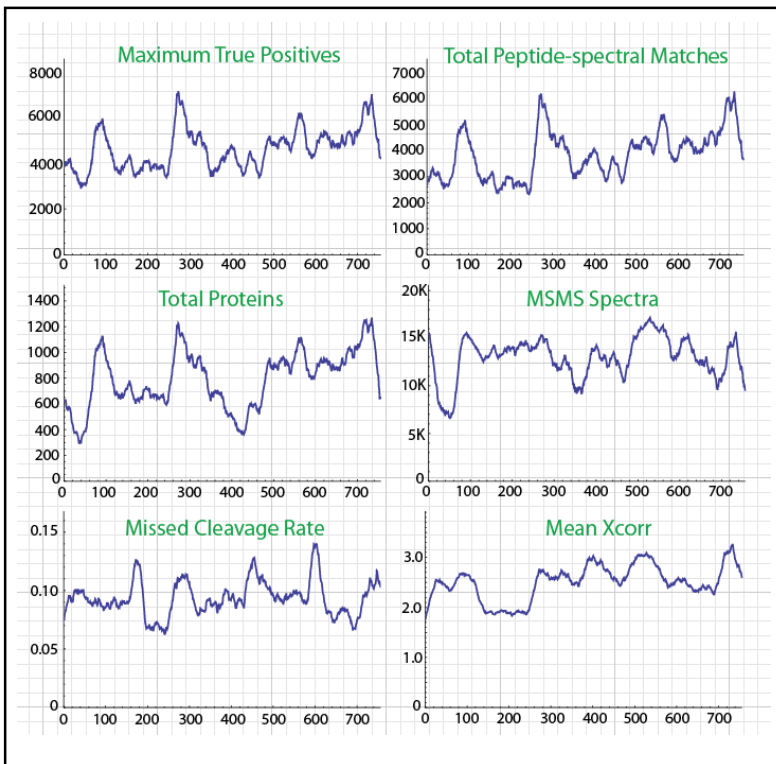
**4.**

How to rapidly and accurately identify interactors?

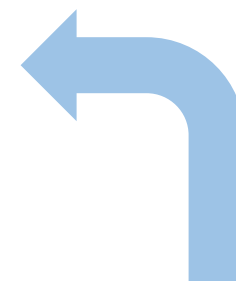
# Data processing overview



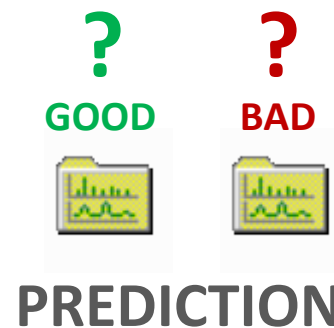
# Automated and Adaptive Run QC



Cycle repeats  
For every run.



PREDICTIVE MODEL



# Automatic Notification of Problematic Runs

## Problems with search 71455



SewerRat@harperfs.med.harvard.edu

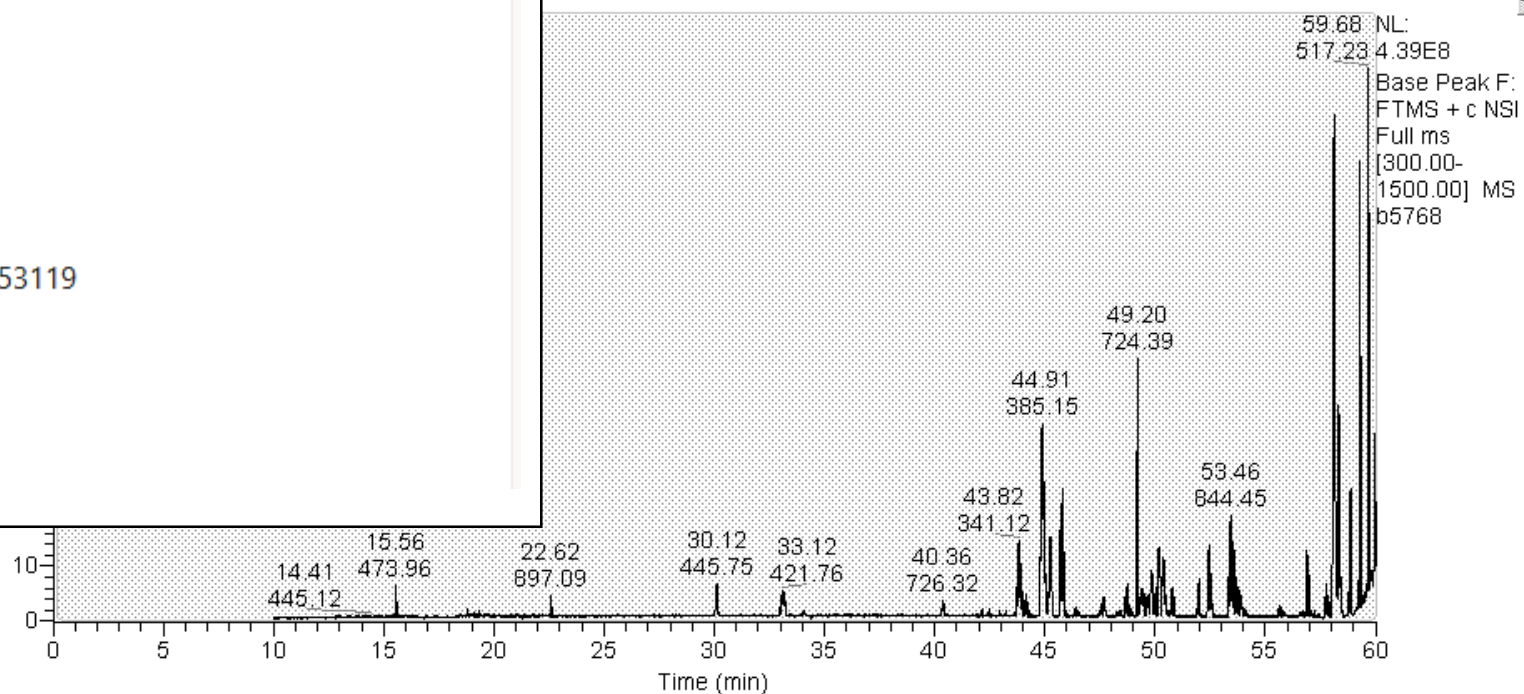
Mon 6/3/2019 1:44 AM

Huttlin, Edward Leo; Paulo, Joao A; joe\_cannon@hms.harvard.edu + 2 others

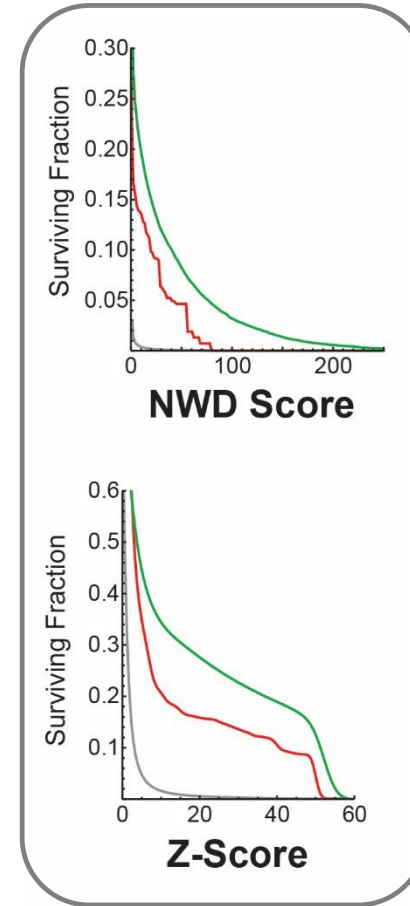
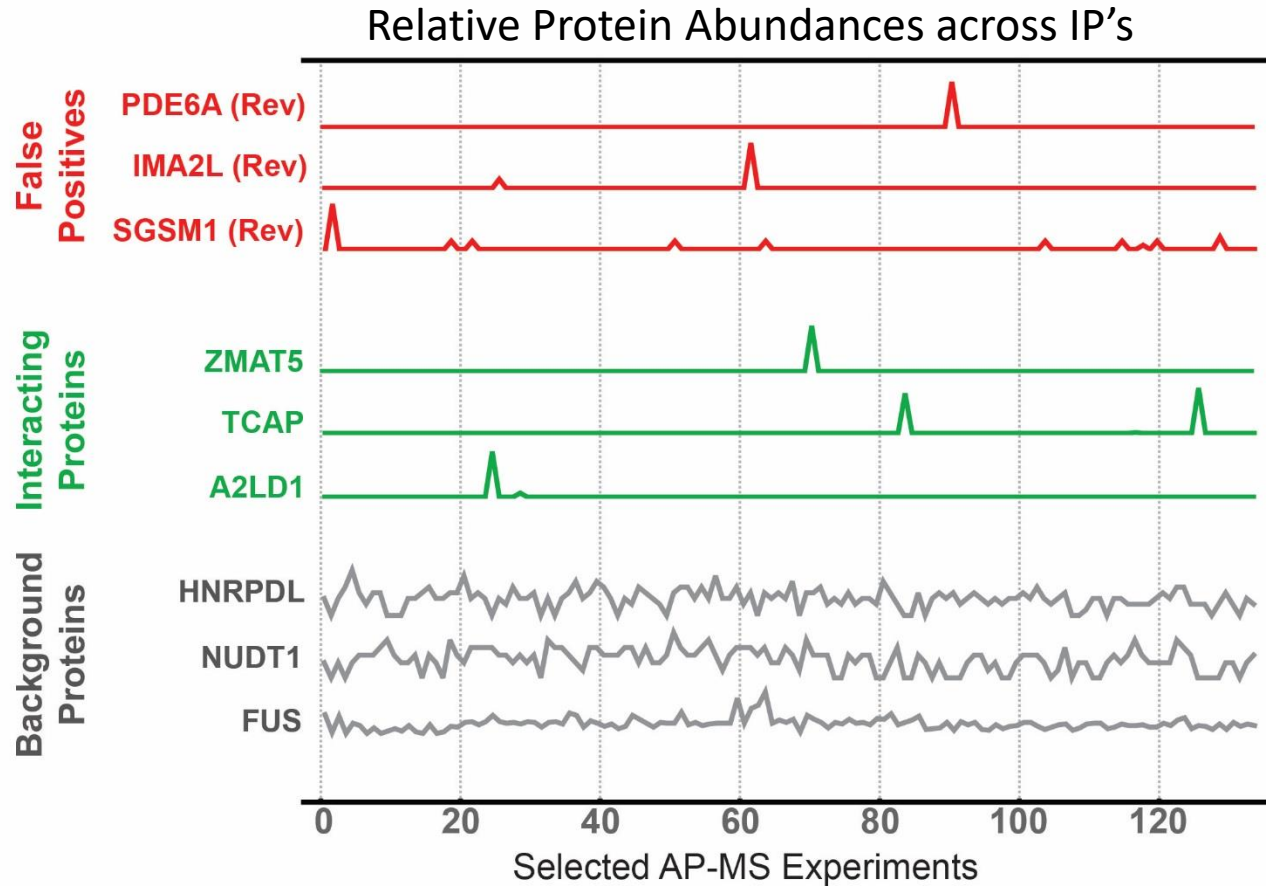


Problems with search 71455: qb01088\_HCT-REPEAT03\_A03\_A\_00671\_053119.

Parameter	Value
tpMax	553
totalPSMs	553
totalProteins	100
ratioUPepsToTPeps	0.47377852842942
sensitivity	0.99999819168501
msmsSpectra	14565
successRate	0.037967730861655
aveXcorr	2.5170851717902365
MCRate	0.094032379688283
search_name	qb01088_HCT-REPEAT03_A03_A_00671_053119
score	22.6033451322409
manualTraining	no
class	bad
notes	

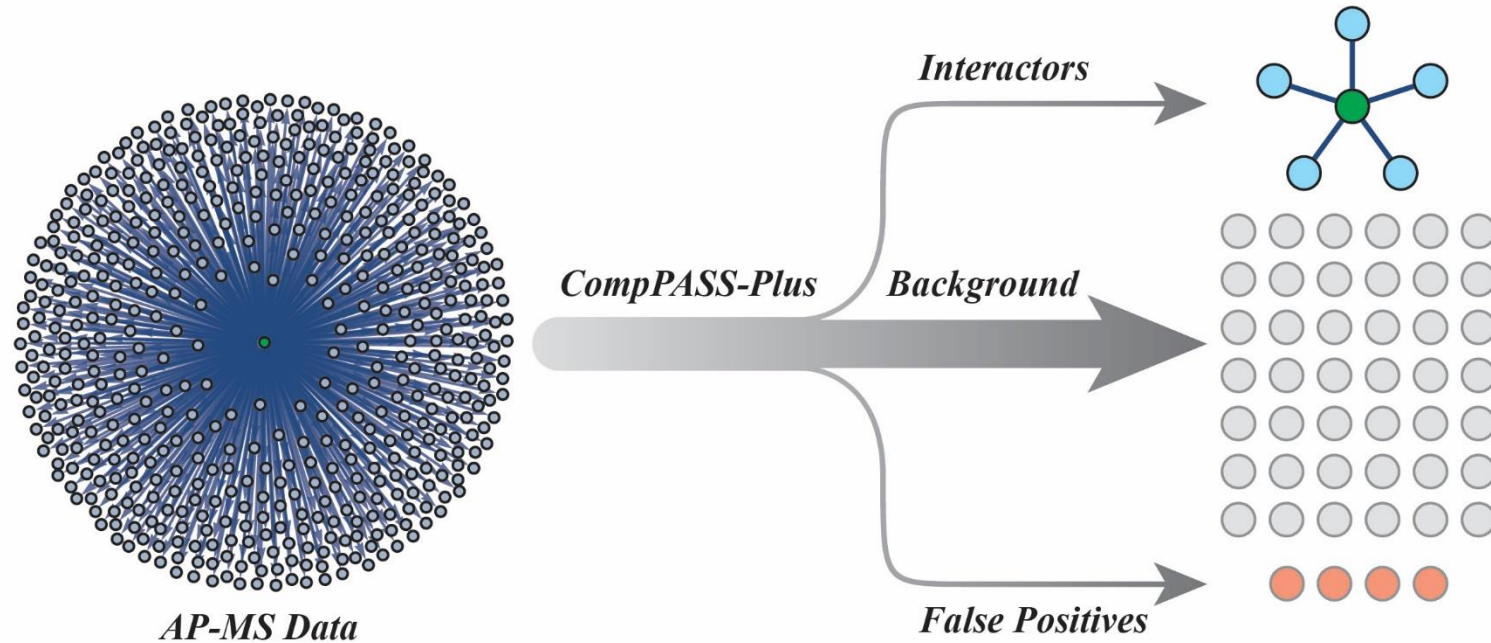


# Distinguishing Interactors from Background and False Positives



Both *bona fide* interactors and false positives appear in AP-MS experiments as “rare” events that score as likely interactions. Unless precautions are taken, enrichment of these false positives can cause surprisingly high false discovery rates among AP-MS datasets.

# Distinguishing Interactors from Background and False Positives



- A key challenge of AP-MS is distinguishing the few **true interacting proteins** (~1-2%) from a much larger number of **background proteins** (~98%), **false positives** (~1%), and other **experimental artifacts**.
- While existing algorithms for AP-MS distinguish enriched interacting proteins from nonspecific background, CompPASS-Plus uniquely accounts for false positive ID's.



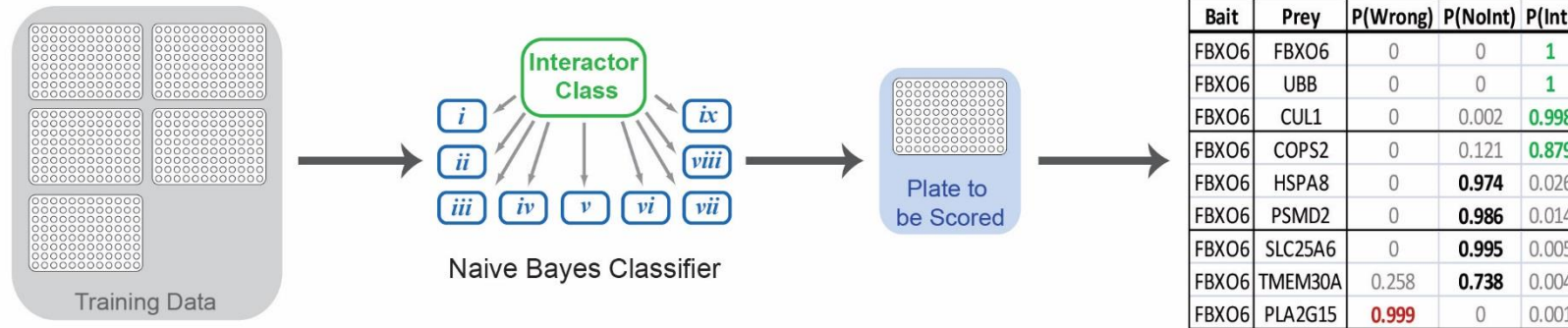
# CompPASS-Plus for Interactor Identification

## Initial Training Data

Class	Description	Percent
Likely Wrong ID	Modeled by distribution of decoy prey ID's.	0.05
Non-Interaction	Bait-prey pairs not reported in STRING/GeneMania	98.6
Likely Interaction	Bait-prey pairs confirmed by STRING/GeneMania.	1.35

## Features

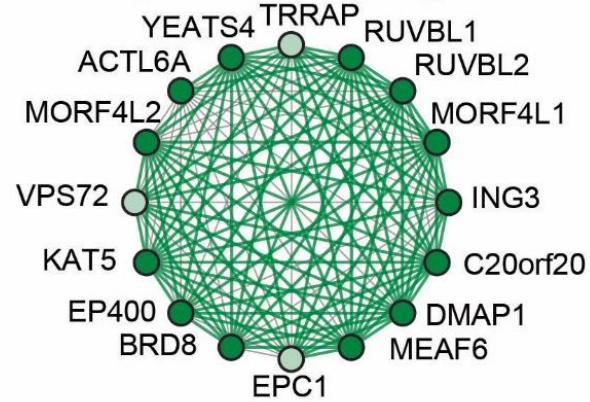
i. NWD Score	vi. Ratio
ii. Z-Score	vii. Total PSM's
iii. Plate Z-Score	viii. Ratio Total PSM's
iv. Entropy	ix. Unique:Total Peptide Ratio
v. Unique Peptides (Binned)	



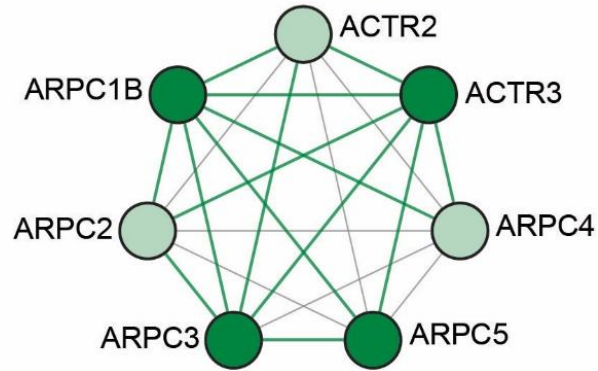
- CompPASS-Plus is a Naive Bayes classifier that extends CompPASS for improved Identification of interacting proteins.
- Features include standard CompPASS scores as well as customized scores.
- Training data is obtained from STRING/GeneMania and from Target/Decoy methods
- Leave-one-out cross-validation is incorporated at the 96-well-plate level for classification.

# CORUM Validation

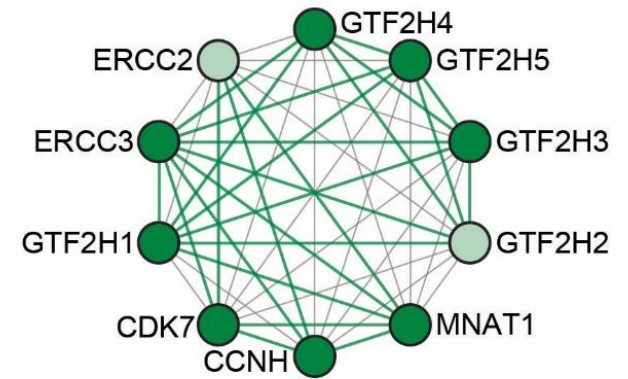
**NuA4/Tip60-HAT Cmplx. ( $p = 5.7e^{-240}$ )**



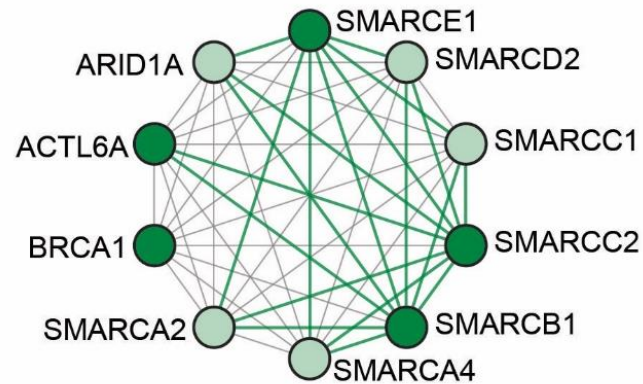
**Arp2/3 Complex ( $p = 1.1e^{-35}$ )**



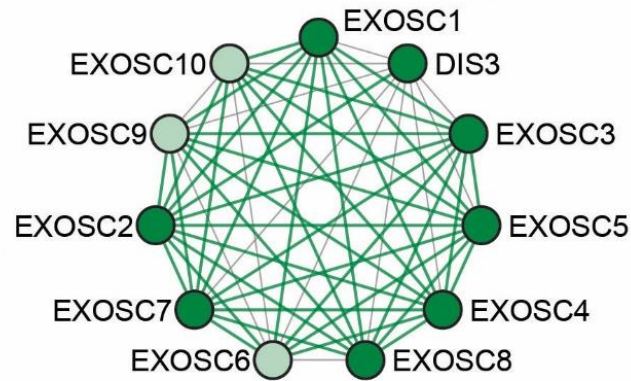
**TFIIH Cmplx. ( $p = 3.8e^{-60}$ )**



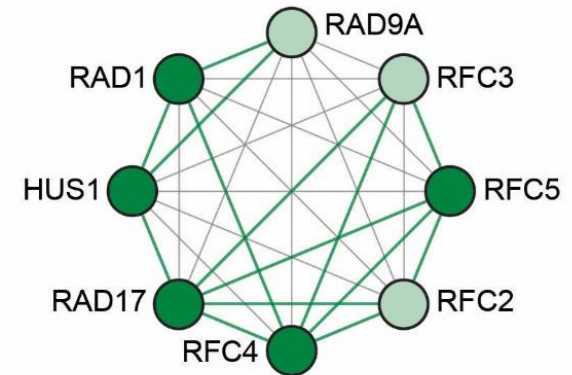
**SWI-SNF Complex ( $p = 1.1e^{-45}$ )**



**Exosome ( $p = 9.5e^{-110}$ )**

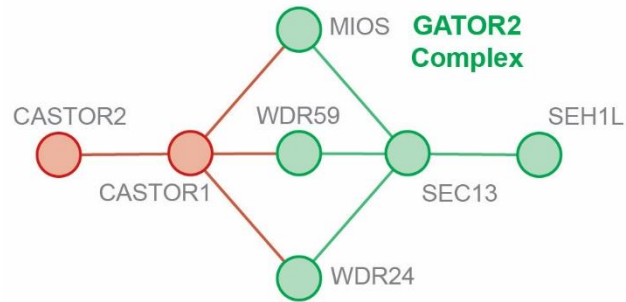


**Checkpoint Rad Complex ( $p = 3.5e^{-33}$ )**

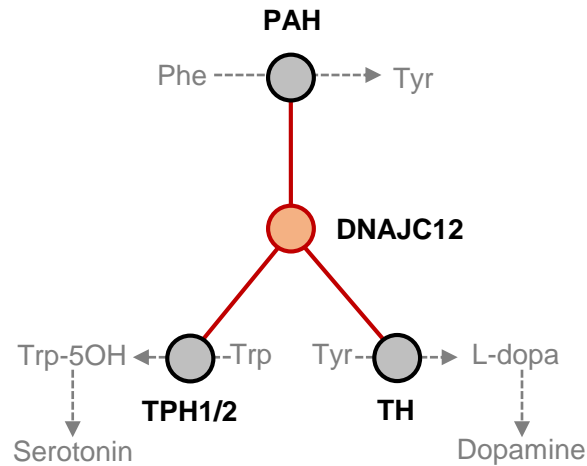




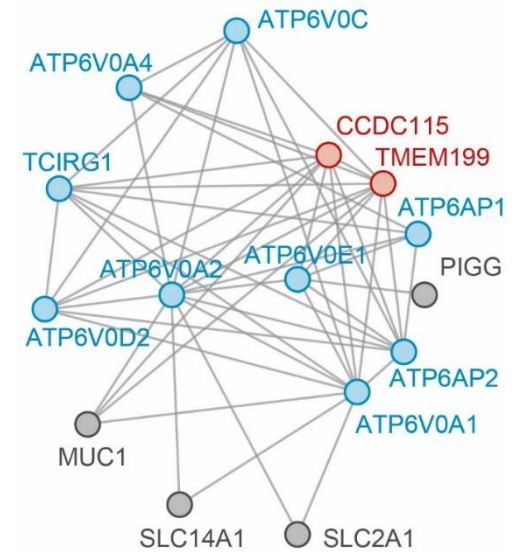
# BioPlex Associates New Proteins with Known Complexes



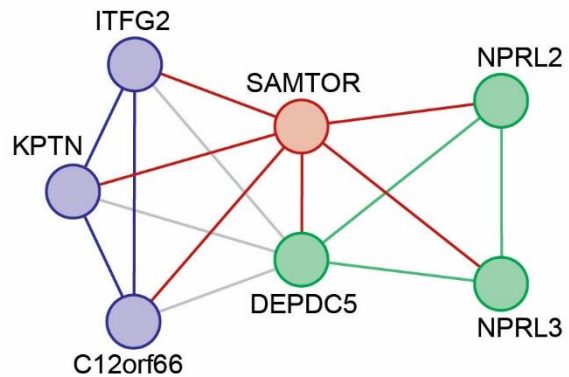
Chantranupong (2016) *Cell*, 165:153.  
Saxton (2016) *Nature*, 536:229.



Anikster et al. (2017)  
*Am. J. Hum. Genet.* 100, 257.

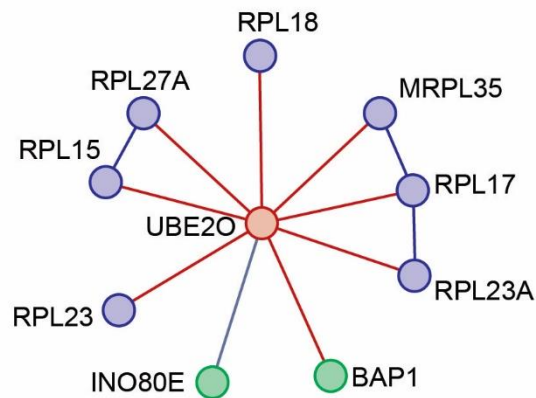


Miles (2017) *eLife*, 6:e22693.



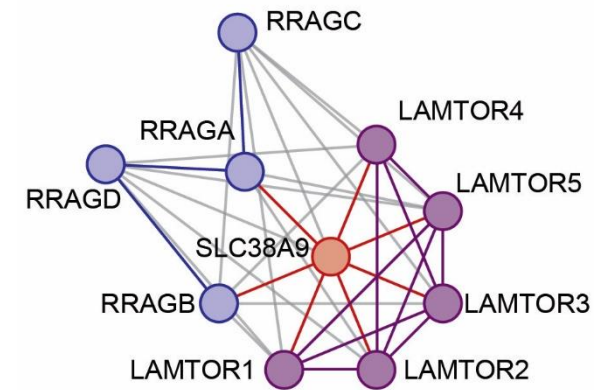
KICSTOR Complex GATOR1 Complex

Gu et al. (2017) *Science*, 358:813.



Ribosome Components  
Known Substrates

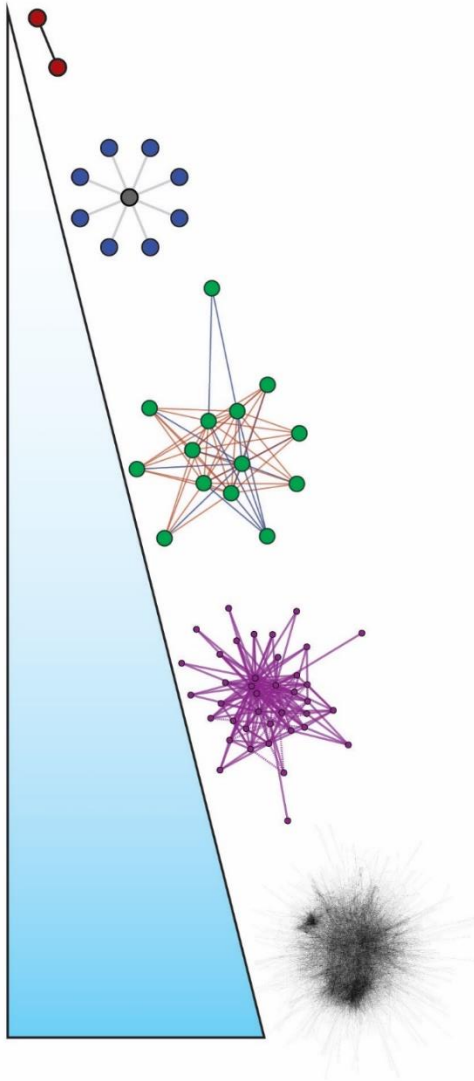
Nguyen (2017) *Science*, 357:eaan0218.



Rag GTPases Ragulator Complex

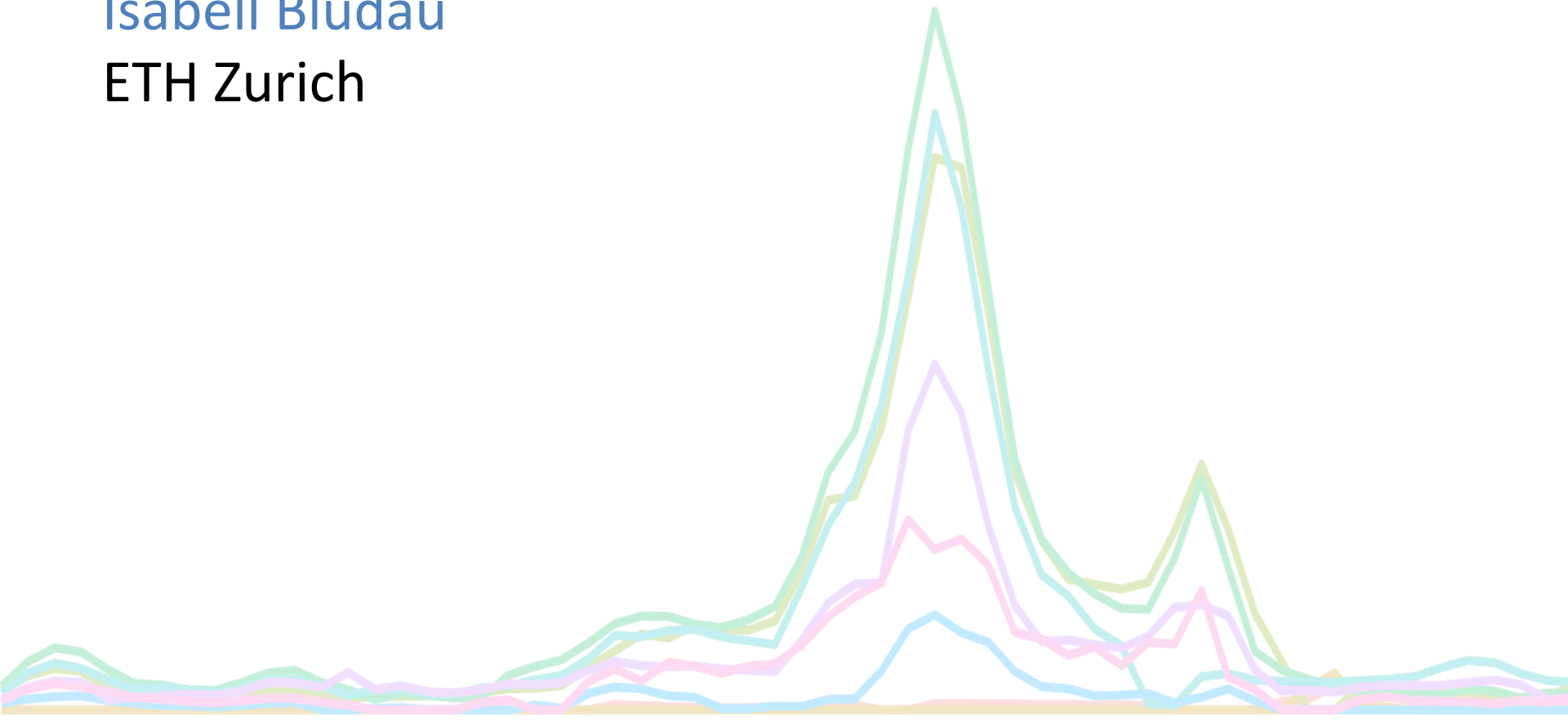
Rebsamen (2015) *Nature*, 519:477.

# Summary



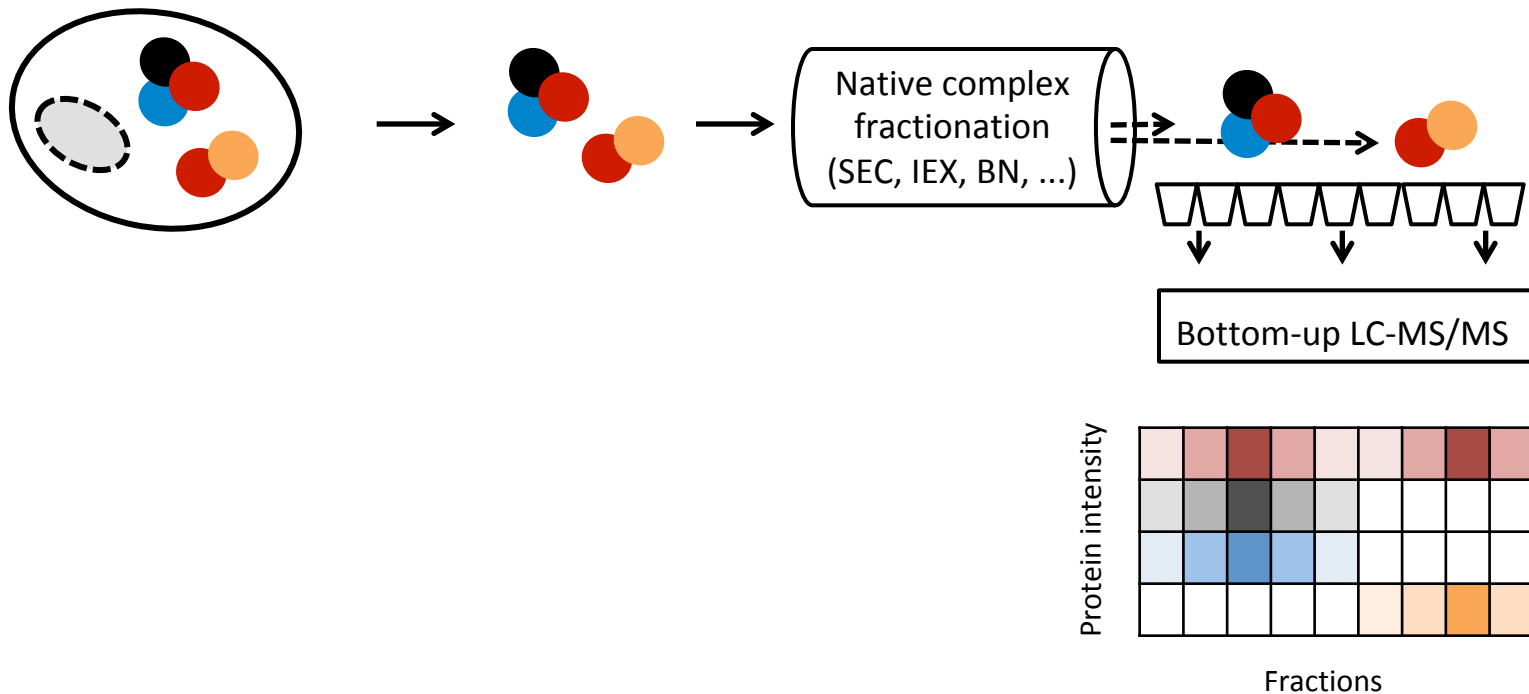
- Separating true interacting protein partners from background is a serious challenge for AP-MS
- A variety of algorithms have been developed to address this that accommodate a variety of experimental designs
- Performing AP-MS at truly large scale leads to additional challenges that must be addressed.
- Our CompPASS-Plus algorithm has enabled us to efficiently identify interacting proteins and produce reliable maps of the human interactome in multiple cell types.

Isabell Bludau  
ETH Zurich



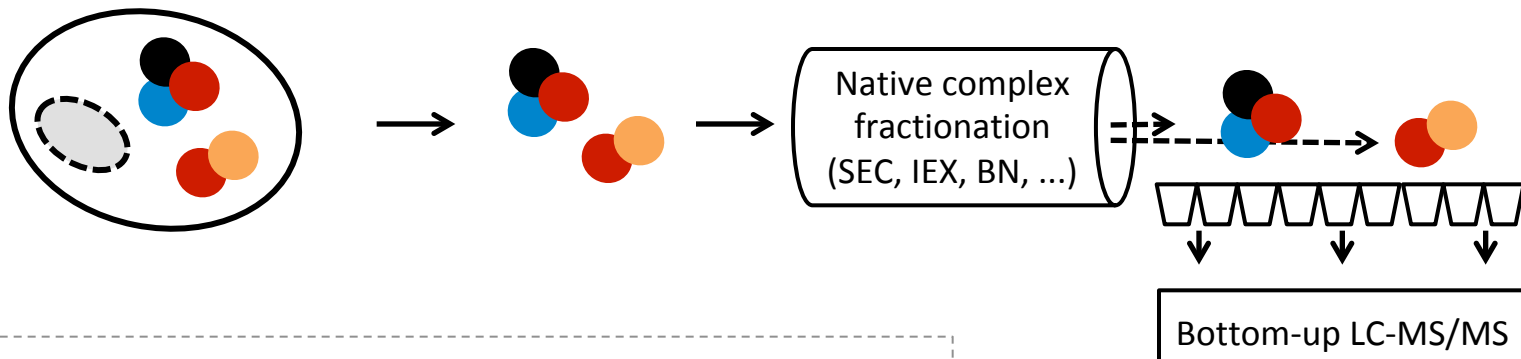
Underlying technology for data acquisition:

## Protein co-fractionation mass spectrometry



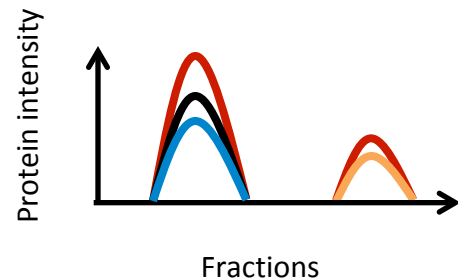
Underlying technology for data acquisition:

## Protein co-fractionation mass spectrometry

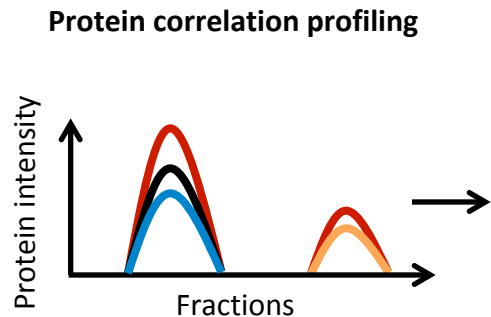


### Benefits:

- + Independent of genetic engineering or antibody availability
- + Parallel detection of protein complexes on a proteome-wide scale

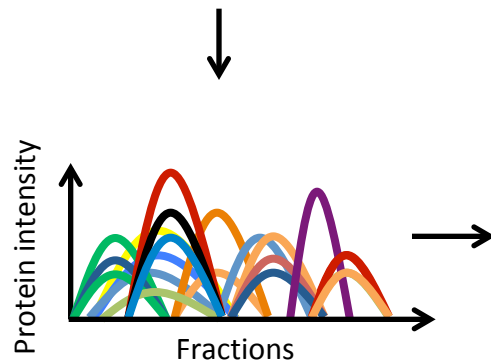
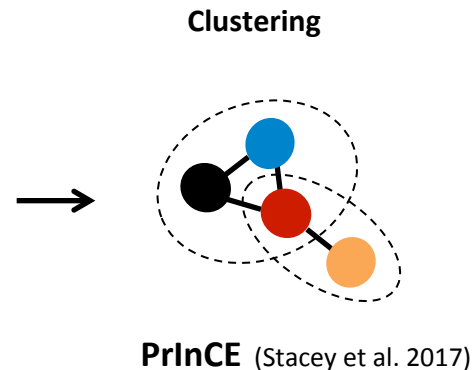


# Established data analysis strategy: Protein correlation profiling



**Pairwise scoring**

	●	●	●	●
●	1	0.9	0.8	0.9
●	0.9	1	0.9	0.1
●	0.8	0.9	1	0.2
●	0.9	0.1	0.2	1



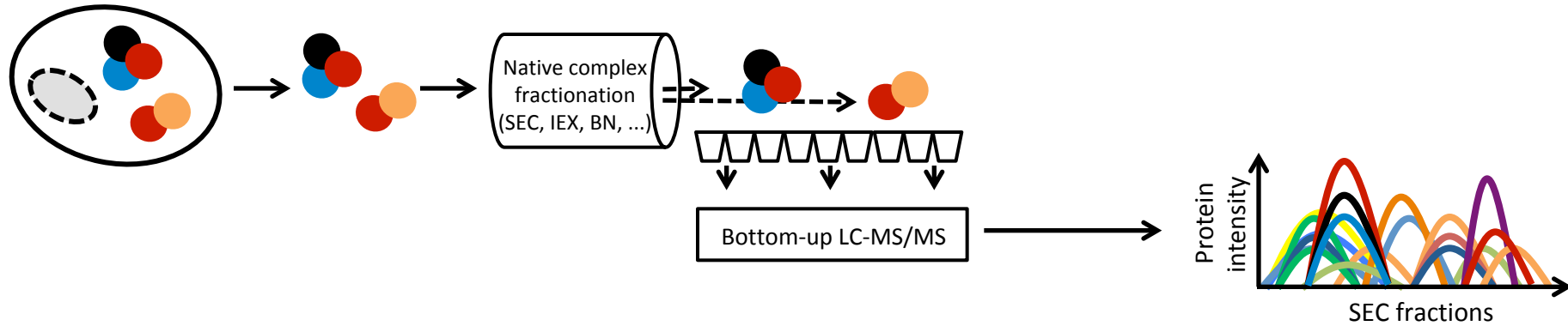
## Data properties:

- Profiles for thousands of proteins (~5000)
- Limited peak capacity (~20-30)

## Consequences:

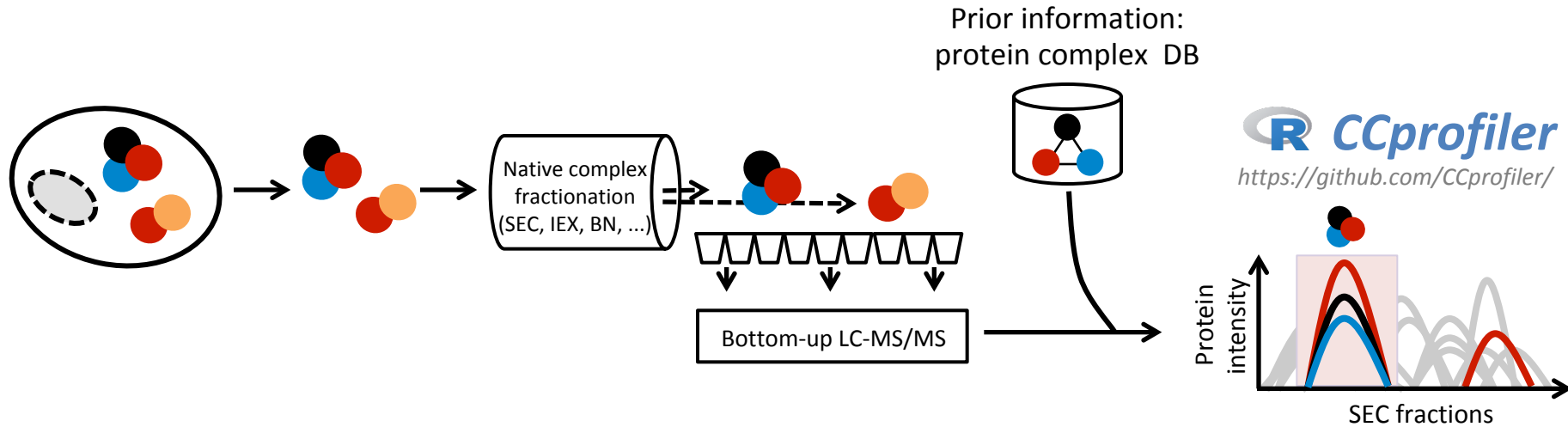
- Random co-elution of proteins
- Limited selectivity and sensitivity

Targeted analysis strategy:  
Complex-centric proteome profiling



Targeted analysis strategy:

## Complex-centric proteome profiling



- ✓ Automated software for targeted complex-centric analysis
- ✓ Parallel and sensitive protein complex detection
- ✓ Complex-level FDR estimation



# Complex-centric proteome profiling: Decoy based FDR estimation

## Targets

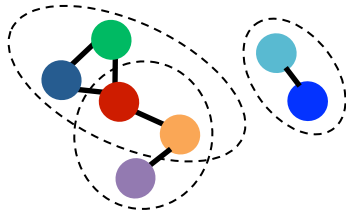
- a) Defined hypotheses (e.g. CORUM)



- b) Interaction network

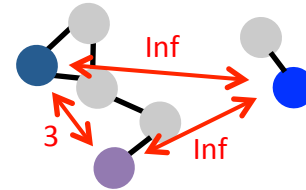
(e.g. AP-MS, BioPlex, StringDB)

Select 1<sup>st</sup> degree neighbors of each protein as one target hypothesis

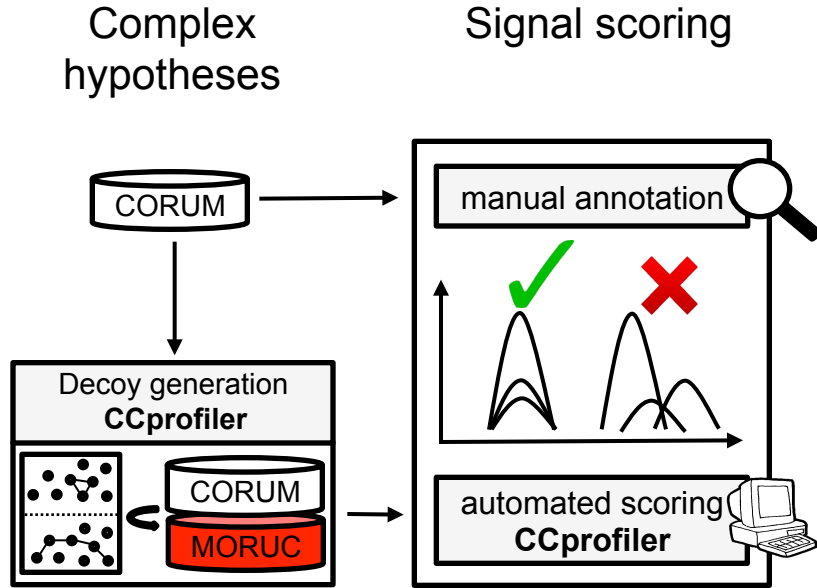


## Decoys

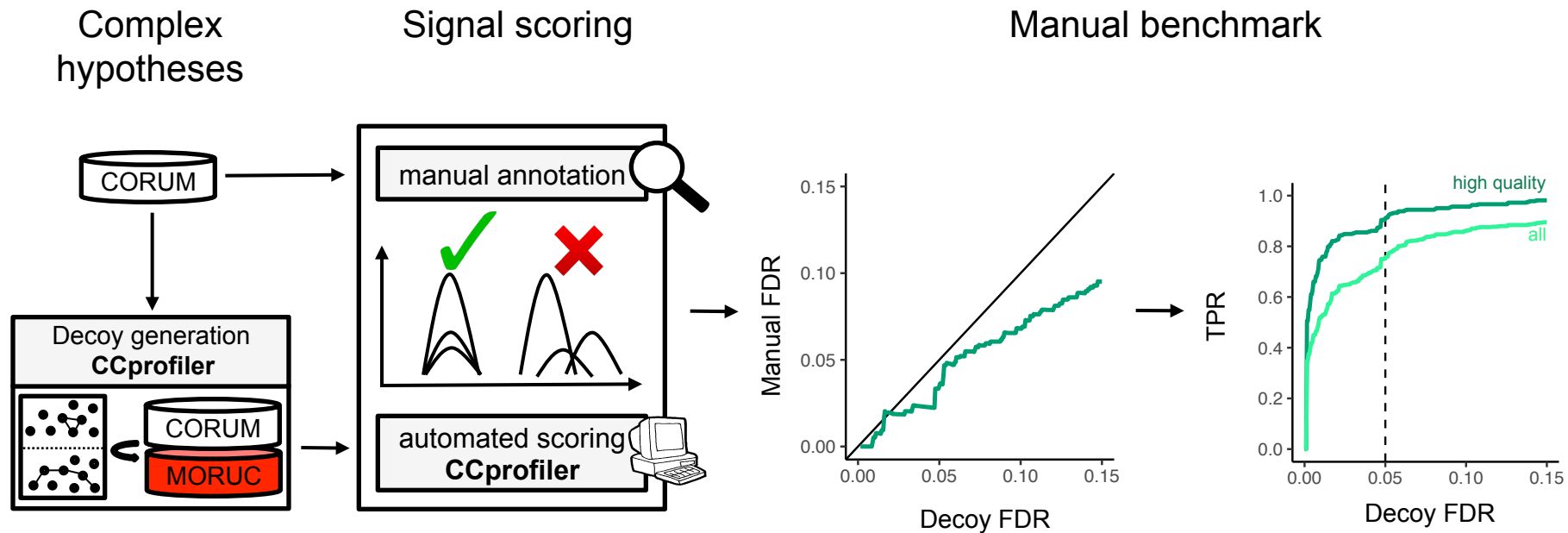
- Generate matching decoy for each target (same complex size distribution)
- Require minimum network distance between decoy proteins to avoid overlap with targets



# Complex-centric proteome profiling: Benchmark

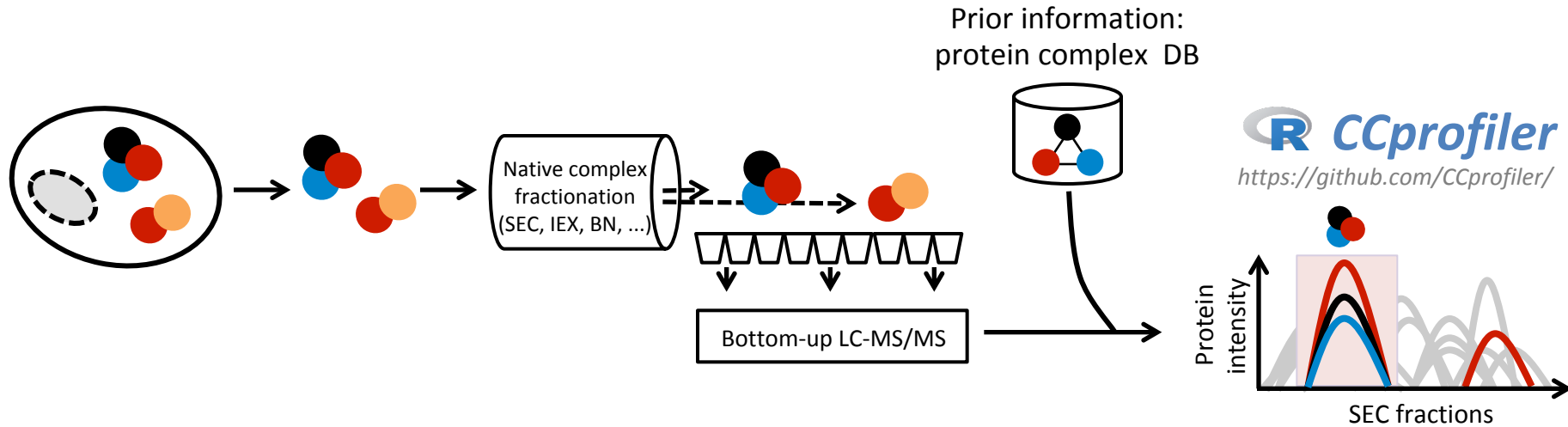


# Complex-centric proteome profiling: Benchmark



Targeted analysis strategy:

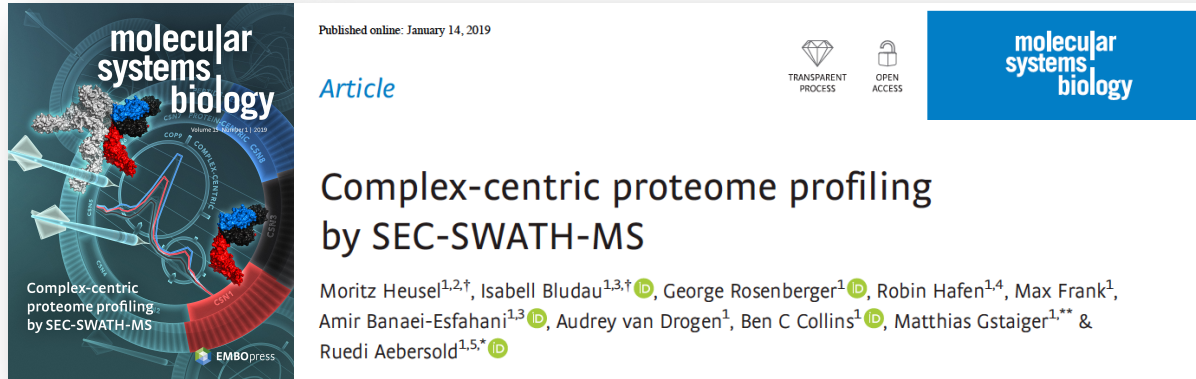
## Complex-centric proteome profiling



- ✓ Automated software for targeted complex-centric analysis
- ✓ Parallel and sensitive protein complex detection
- ✓ Complex-level FDR estimation

Targeted analysis strategy:

## Complex-centric proteome profiling



<sup>†</sup> These authors contributed equally to this work.

- ✓ Automated software for targeted complex-centric analysis
- ✓ Parallel and sensitive protein complex detection
- ✓ Complex-level FDR estimation

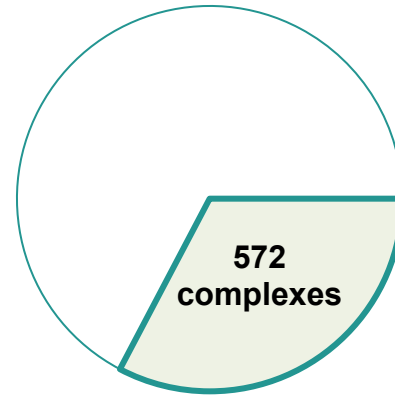
# Complex-centric proteome profiling by SEC-SWATH-MS

- ✓ **Detect and quantify hundreds of protein complexes**

## HEK293 soluble proteome

- SEC fractionation: 1 mg, Yarra-SEC-4000, 81 fractions
- SWATH-MS: Triple-TOF 5600, 64 vw, 120 min gradient
- Consistent quantification of 4916 proteins

Detection of 572 out of 1753  
CORUM complexes (5% FDR)

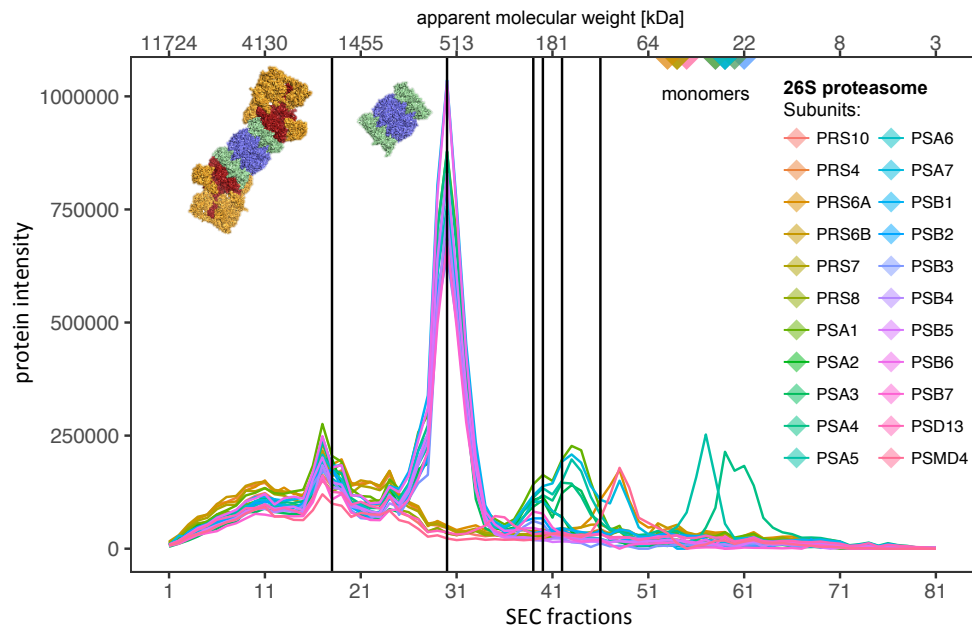


# Complex-centric proteome profiling by SEC-SWATH-MS

- ✓ Detect and quantify hundreds of protein complexes
- ✓ Investigate sub-complexes and assembly intermediates

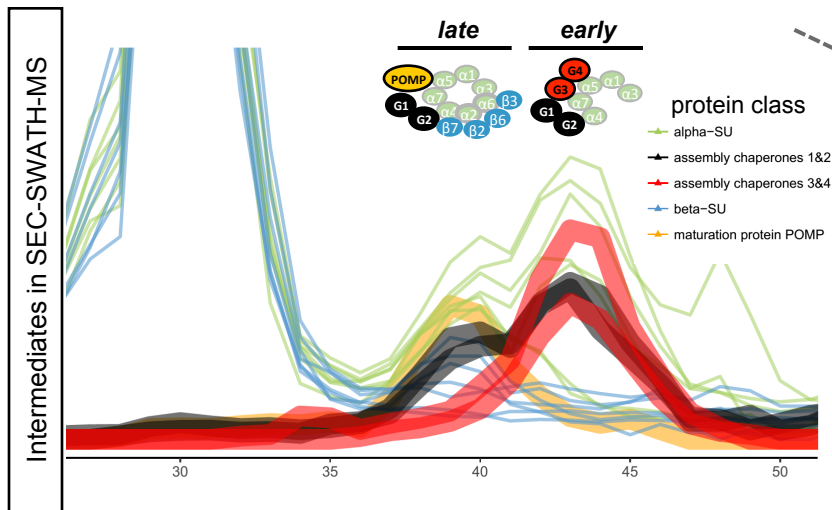
HEK293 soluble proteome

## Proteasome assembly



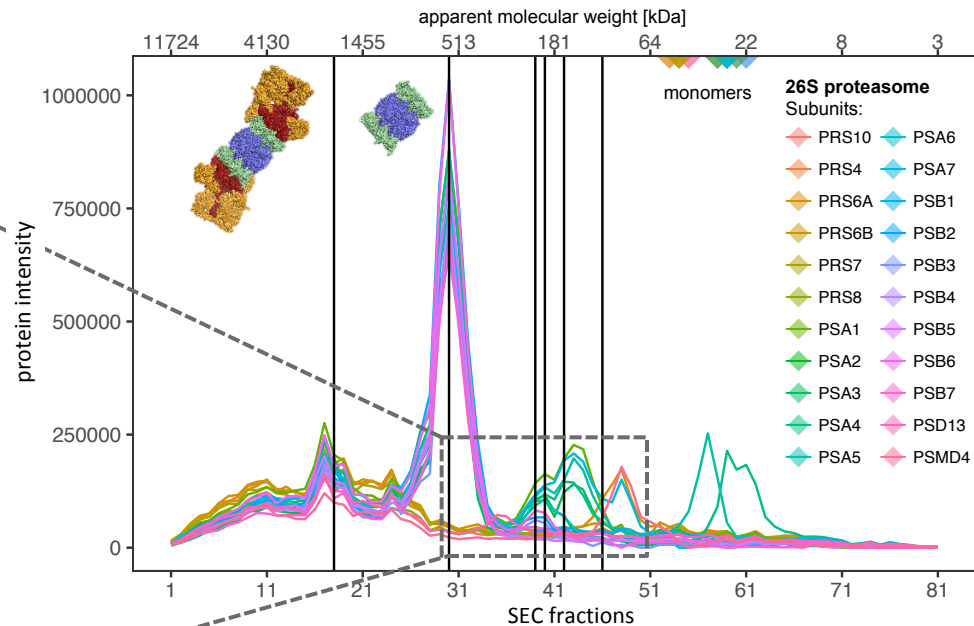
# Complex-centric proteome profiling by SEC-SWATH-MS

- ✓ Detect and quantify hundreds of protein complexes
- ✓ Investigate sub-complexes and assembly intermediates



HEK293 soluble proteome

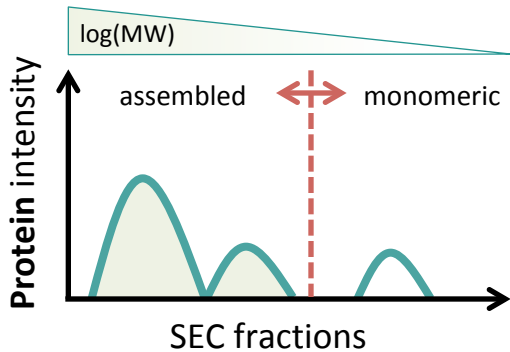
## Proteasome assembly



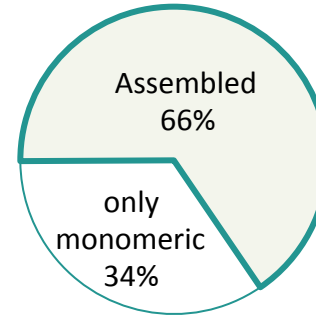


# Complex-centric proteome profiling by SEC-SWATH-MS

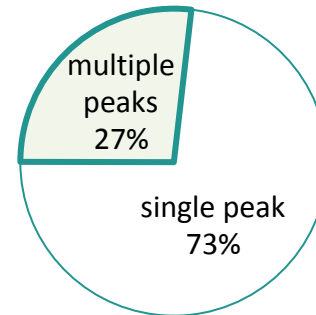
- ✓ Detect and quantify hundreds of protein complexes
- ✓ Investigate sub-complexes and assembly intermediates
- ✓ **Evaluate global proteome assembly characteristics**



## HEK293 soluble proteome



- The majority of the proteins appear in at least one assembled state

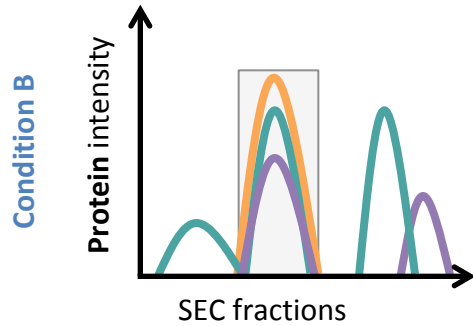
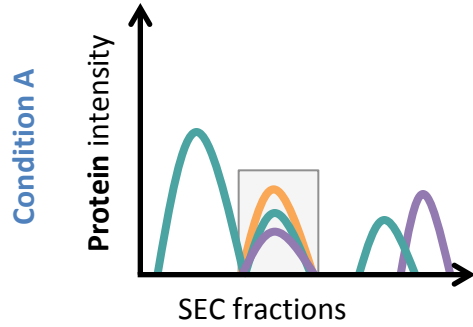


5503 elution peaks for 4065 proteins

- Many proteins are observed in multiple distinct assembly states

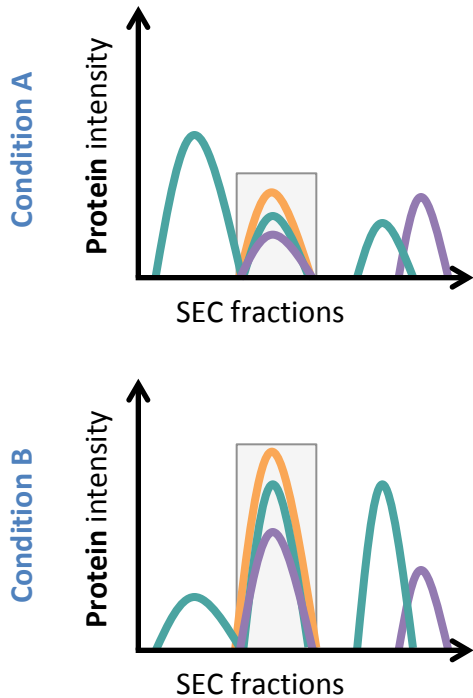
Current developments:

## Quantitative comparison of protein complexes across conditions



Current developments:

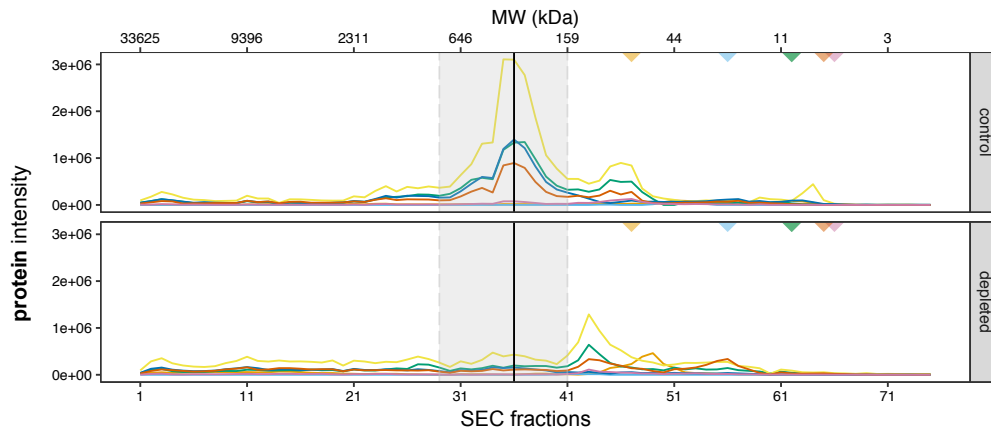
# Quantitative comparison of protein complexes across conditions



## Splicing efficient and PRPF8 depleted human cells

- 110 out of 553 complexes are differentially abundant (FDR < 0.05)
- Spliceosome biogenesis is down-regulated

### SMN complex



Current developments:

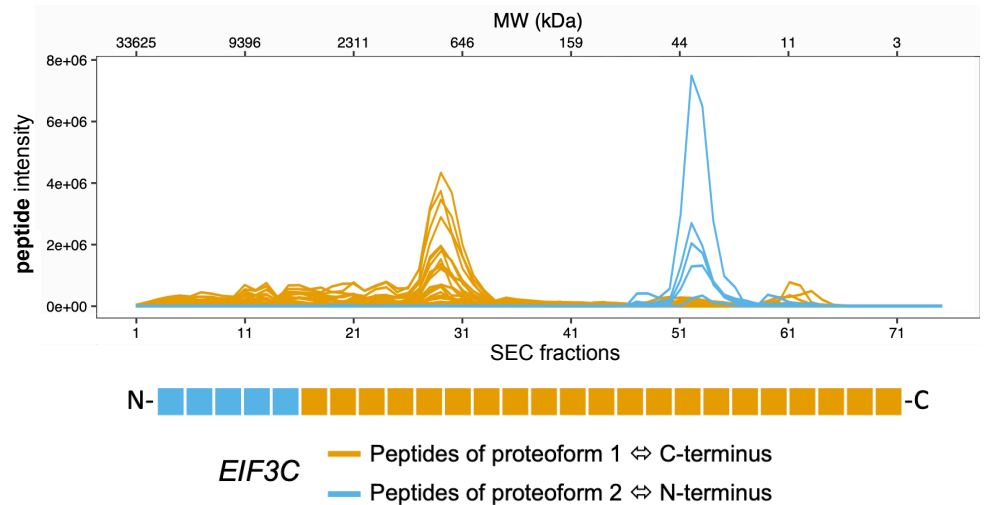
## Data-driven detection of assembly specific proteoforms

‘**Proteoforms**’ = protein variants that originate from the **same gene**,  
but that have a **unique amino acid sequence and post-translational modifications**

- Make use of peptide-level information available in SEC-SWATH-MS
- Distinguish assembly specific proteoforms based on unique peptides
- ✓ Parallel detection of 1,378 assembly specific proteoforms

Poster *ThP626*

*Eukaryotic translation initiation factor 3 subunit C (EIF3C)*



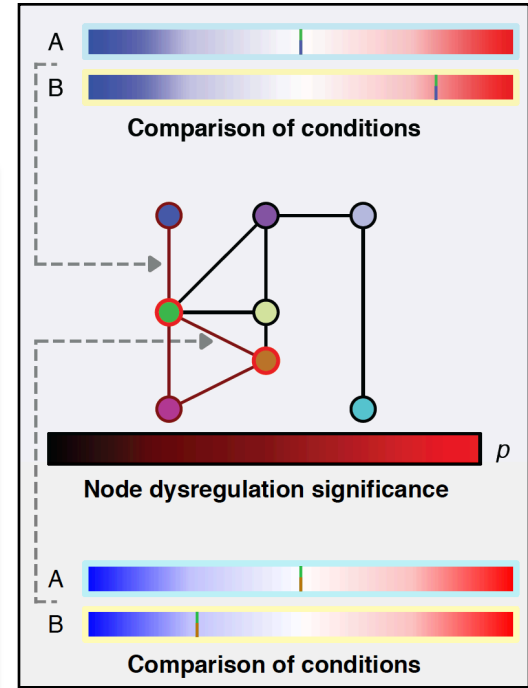
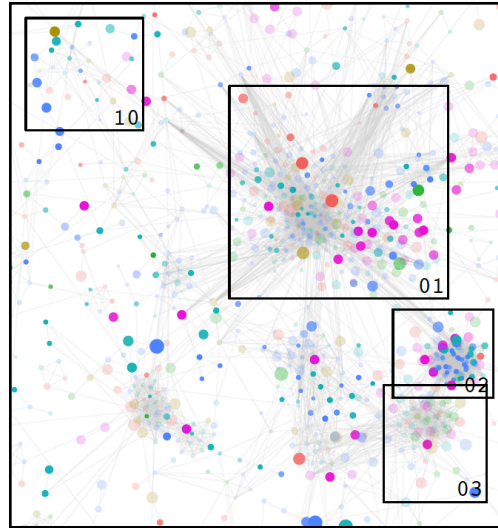
Current developments:

## Network-centric analysis

### SEC analysis toolkit (SECAT)

- Determine perturbed nodes in the PPI interaction network

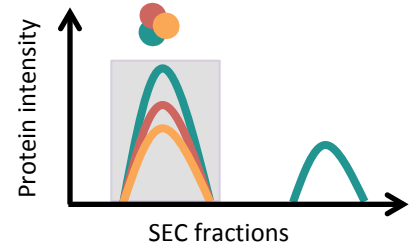
Rosenberger et al. (*in preparation*)



Take home message:

## Complex-centric analysis of CoFrac-MS data

- ✓ Consistent detection and quantification of 100s protein complexes on a proteome-wide scale
- ✓ Controlled complex-level FDR
- ✓ Sub-complex resolution



### Requirements:

- Quantitative peptide or protein matrix (DIA or DDA)
- Annotation table that matches MS runs to the sampled fractions
- Prior protein connectivity information (e.g. CORUM, BioPlex, StringDB)



- ✓ Coming next: quantitative complex comparison, proteoform detection, network-centric analysis

# Thank you for your attention!

## Aebersold lab

- Ruedi Aebersold
- Ben Collins
- **Moritz Heusel**
- **Max Frank**
- George Rosenberger
- Claudia Martinelli
- Peng Xue
- Robin Hafen
- Amir Banaei-Esfahani
- Audrey van Drogen
- Yansheng Liu
- Matthias Gstaiger



**ETH** zürich

life science zurich  
graduate school

## External collaborators

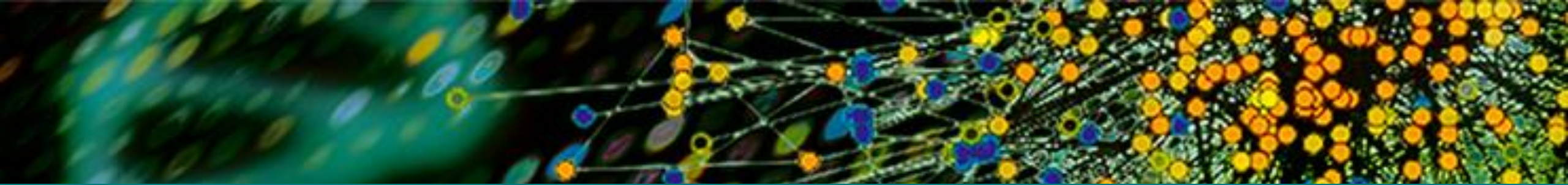
- Hannes Röst
- Yuija Cai
- Vihandha Wickramasinghe
- Ashok Venkitaraman



Dominic Helm

EMBL Heidelberg

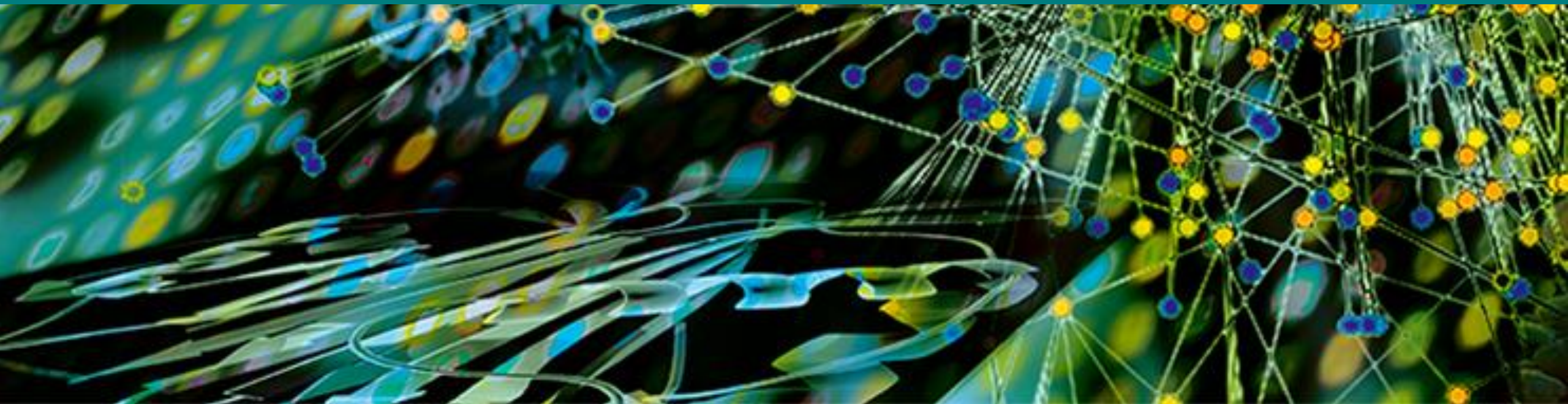




# Thermal proteome profiling for interactomics

ASMS 2019 – MS-Based interactomics

dominic helm





# Acknowledgement

Martin Beck

Peer Bork

Nassos Typas

Wolfgang Huber

Eileen Furlong

Christoph Mueller

Natalie Romanov

Malte Paulsen

Marcus Bantscheff

Gerard Drewes

Thilo Werner

Friedrich Reinhard

Holger Franken

Celia Berkers (Utrecht University)

Proteomics Core Facility

Flow Cytometry Core Facility

Advanced Light Microscopy Core Facility

Mechanical Workshops

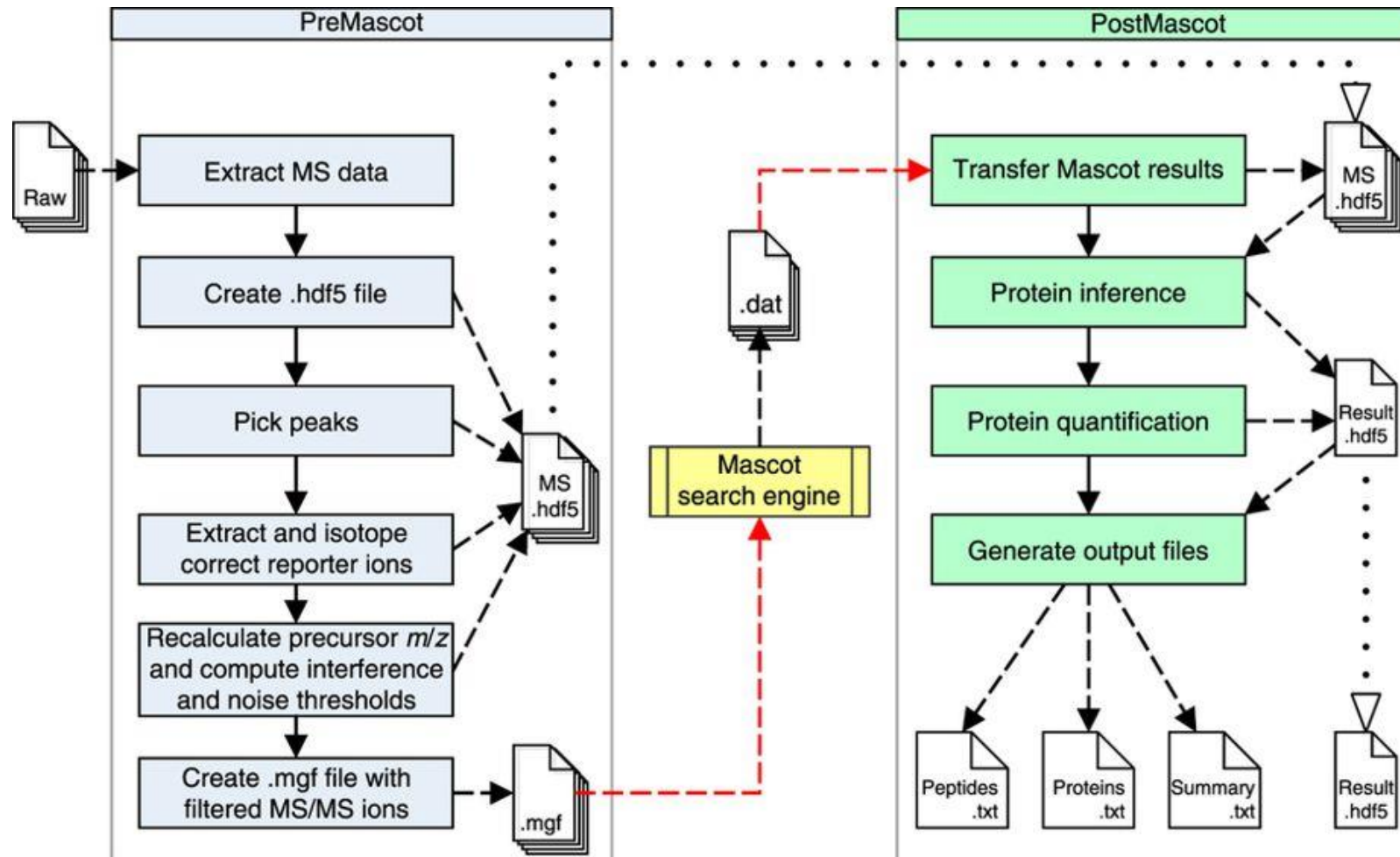


**Sindhuja Sridharan, Nils Kurzawa, Isabelle Becher, Frank Stein, Dominic Helm, Tomi Määttä, Andre Mateus, Per Haberkant, Mandy Rettel, Jianguo Zhu, Henrik Hammaren, Clement Potel, Malay Sha, Vallo Varik, Kerri Malone, Cecilia Perez-Borrajero, Matteo Perino**

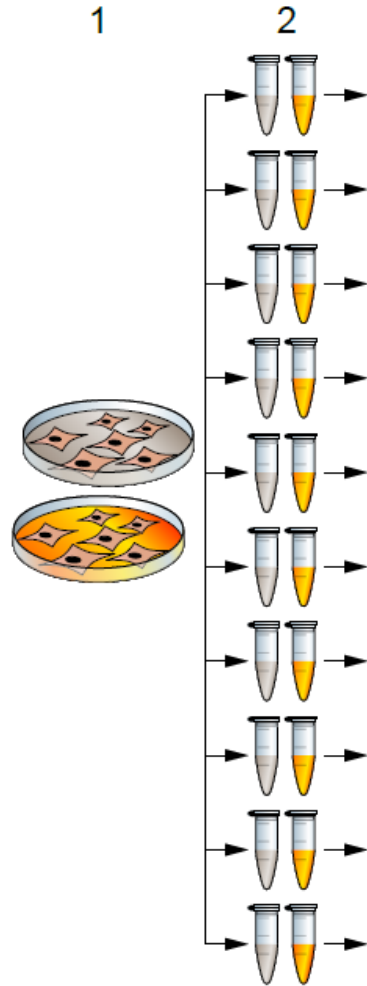
**Former members: Johannes Hevler, Maike Schramm EMBL**



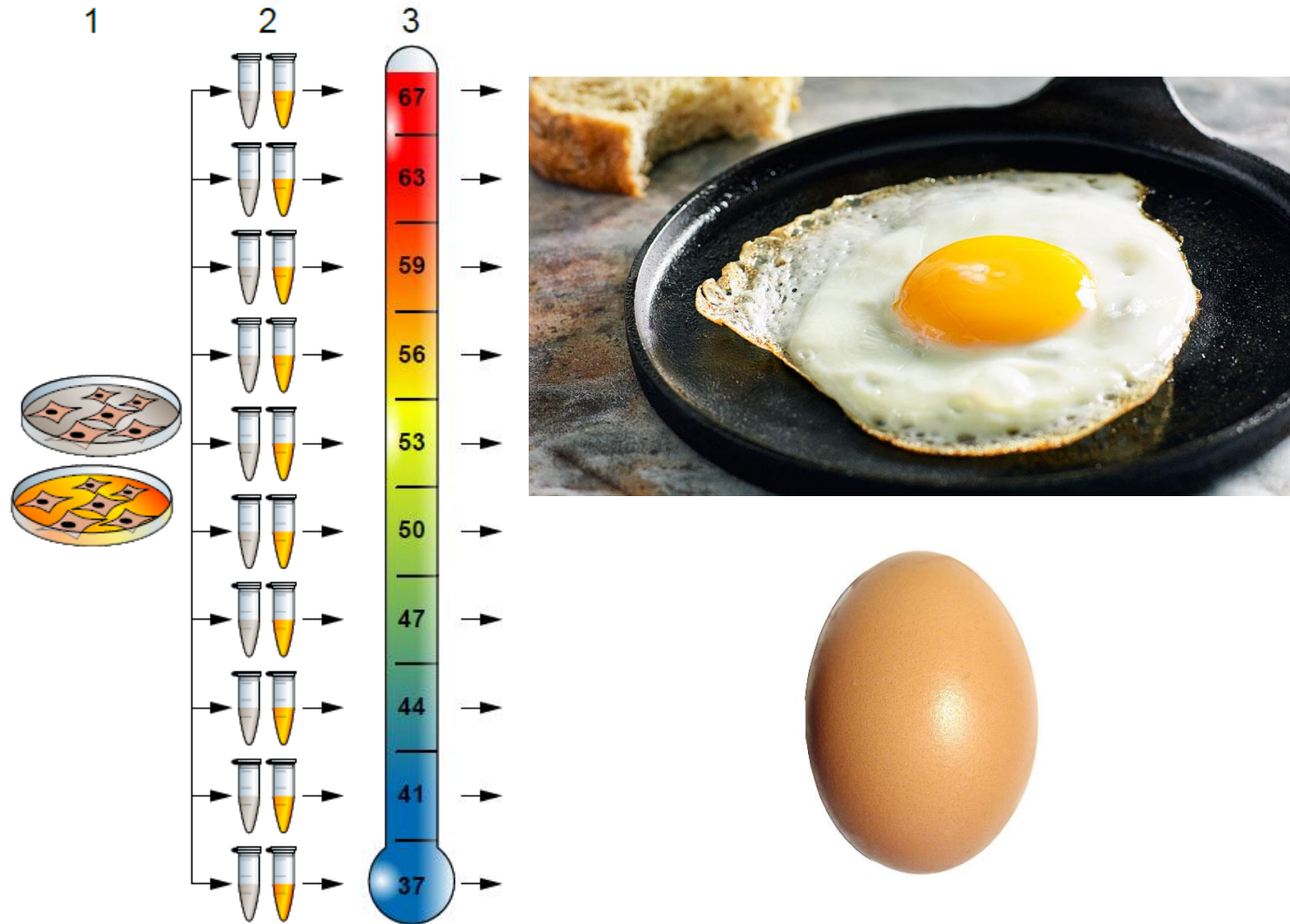
# Prelude: Isobarquant



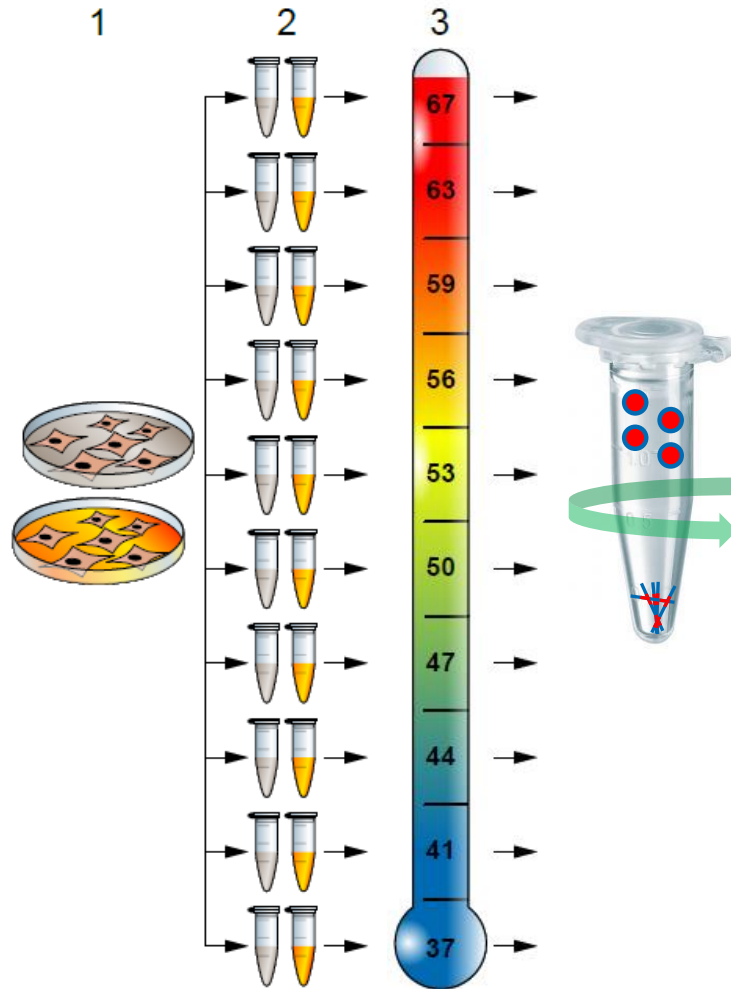
# Protein/drug interactions in living cells proteome-wide



# Protein/drug interactions in living cells proteome-wide

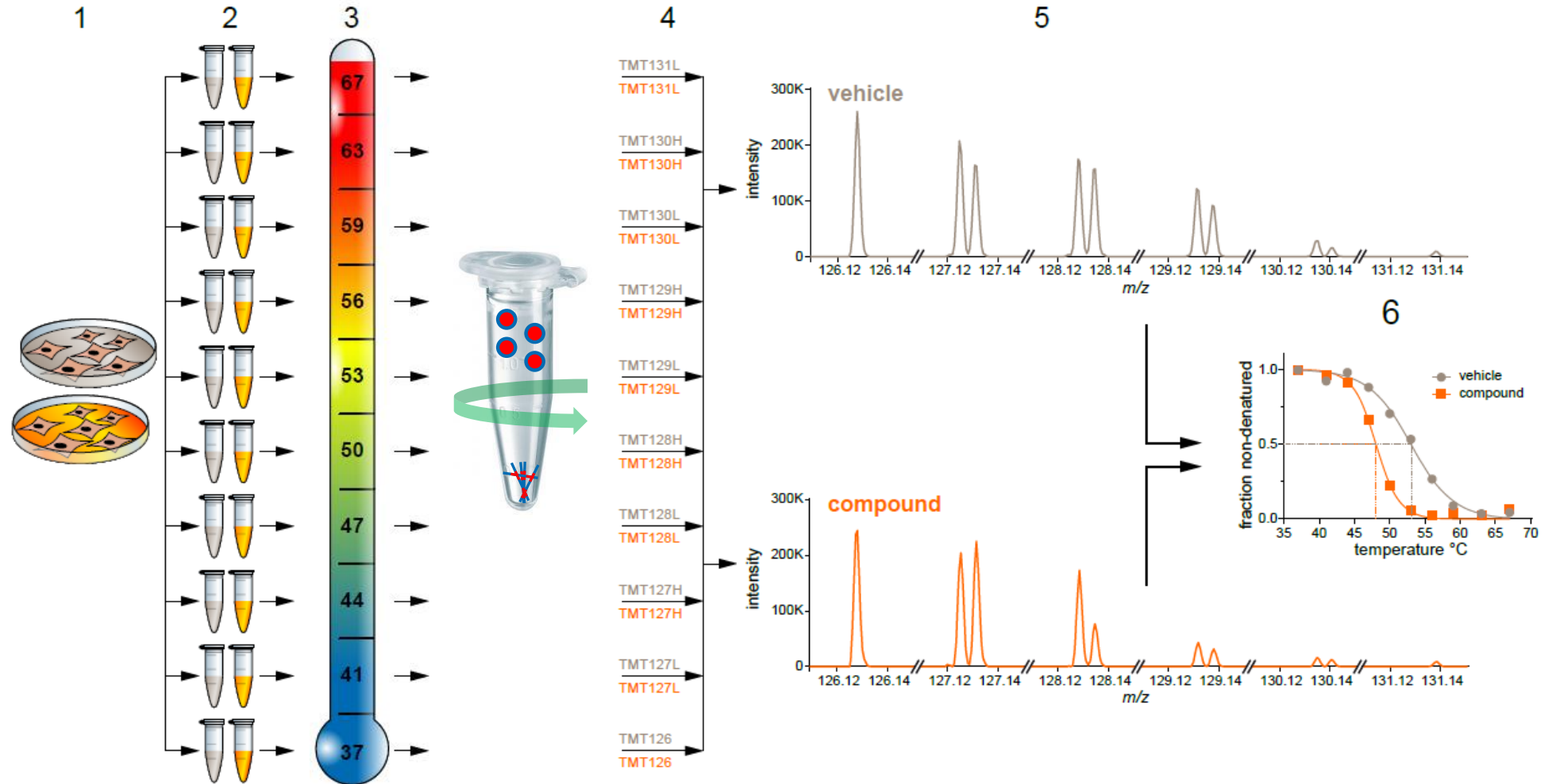


# Protein/drug interactions in living cells proteome-wide

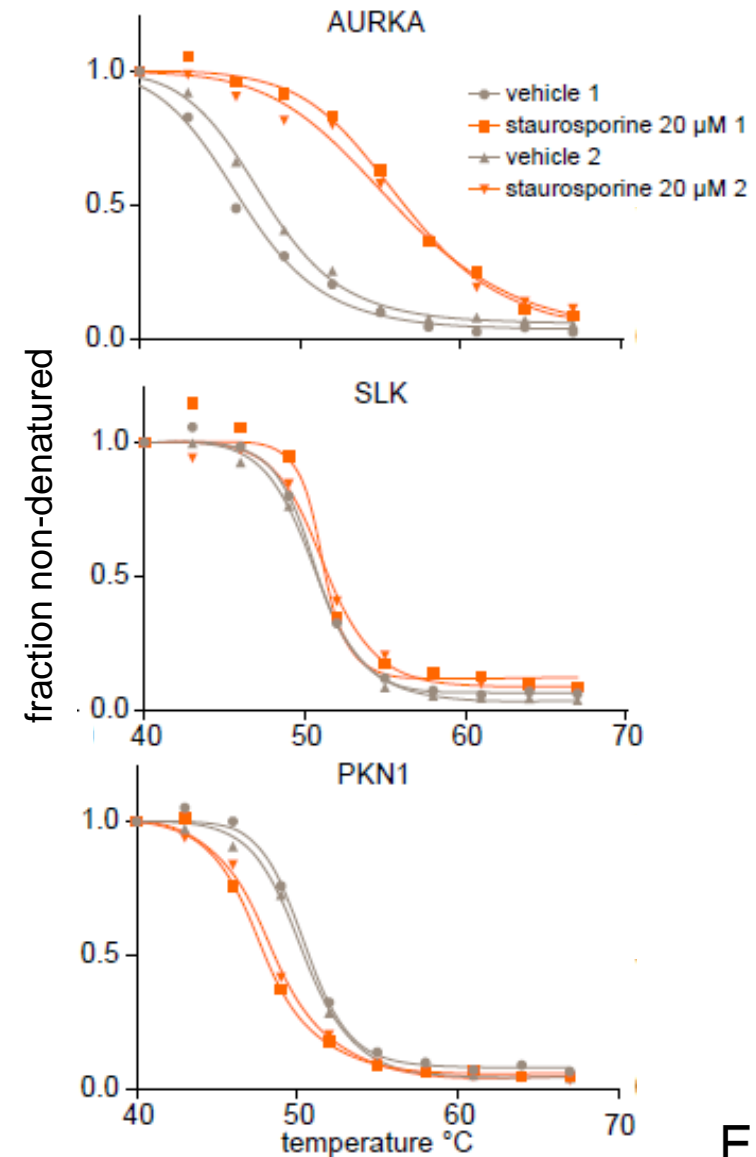
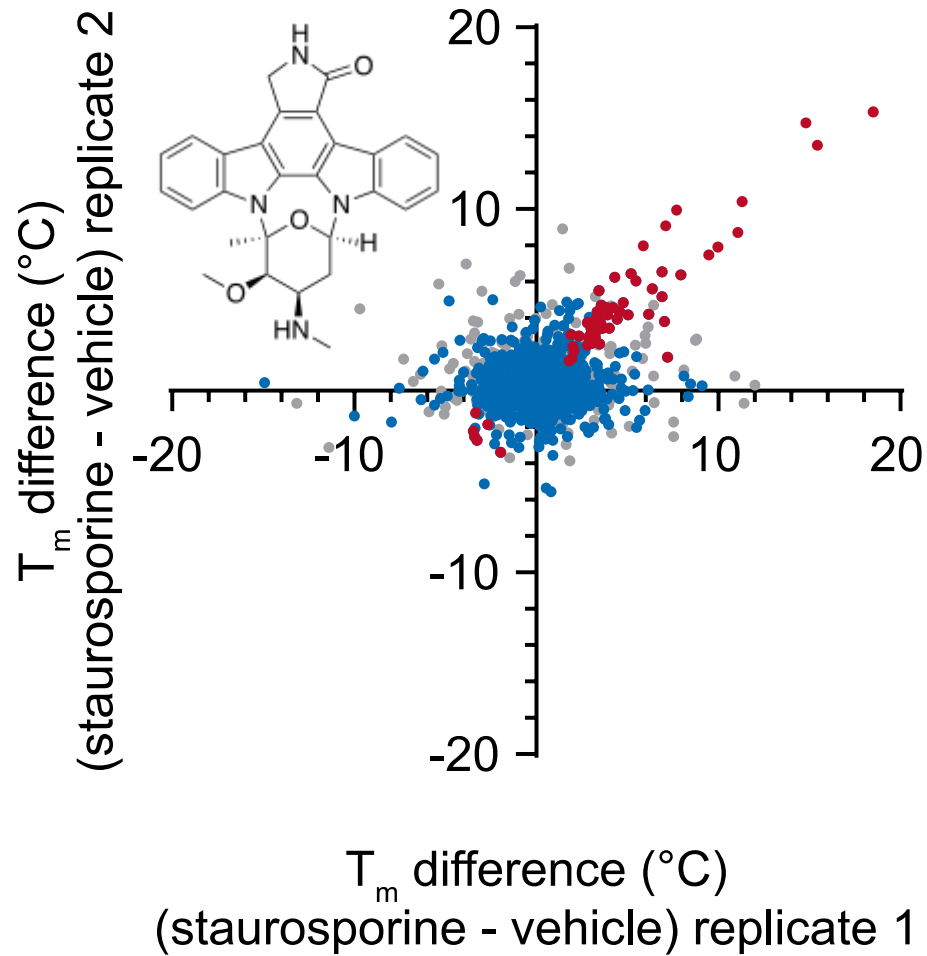




# Protein/drug interactions in living cells proteome-wide



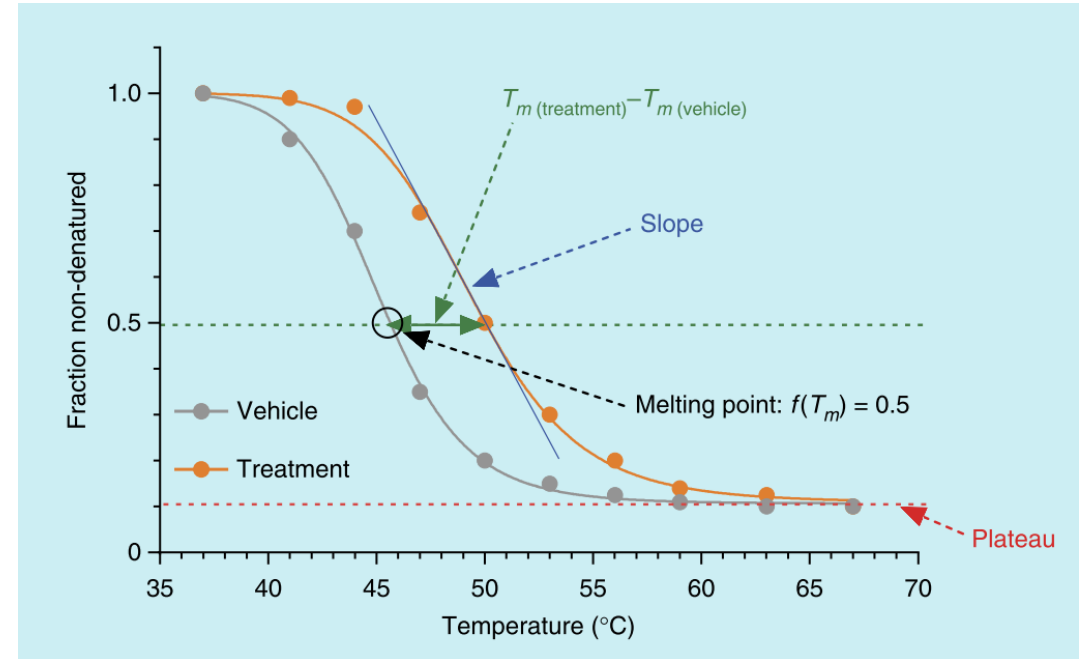
# Protein-Drug interaction: Staurosporine targets





# Melting curve fits

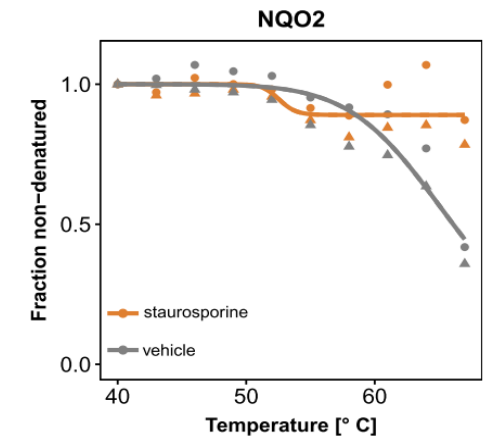
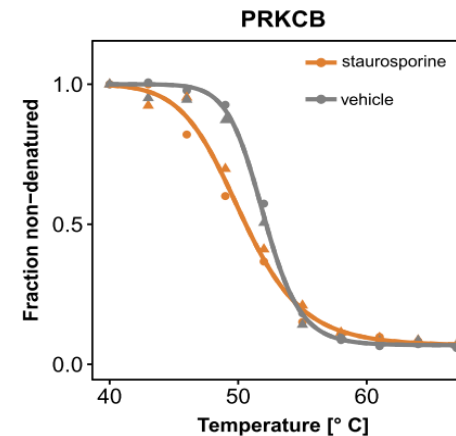
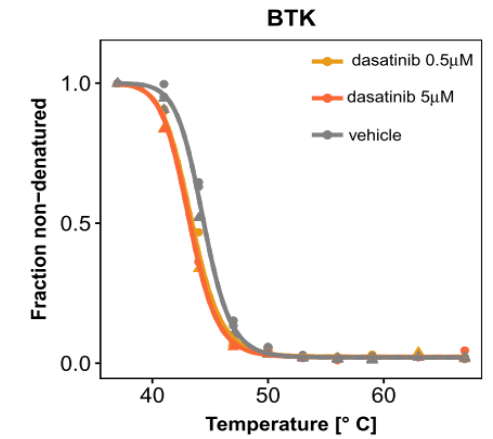
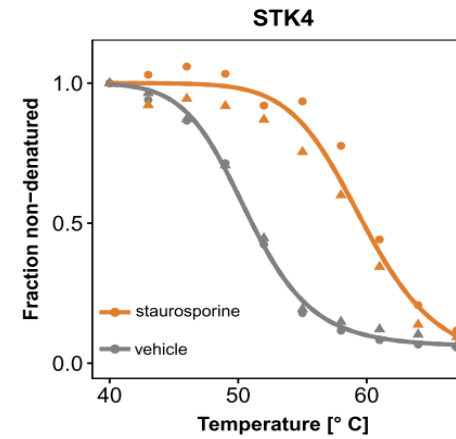
- Fit parametric sigmoidal curves for each condition
- Estimate melting points
- Compare melting points between the treatment conditions using a z-Test



# Problems with the melting point ( $T_m$ ) comparison

Several reasons can lead to  $T_m$  shift being insufficient to detect ligand effect:

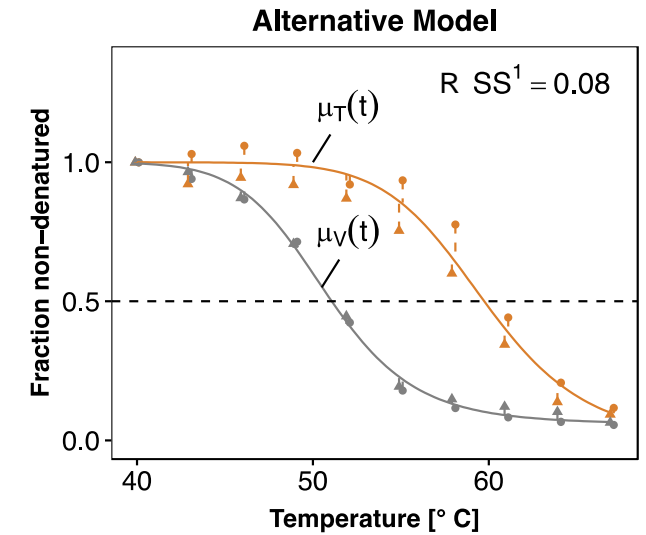
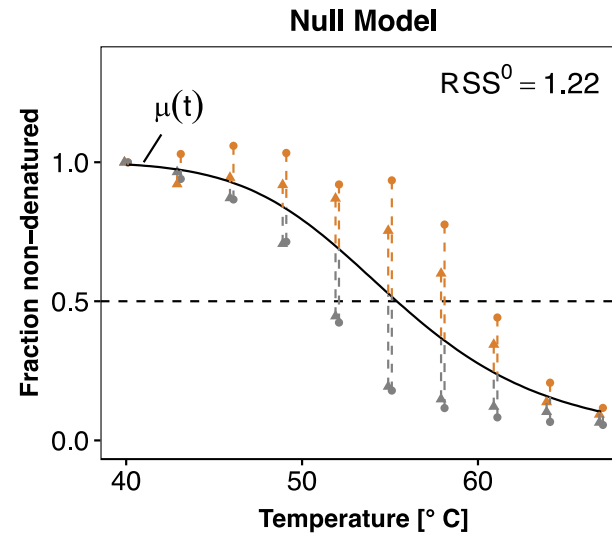
- Small but reproducible shifts (BTK)
- Shifts in non-centered curve areas (PRKCB)
- Melting points outside the temperature range (NQO2)



# Our solution: Functional melting curve analysis

Fit two competing models per protein  
Compute F-statistic:

$$F = \frac{RSS0 - RSS1}{RSS1}$$



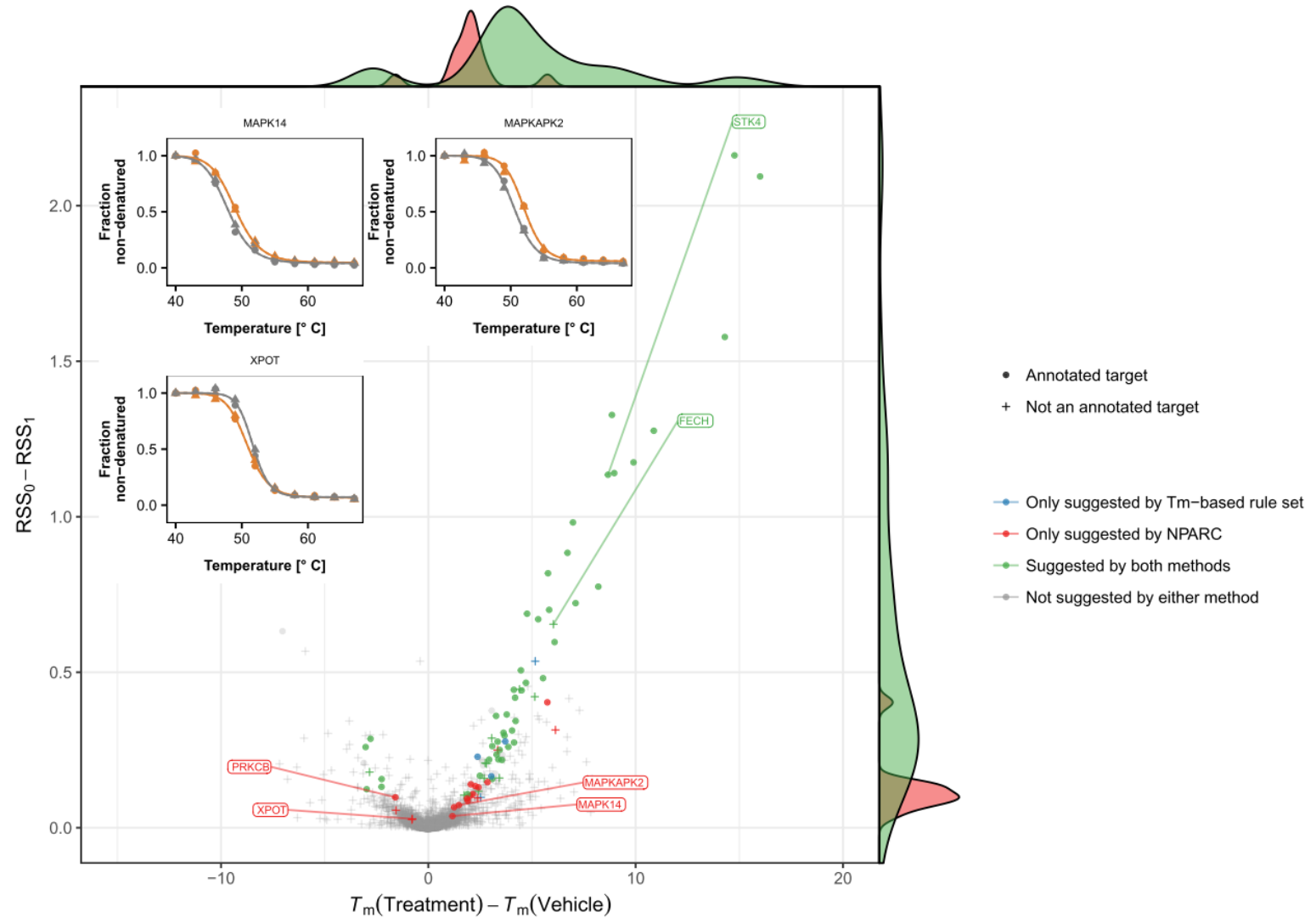
RSS: residual sum of squares



Karsten Bach Dorothee Childs Holger Franken

# Functional melting curve analysis

New method captures old targets  
+ cases with more subtle effects

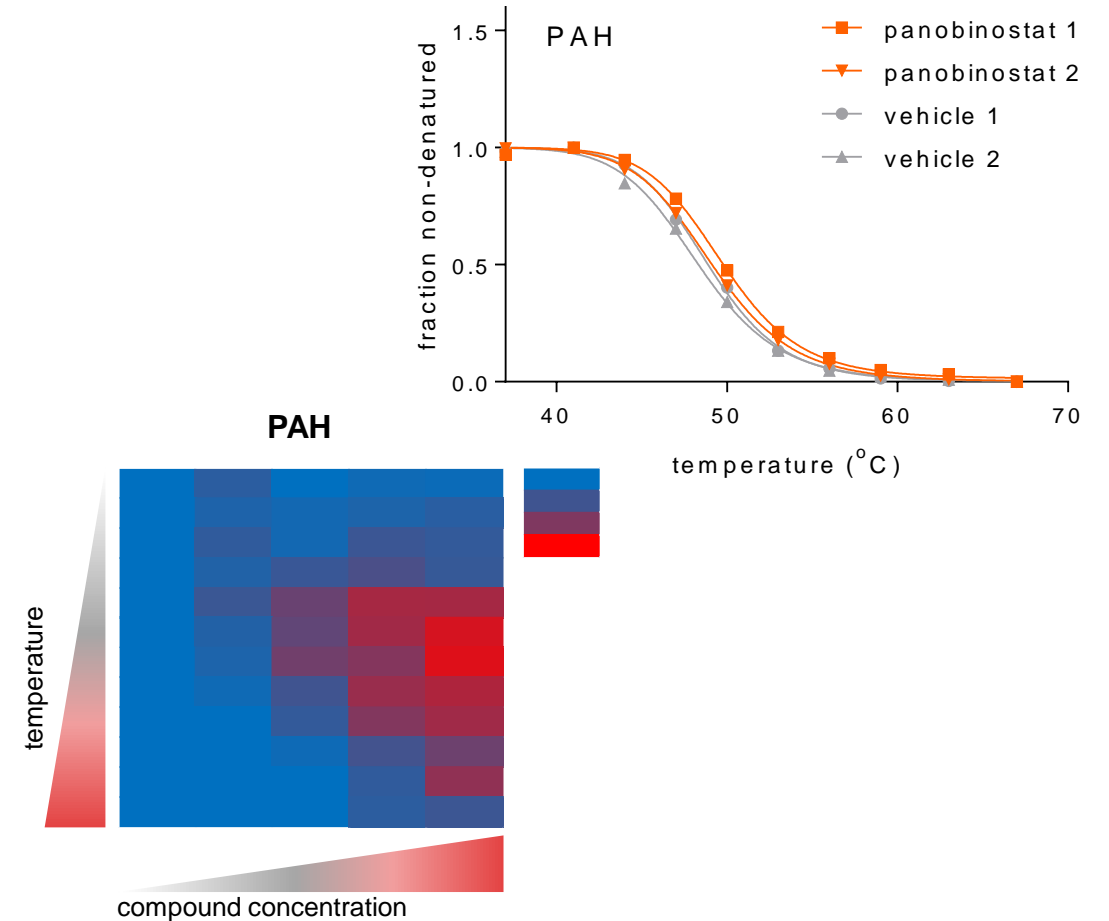
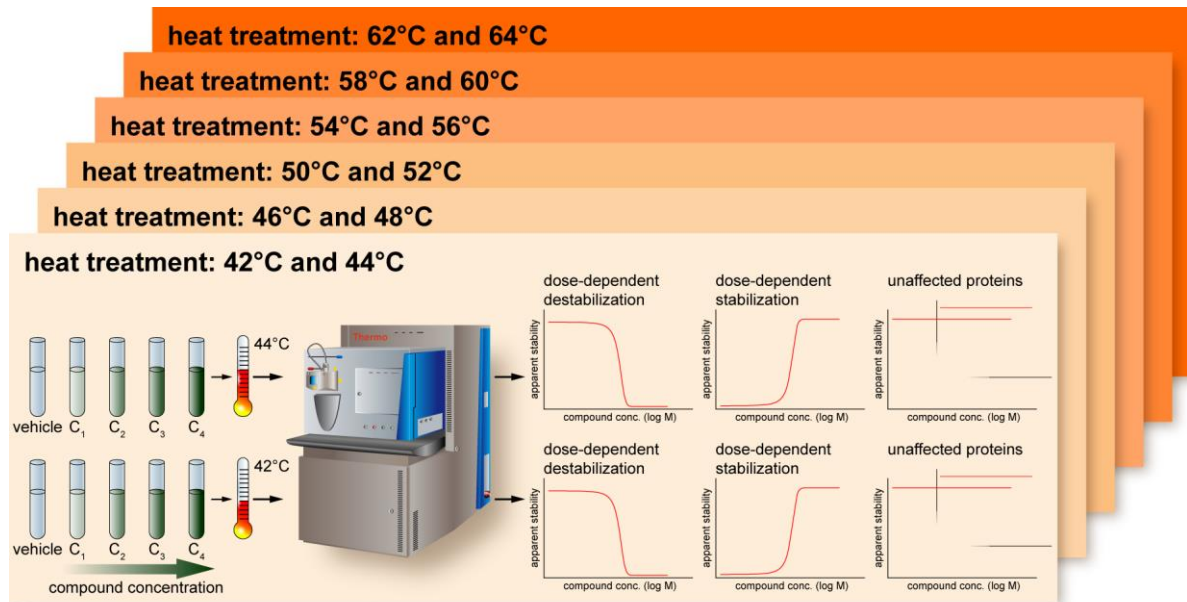


# More sensitive experimental design: 2D-TPP

Compound concentration-dependent profiling over a range of temperatures



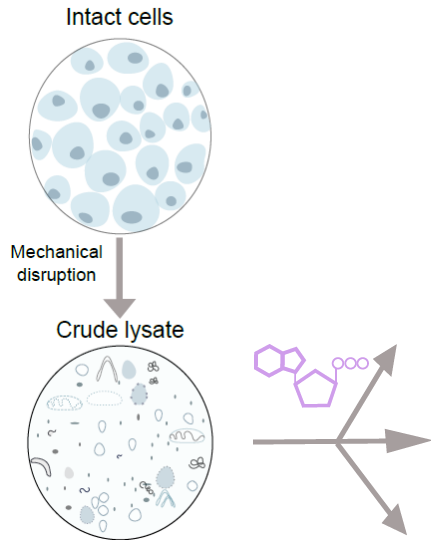
Instead of statistics: More experiments – more data – more targets



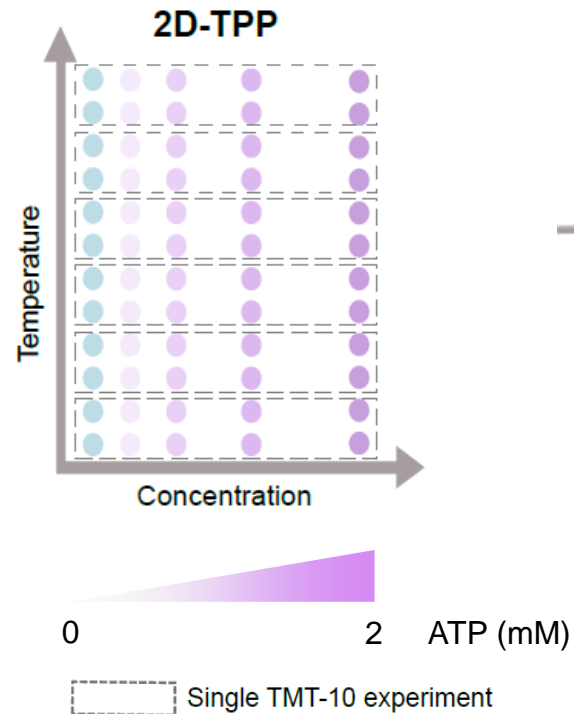
# Protein-Metabolite interaction: Adenosine triphosphate (ATP)



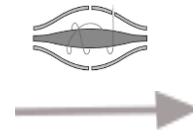
## Model system



## Metabolite and heat treatment

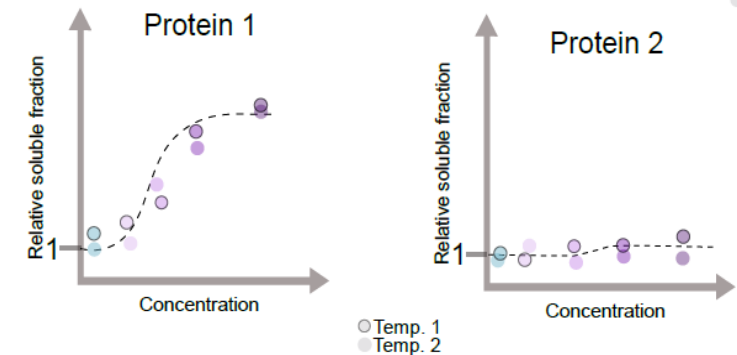


## MS-analysis



## Data analysis

### FDR controlled dose-response assessment

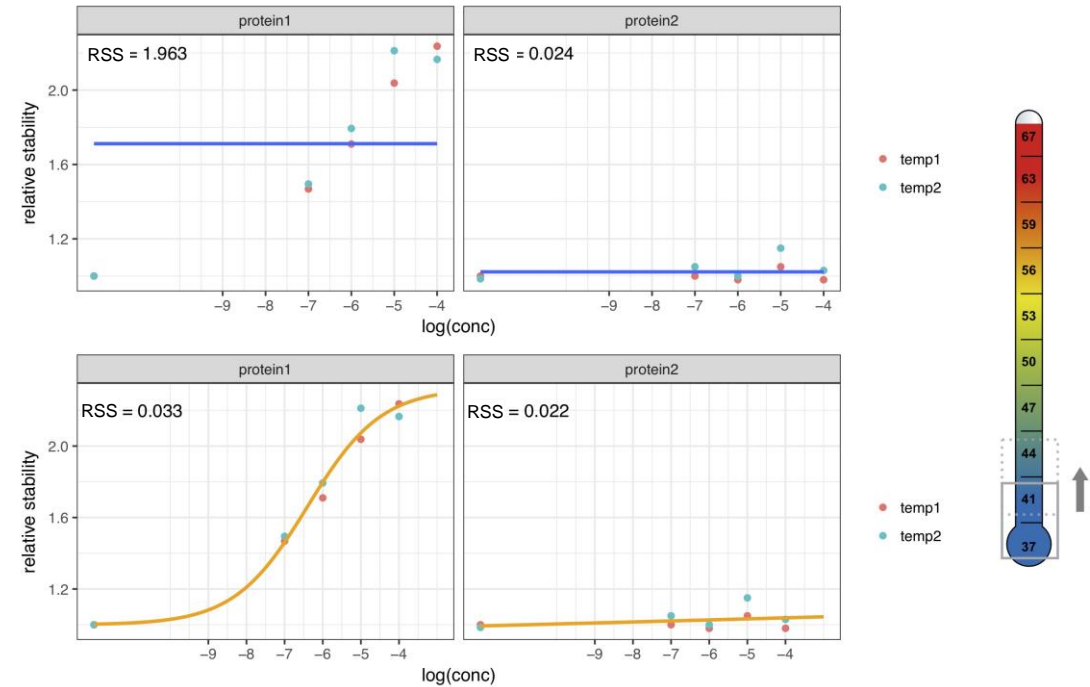


pEC<sub>50</sub> – measure of affinity



# Functional 2D-TPP data analysis using sliding temperature windows

- Fit models to adjacent temperatures
- **H0 Model:** intercept model (blue)
- **H1 Model:** dose-response curve (orange)
- Compare goodness of fit
- Summarize per protein score  $F^{\text{comb}}$
- Estimate the FDR for given scores by repeating the procedure with permuted data and ranking results jointly



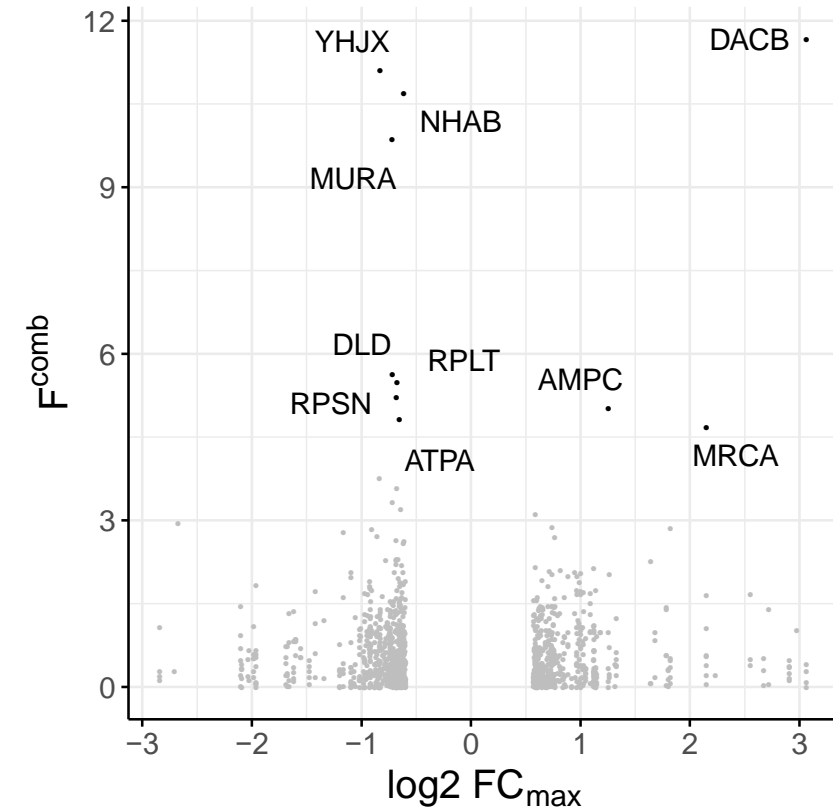
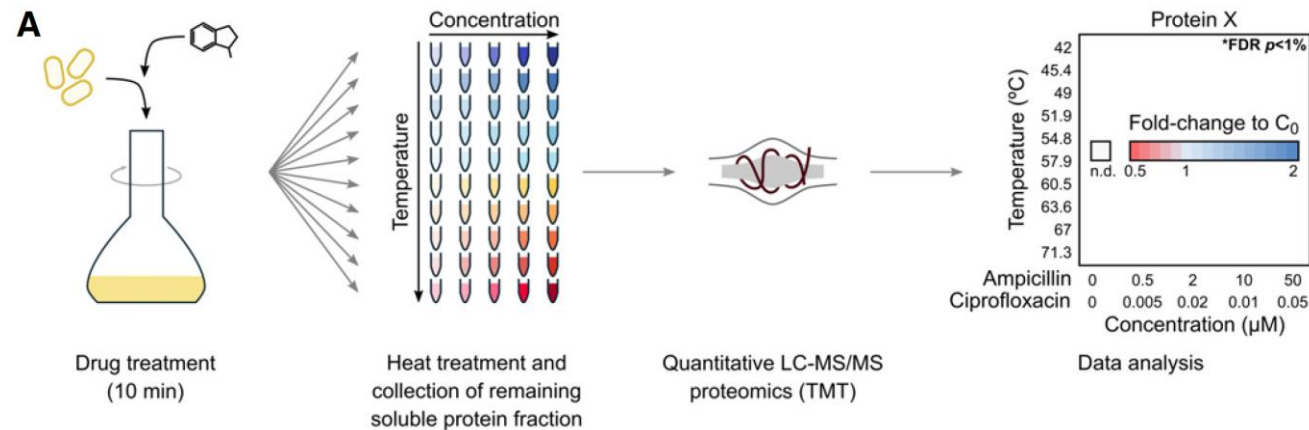
RSS: residual sum of squares

# Validation of the approach on drug datasets with known targets

- Ampicillin treated *E. coli* lysate
- beta-lactam antibiotic, inhibiting bacterial cell-wall synthesis
- Known targets: MrcA, FtsI, DacB & PbpG
- Other binders: AmpC



André Mateus





# Thermal proteome profiling

- Unbiased
- Proteome wide
- Versatile



**Thank you for your attention!**