ASMS 2020 Reboot Workshop

Machine Learning Analysis of Mass Spectrometry Data in the Life Sciences

William Stafford Noble, University of Washington, Seattle

Arzu Tugce Guler, Amsterdam University Medical Center, Amsterdam

Claire O’Donovan, European Bioinformatics Institute, Cambridge

Agenda

- ❖ Overview of machine learning techniques in mass spectrometry data analysis
- ❖ Panel:
  - *Machine-learning for the proteomic masses: learning peptide properties and clustering spectra,* **Lukas Käll, Science for Life Laboratory, Stockholm**
  - *Revolutionizing MS-based proteomics using machine learning techniques: fragmentation prediction, relative peptide intensity prediction, missing value imputation.* **Lars Jensen, University of Copenhagen**
  - *Bayesian learning and MS big data*, **Sam Payne, Brigham Young University, Salt Lake City**
  - *Machine learning applications for real-time analysis*, **Devin Schweppe, Harvard University**
- ❖ Discussion

What is a learning behaviour?

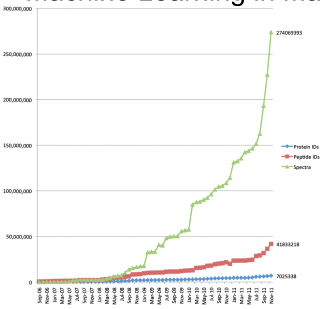
Given a task *T*, a performance criterion *C*, and experience *E*, a system **learns** from *E* if it becomes better at solving task *T*, as measured by criterion *C*, by exploiting the information in *E*. *

We need to know:

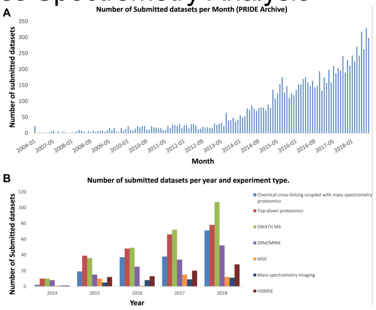
- ➢ which task is going to be performed
- ➢ how the performance on the task is measured
- ➢ what kind of information is used by the system

T. Mitchell. *Machine Learning.* McGraw-Hill 1996

Machine Learning in Mass Spectrometry Analysis



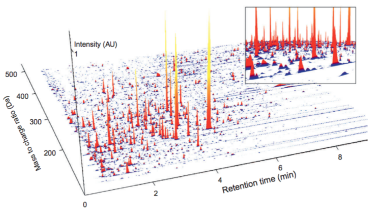
Csordas et al. *Database* 2012

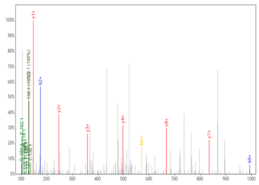


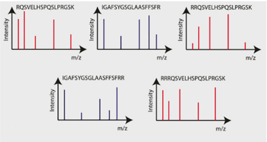
Perez-Riverol et al. *Nucleic Acids Research* 2012

Machine Learning in Mass Spectrometry Analysis

Data is complex, noisy, non-trivial







1.Doana. *Metabolic Phenotyping in Personalized and Public Healthcare* 2016

2.https://phosphopedia.gs.washington.edu/PhosphoproteomicsAssay/index.xhtml

3.Saeed et. al IE*EE/ACM Trans Comput Biol Bionform* 2018

Retention time prediction

Input: peptide sequence

Output: chromatographic retention time

Utility: boost statistical power to detect peptides

Proc. Natl. Acad. Sci. USA
Vol. 77, No. 3, pp. 1632–1636, March 1980
Medical Sciences

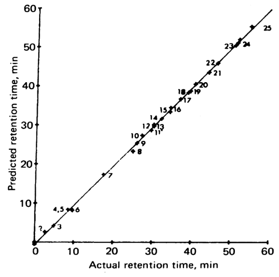
Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition

(lipophilicity/separation techniques)

JAMES L. MEEK

Laboratory of Preclinical Pharmacology, National Institute of Mental Health, Saint Elizabeths Hospital, Washington, D.C. 20032

Communicated by Bruce Merrifield, December 17, 1979

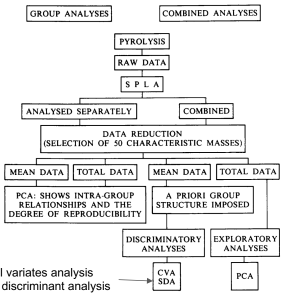


Species identification

Input: MS1 or MS2 data from a given sample

Output: (list of) detected species

Utility:



GROUP ANALYSES

COMBINED ANALYSES

PYROLYSIS

RAW DATA

S P L A

ANALYSED SEPARATELY

COMBINED

DATA REDUCTION
(SELECTION OF 50 CHARACTERISTIC MASSES)

MEAN DATA

TOTAL DATA

MEAN DATA

TOTAL DATA

PCA: SHOWS INTRA-GROUP RELATIONSHIPS AND THE DEGREE OF REPRODUCIBILITY

A PRIORI GROUP STRUCTURE IMPOSED

DISCRIMINATORY ANALYSES

EXPLORATORY ANALYSES

CVA
SDA

PCA

Canonical variates analysis
Stepwise discriminant analysis

Shute et al. *Microbiology* 1984

Classifying peptide-spectrum matches

Input: a vector of features associated with a peptide-spectrum match

Output: Is this peptide responsible for generating this spectrum?

Utility: boost statistical power to detect peptides / proteins

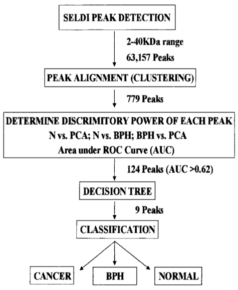
Keller et al. *Analytical Chemistry* 2002

Phenotype / disease classification

Input: MS1 or MS2 data from an individual

Output: predicted phenotype

Utility: early diagnosis or disease prognosis



SELDI PEAK DETECTION

2-40KDa range
63,157 Peaks

PEAK ALIGNMENT (CLUSTERING)

779 Peaks

DETERMINE DISCRIMITORY POWER OF EACH PEAK
N vs. PCA; N vs. BPH; BPH vs. PCA
Area under ROC Curve (AUC)

124 Peaks (AUC >0.62)

DECISION TREE

9 Peaks

CLASSIFICATION

CANCER

BPH

NORMAL

Adam et al. *Cancer Research* 2002

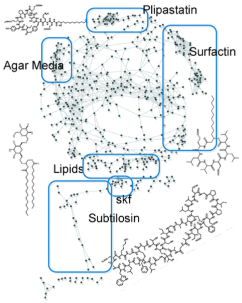
Clustering of mass spectra

Input: large collection of mass spectra

Output: cluster assignments

Utility:

- boost statistical power to detect peptides
- speed up database search



Tabb et al. *Analytical Chemistry* 2003

Guthals et al. *Molecular BioSystems* 2012

SciLifeLab

MS and ML

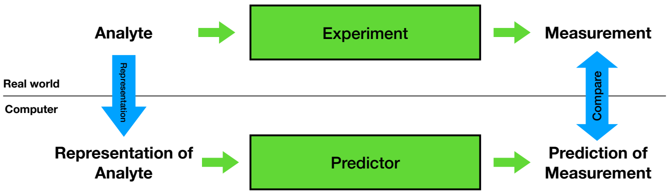
Lukas Käll
Royal Institute of Technology - KTH
School of Biotechnology
Stockholm, Sweden

KTH VETENSKAP OCH KONST

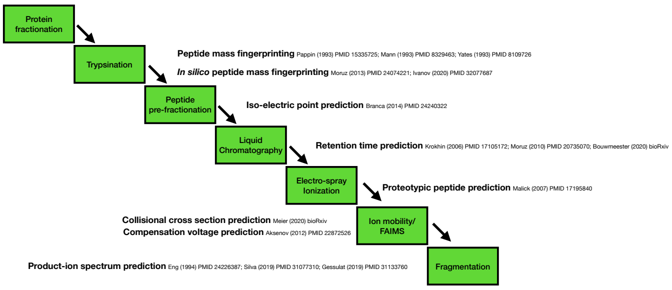
http://percolator.ms
http://kaell.org

Background on machine learning for MS-based proteomics

Identification by comparing observations to predictions



MS-related separation techniques and their predictors



Peptide mass fingerprinting Pappin (1993) PMID 15335725; Mann (1993) PMID 8329463; Yates (1993) PMID 8109726

***In silico* peptide mass fingerprinting** Moruz (2013) PMID 24074221; Ivanov (2020) PMID 32077687

Iso-electric point prediction Branca (2014) PMID 24240322

Retention time prediction Krokhin (2006) PMID 17105172; Moruz (2010) PMID 20735070; Bouwmeester (2020) bioRxiv

Proteotypic peptide prediction Malick (2007) PMID 17195840

Collisional cross section prediction Meier (2020) bioRxiv

Compensation voltage prediction Aksenov (2012) PMID 22872526

Product-ion spectrum prediction Eng (1994) PMID 24226387; Silva (2019) PMID 31077310; Gessulat (2019) PMID 31133760

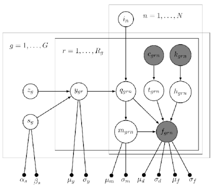
Mining Repositories

- Clustering offers a nice way to condense information from prior experiments
  Frank et al 2008;Griss et al
- Yade yade

1 case

Quandenser:

Combining quantification and clustering



Input: Raw-files from LFQ experiments

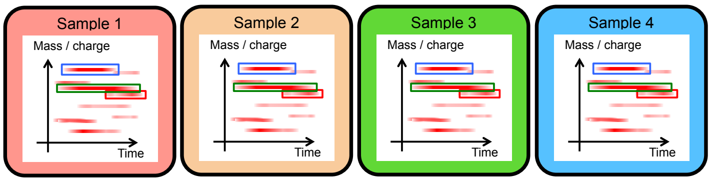
Output: proteins with *Pr(diff. exp.)*

Binaries: `https://github.com/statisticalbiotechnology/quandenser`

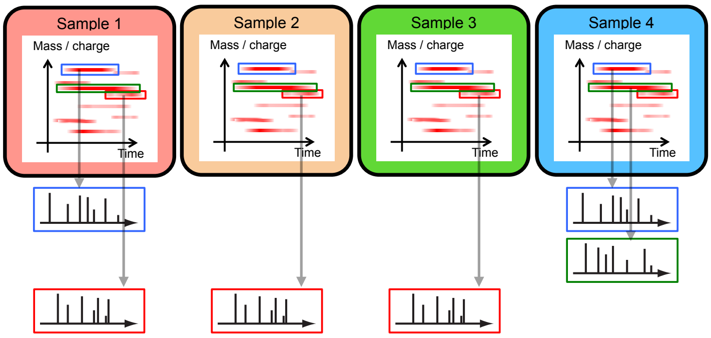
Paper: The & Käll Nature Communications (in press)

Dr. Matthew The

MS1 features can often be retrieved across samples…

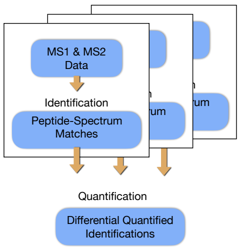


… but MS2 spectra are sparse



Flipping the pipeline

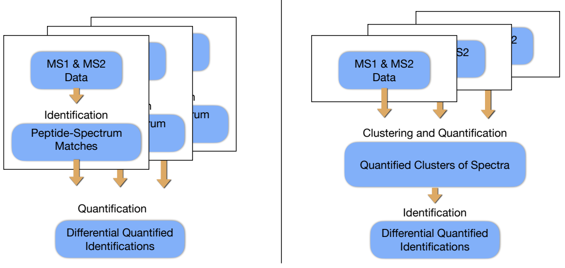
Identification-first



Flipping the pipeline

Identification-first

Quantification-first

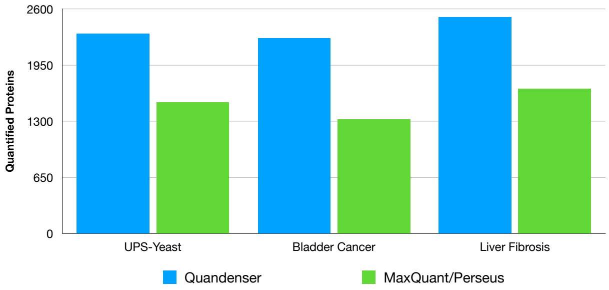

MS1 & MS2 Data
Identification
Peptide-Spectrum Matches
Quantification
Differential Quantified Identifications
MS1 & MS2 Data
Clustering and Quantification
Quantified Clusters of Spectra
Identification
Differential Quantified Identifications

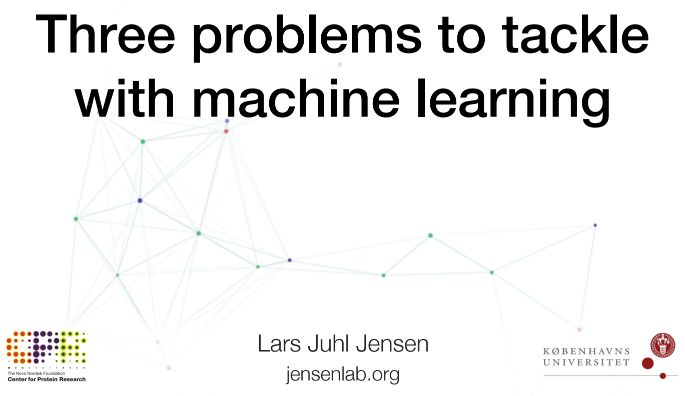
Benefits of quantification-first

- No need to rediscover the same peptides for each run
- Lowering the number of spectra
  - Enables more advanced identification strategies
  - Faster identification
  - Fewer hypothesis tests

Focus on the spectra that matters!

Dramatic increase in number of differentially quantified proteins at 5% FDR

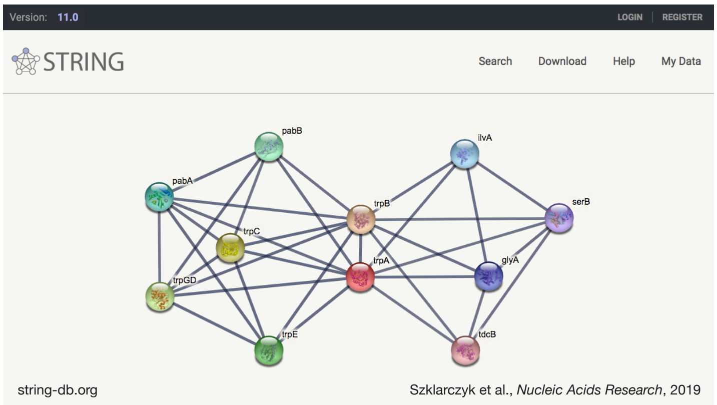




Three problems to tackle with machine learning

Lars Juhl Jensen
jensenlab.org

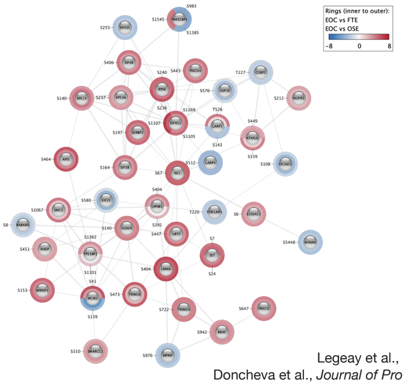
network biology



string-db.org

Szklarczyk et al., *Nucleic Acids Research*, 2019

data visualization



Legeay et al., *F1000 Research*, 2020
Doncheva et al., *Journal of Proteome Research*, 2019

1

peptide fragmentation

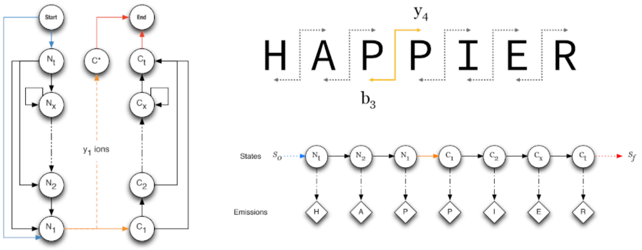
fragment ion spectrum

theoretical spectra

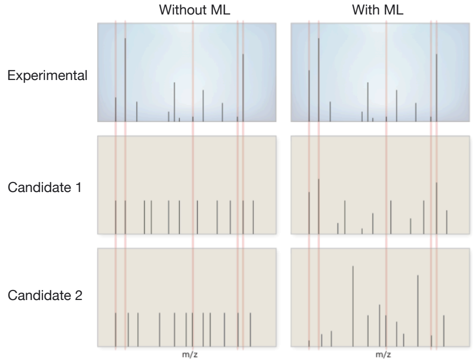
count matches

assume all are equal

fragmentation is predictable



Kirik, Refsgaard & Jensen, *Journal of Proteome Research*, 2019 better identification



Kirik, Refsgaard & Jensen, *Journal of Proteome Research*, 2019

better PTM localization

2

peptide abundance

one protein

multiple unique peptides

equal abundance

different intensities

plenty of training data

peptides + MS parameters

relative intensities

better protein quantification

cross-sample comparison

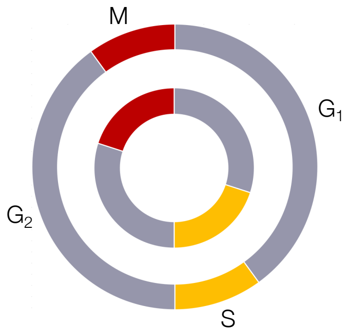
3 confounding effects

phosphoproteomics

kinase motif analysis

regulation of CDK activity

different growth rates



a few common effects

affect numerous peptides

in a consistent manner

many unrelated experiments

dimensionality reduction

auto encoder

learn effect signatures

quantify their impact

residual signals

?

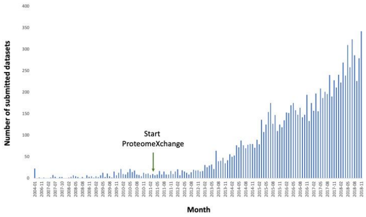
questions

Bayesian Learning and MS Big Data

Samuel Payne
Brigham Young University
ASMS 2020 - Bioinformatics Interest Group

Data, Data Everywhere. Are we learning?

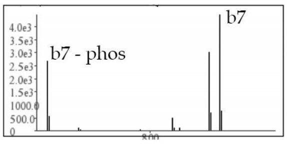
How can we learn from public mass spectrometry data?

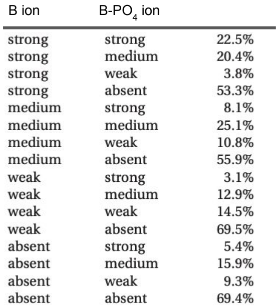


"We have seen this spectrum a lot of times. Why do we always pretend to not know anything about it?"
--- Mike MacCoss

Learning from our data

What can we learn about expected intensity from data mining?





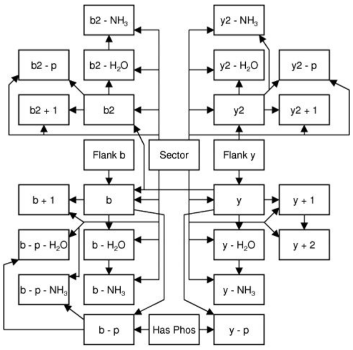
B ion	B-PO_4 ion	
strong	strong	22.5%
strong	medium	20.4%
strong	weak	3.8%
strong	absent	53.3%
medium	strong	8.1%
medium	medium	25.1%
medium	weak	10.8%
medium	absent	55.9%
weak	strong	3.1%
weak	medium	12.9%
weak	weak	14.5%
weak	absent	69.5%
absent	strong	5.4%
absent	medium	15.9%
absent	weak	9.3%
absent	absent	69.4%

Payne et al. 2008
https://doi.org/10.1021/pr800129m

Building your Bayes Net

Identify mutual information

More nodes is not always better



Training

Inspect (deprecated) trained on 170,000 phospho-peptide spectra (2008) to learn the probabilities (and joint probabilities) of fragment ion intensity

MSGF+ trained on 2.8 million spectra (2014) to learn the probabilities of fragment ion intensities

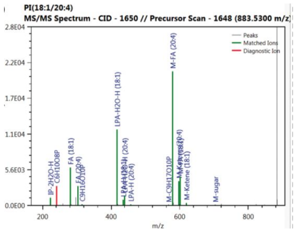
New Areas for Bayesian Scoring in MS id

"We have seen this spectrum a lot of times. Why do we always pretend to not know anything about it?"
--- Mike MacCoss

Spectral Library Matching (DIA or metabolomics)
- are my peak relative intensities as expected?
- how often does this peak have interference?

Lipids and Bayes Learning

Learn the fragment ions
Learn their intensity, relative intensity



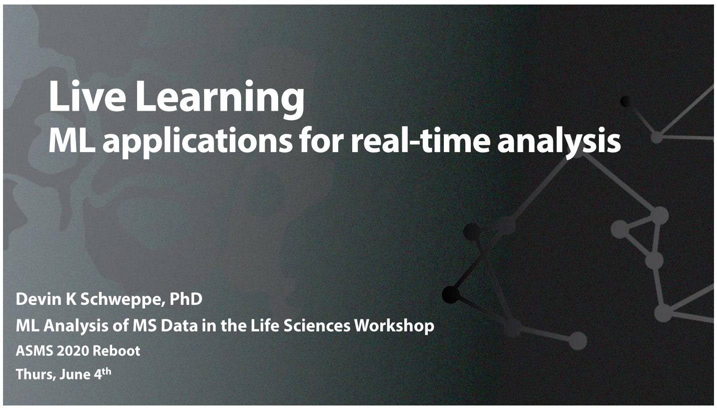
Kyle et al. 2017
https://doi.org/10.1093/bioinformatics/btx046

New Areas for Bayesian Scoring

Protein Inference =? Bayesian Inference

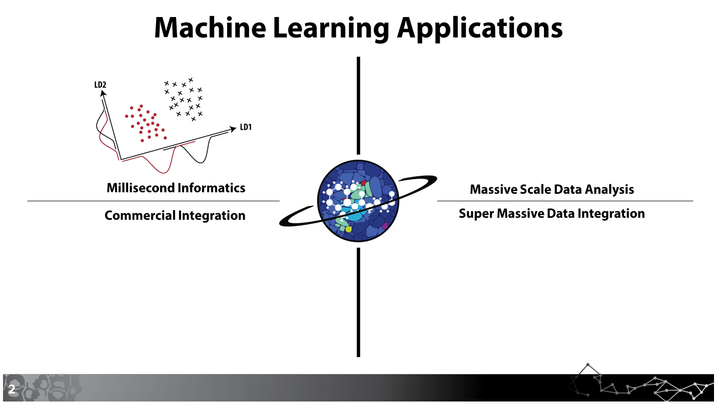
- Do peptide intensities within a protein have a reliable relationship?

- Do we need to learn for each tissue or cell line?



Live Learning
ML applications for real-time analysis

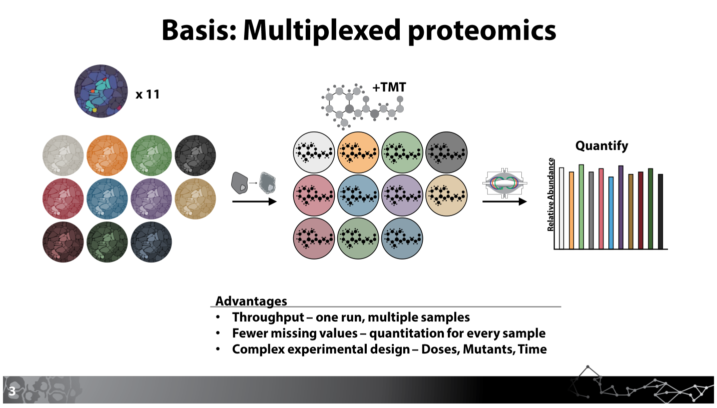
Devin K Schweppe, PhD
ML Analysis of MS Data in the Life Sciences Workshop
ASMS 2020 Reboot
Thurs, June 4th



Machine Learning Applications

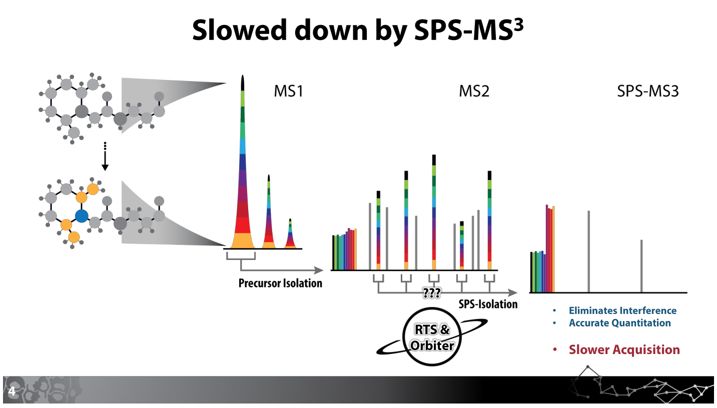
Millisecond Informatics
Commercial Integration

Massive Scale Data Analysis
Super Massive Data Integration



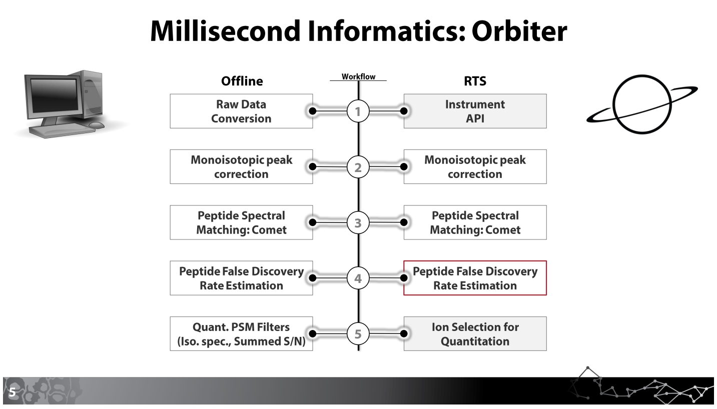
Basis: Multiplexed proteomics

Advantages
- Throughput – one run, multiple samples
- Fewer missing values – quantitation for every sample
- Complex experimental design – Doses, Mutants, Time



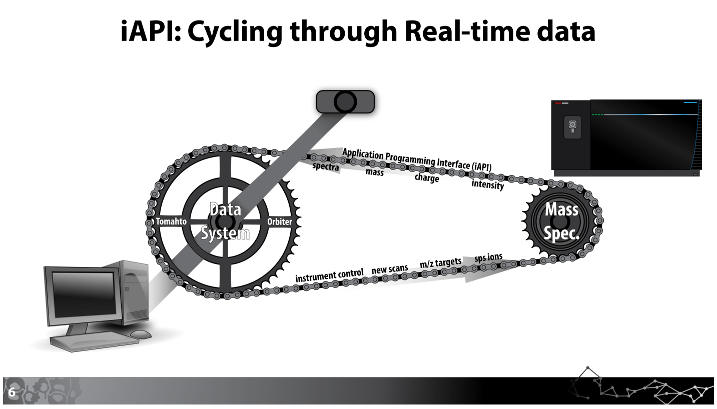
Slowed down by SPS-MS^3

- Eliminates Interference
- Accurate Quantitation
- Slower Acquisition



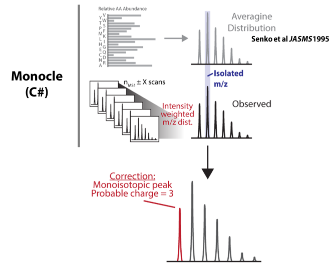
Millisecond Informatics: Orbiter

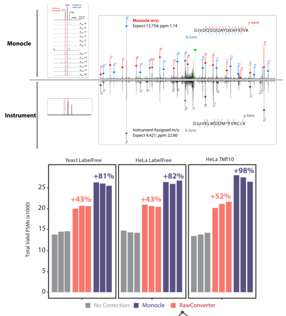
Offline	Workflow	RTS
Raw Data Conversion	1	Instrument API
Monoisotopic peak correction	2	Monoisotopic peak correction
Peptide Spectral Matching: Comet	3	Peptide Spectral Matching: Comet
Peptide False Discovery Rate Estimation	4	Peptide False Discovery Rate Estimation
Quant. PSM Filters (Iso. spec., Summed S/N)	5	Ion Selection for Quantitation



iAPI: Cycling through Real-time data

Monoisotopic Peak Correction





Rad et al *(in preparation)*

Comet: Open source search engine

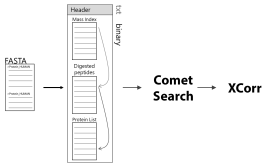
22 DOI 10.1002/pmic.201200439 Proteomics 2013, 13, 22–24

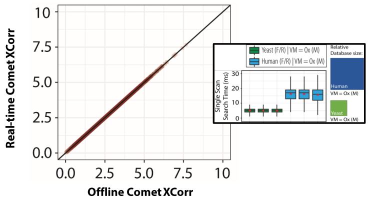
TECHNICAL BRIEF

Comet: An open-source MS/MS sequence database search tool

Jimmy K. Eng[1], Tahmina A. Jahan[1] and Michael R. Hoopmann[2]

[1] Department of Genome Sciences, University of Washington, Seattle, WA, USA
[2] Institute for Systems Biology, Seattle, WA, USA

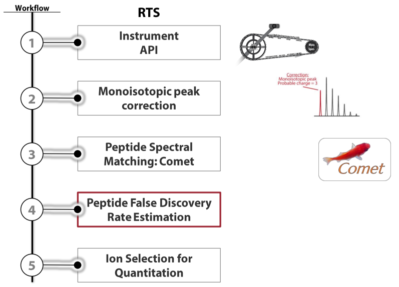




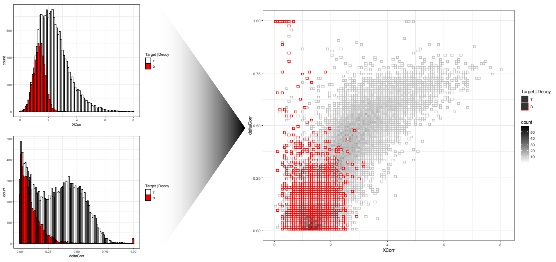
Real-time Comet scoring with indexed databases shows consistent XCorr results compared to offline search.

Jimmy Eng
Schweppe et al *JPR* 2020

Millisecond Informatics: Orbiter

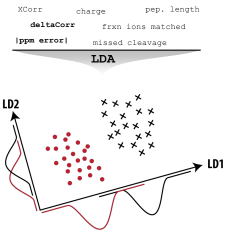

Workflow
RTS
1
Instrument API
2
Monoisotopic peak correction
3
Peptide Spectral Matching: Comet
4
Peptide False Discovery Rate Estimation
5
Ion Selection for Quantitation
Correction:
Monoisotopic peak
Probable charge = 3
Comet

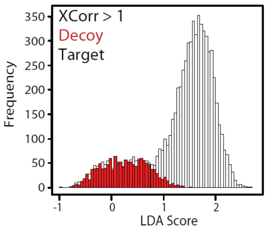
Towards real-time FDR filtering



Combining peptide spectral match scores (e.g. XCorr and deltaCorr) aids separation of target and decoy populations.

Linear Discriminant Analysis (LDA)

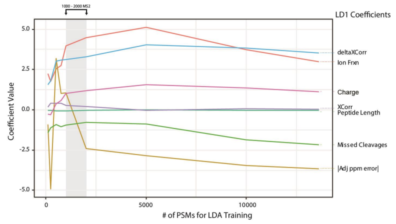




LDA enables fast target-decoy classification and FDR estimation

Schweppe et al *JPR* 2020

Accumulate Training Data

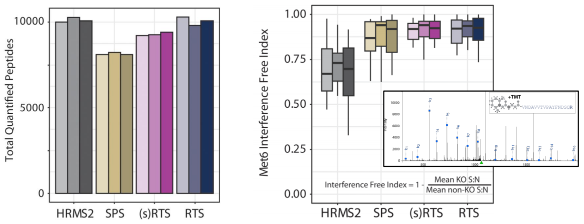


XCorr/deltaCorr/PPM ↔

LDA + FDR ↔

1000-2000 total PSMs collected
At least 10 decoy PSMs*
*Huttlin et al 2007

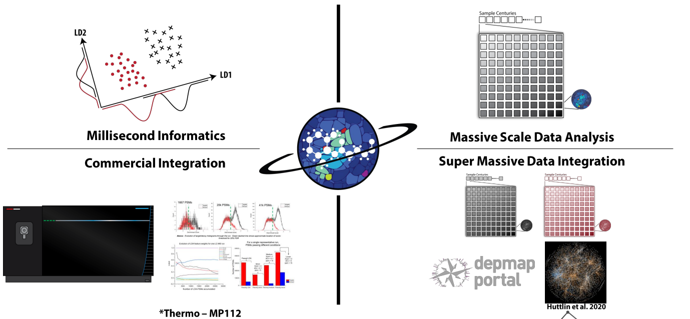
Improved sensitivity & quantitative accuracy



Real-time FDR filtering increases the number of quantified peptides and eliminated isobaric tag interference.

Schweppe et al *JPR* 2020

Moving Forward with ML



Millisecond Informatics

Commercial Integration

Massive Scale Data Analysis

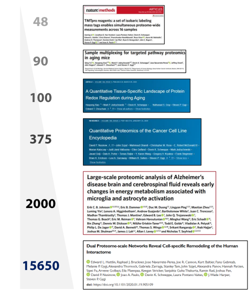
Super Massive Data Integration

Huttlin et al. 2020

*Thermo – MP112

As sample size grows…

# of samples



48 — nature methods ARTICLES

TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples

Jiaming Li, Jonathon G. Van Vranken, Laura Pontano Vaites, Devin K. Schweppe, Edward L. Huttlin, Chris Etienne, Premchendar Nandhikonda, Rosa Viner, Aaron M. Robitaille, Andrew H. Thompson, Karsten Kuhn, Ian Pike, Ryan D. Bomgarden, John C. Rogers, Steven P. Gygi and Joao A. Paulo

90 — Sample multiplexing for targeted pathway proteomics in aging mice

Qing Yu, Haopeng Xiao, Mark P. Jedrychowski, Devin K. Schweppe, Jose Navarrete-Perea, Jeffrey Knott, John Rogers, Edward T. Chouchani, and Steven P. Gygi

[illegible]

100 — ARTICLE | VOLUME 180, ISSUE 5, P968-983.E24, MARCH 05, 2020

A Quantitative Tissue-Specific Landscape of Protein Redox Regulation during Aging

Haopeng Xiao • Mark P. Jedrychowski • Devin K. Schweppe • … Nathanael S. Gray • Steven P. Gygi • Edward T. Chouchani • Show all authors • Show footnotes

375 — RESOURCE | VOLUME 180, ISSUE 2, P387-402.E16, JANUARY 23, 2020

Quantitative Proteomics of the Cancer Cell Line Encyclopedia

David P. Nusinow • John Szpyt • Mahmoud Ghandi • Christopher M. Rose • E. Robert McDonald III • Marian Kalocsay • Judit Jané-Valbuena • Ellen Gelfand • Devin K. Schweppe • Mark Jedrychowski • Javad Golji • Dale A. Porter • Tomas Rejtar • Y. Karen Wang • Gregory V. Kryukov • Frank Stegmeier • Brian K. Erickson • Levi A. Garraway • William R. Sellers • Steven P. Gygi • Show less • Show footnotes

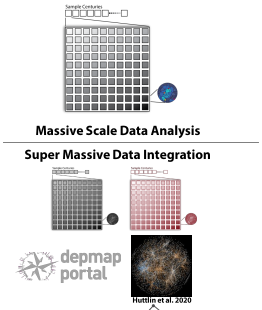
2000 — Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation

Erik C. B. Johnson, Eric B. Dammer, Duc M. Duong, Lingyan Ping, Maotian Zhou, Luming Yin, Lenora A. Higginbotham, Andrew Guajardo, Bartholomew White, Juan C. Troncoso, Madhav Thambisetty, Thomas J. Montine, Edward B. Lee, John Q. Trojanowski, Thomas G. Beach, Eric M. Reiman, Vahram Haroutunian, Minghui Wang, Eric Schadt, Bin Zhang, Dennis W. Dickson, Nilüfer Ertekin-Taner, Todd E. Golde, Vladislav A. Petyuk, Philip L. De Jager, David A. Bennett, Thomas S. Wingo, Srikant Rangaraju, Ihab Hajjar, Joshua M. Shulman, James J. Lah, Allan I. Levey and Nicholas T. Seyfried

15650 — Dual Proteome-scale Networks Reveal Cell-specific Remodeling of the Human Interactome

Edward L. Huttlin, Raphael J. Bruckner, Jose Navarrete-Perea, Joe R. Cannon, Kurt Baltier, Fana Gebreab, Melanie P. Gygi, Alexandra Thornock, Gabriela Zarraga, Stanley Tam, John Szpyt, Alexandra Panov, Hannah Parzen, Sipei Fu, Arvene Golbazi, Eila Maenpaa, Keegan Stricker, Sanjukta Guha Thakurta, Ramin Rad, Joshua Pan, David P. Nusinow, Joao A. Paulo, Devin K. Schweppe, Laura Pontano Vaites, J. Wade Harper, Steven P. Gygi

doi: https://doi.org/10.1101/2020.01.19.905109

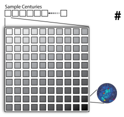


Massive Scale Data Analysis

Super Massive Data Integration

Huttlin et al. 2020

Latent representations of graphs with DeepWalk



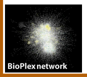
# of samples

15650

Dual Proteome-scale Networks Reveal Cell-specific Remodeling of the Human Interactome

Edward L. Huttlin, Raphael J. Bruckner, Jose Navarrete-Perea, Joe R. Cannon, Kurt Baltier, Fana Gebreab, Melanie P. Gygi, Alexandra Thornock, Gabriela Zarraga, Stanley Tam, John Szpyt, Alexandra Panov, Hannah Parzen, Sipei Fu, Arvene Golbazi, Eila Maenpaa, Keegan Stricker, Sanjukta Guha Thakurta, Ramin Rad, Joshua Pan, David P. Nusinow, Joao A. Paulo, Devin K. Schweppe, Laura Pontano Vaites, J. Wade Harper, Steven P. Gygi

doi: https://doi.org/10.1101/2020.01.19.905109

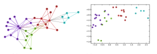


DeepWalk: Online Learning of Social Representations

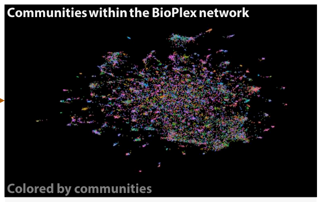
Bryan Perozzi, Stony Brook University, Department of Computer Science

Rami Al-Rfou, Stony Brook University, Department of Computer Science

Steven Skiena, Stony Brook University, Department of Computer Science



(a) Input: Karate Graph (b) Output: Representation

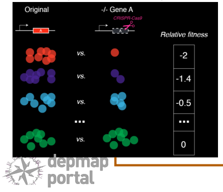


What is the interplay of protein communities within the BioPlex network?

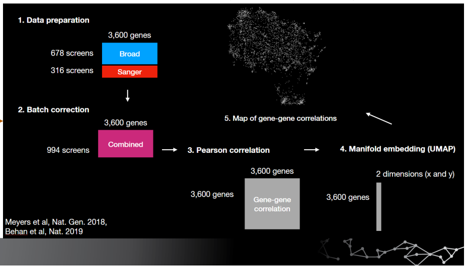
Josh Pan
Ed Huttlin

Manifold embedding of cellular fitness profiles

Dependency Map Project
(Cellular Fitness)



Related genes co-vary in their dependency



1. Data preparation

2. Batch correction

3. Pearson correlation

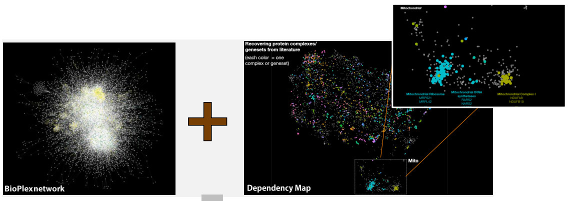
4. Manifold embedding (UMAP)

5. Map of gene-gene correlations

Meyers et al, Nat. Gen. 2018,
Behan et al, Nat. 2019

Josh Pan

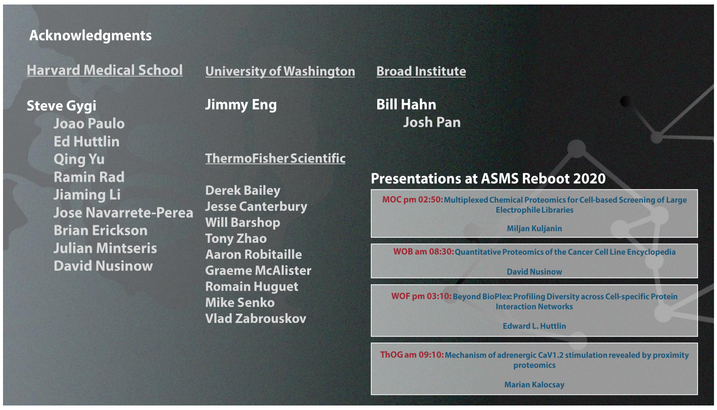
Manifold Embedding to determine latent cell states



Incorporating multiple large-scale analyses to identify novel complex members, pathways or allied proteins, particularly for uncharacterized proteins.

Other datasets:
CCLE Proteome (Nusinow et al *Cell* 2020)
Cell perturbation
OxiMouse (Xiao et al *Cell* 2020)

Josh Pan
Ed Huttlin



Acknowledgments

Harvard Medical School

Steve Gygi
Joao Paulo
Ed Huttlin
Qing Yu
Ramin Rad
Jiaming Li
Jose Navarrete-Perea
Brian Erickson
Julian Mintseris
David Nusinow

University of Washington

Jimmy Eng

ThermoFisher Scientific

Derek Bailey
Jesse Canterbury
Will Barshop
Tony Zhao
Aaron Robitaille
Graeme McAlister
Romain Huguet
Mike Senko
Vlad Zabrouskov

Broad Institute

Bill Hahn
Josh Pan

Presentations at ASMS Reboot 2020

MOC pm 02:50: Multiplexed Chemical Proteomics for Cell-based Screening of Large Electrophile Libraries

Miljan Kuljanin

WOB am 08:30: Quantitative Proteomics of the Cancer Cell Line Encyclopedia

David Nusinow

WOF pm 03:10: Beyond BioPlex: Profiling Diversity across Cell-specific Protein Interaction Networks

Edward L. Huttlin

ThOG am 09:10: Mechanism of adrenergic CaV1.2 stimulation revealed by proximity proteomics

Marian Kalocsay