Deep Learning for DIA Based Mass Spectrometry

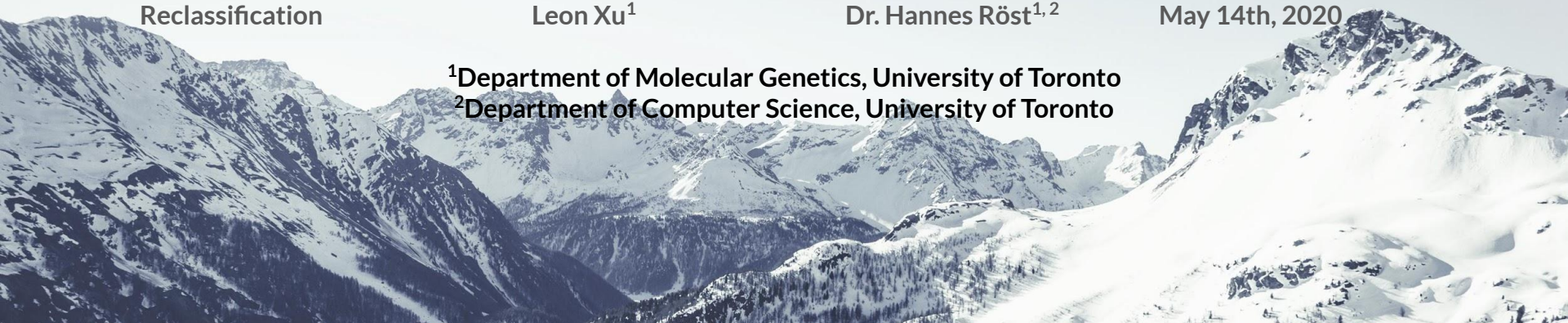
Reclassification

Leon Xu[1]

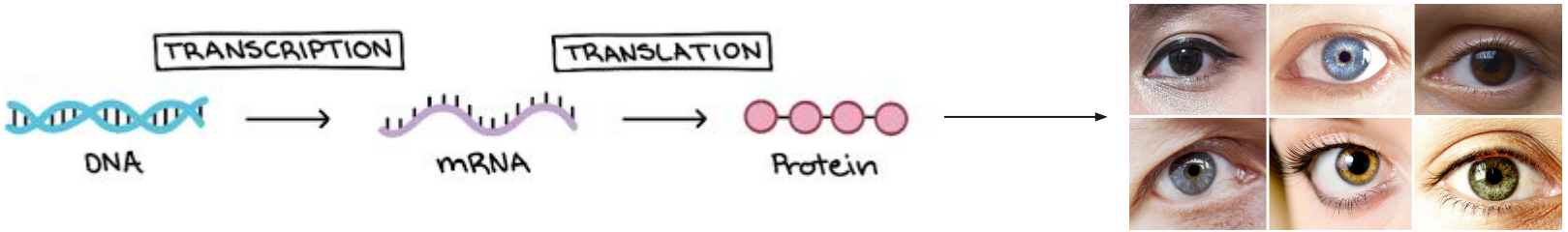
Dr. Hannes Röst[1, 2]

May 14th, 2020

[1]Department of Molecular Genetics, University of Toronto
[2]Department of Computer Science, University of Toronto



Proteins Provide A Window Towards Patient Phenotype



Phenotype

Relatively static → Changing, dynamic

Image from Khan Academy

Blood is Easily Accessible, Minimally Invasive, and Analyte Rich

- Blood is a key element involved in many biological processes
- We can get blood samples fairly easily, ideal as a “liquid biopsy”
- Already have some immunoassays for plasma proteins approved for clinical use
- Thus, plasma proteomics has a high potential for novel disease specific biomarker discovery

MS Allows For Better Characterization of the (Plasma) Proteome

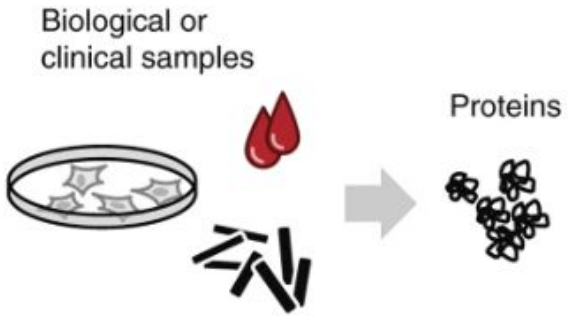


Image from Röst et al. Proteomics 289–307 (2017).

MS Allows For Better Characterization of the (Plasma) Proteome

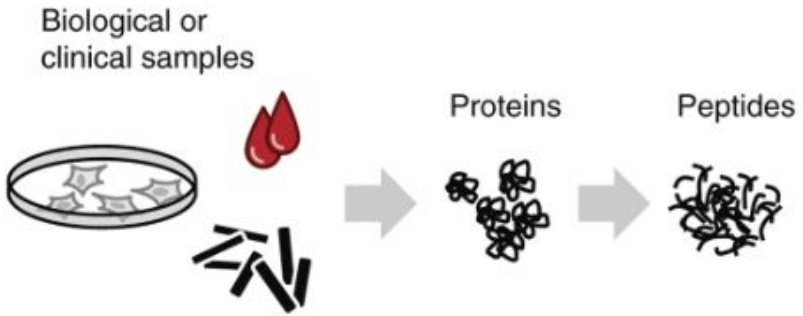


Image from Röst et al. Proteomics 289–307 (2017).

MS Allows For Better Characterization of the (Plasma) Proteome

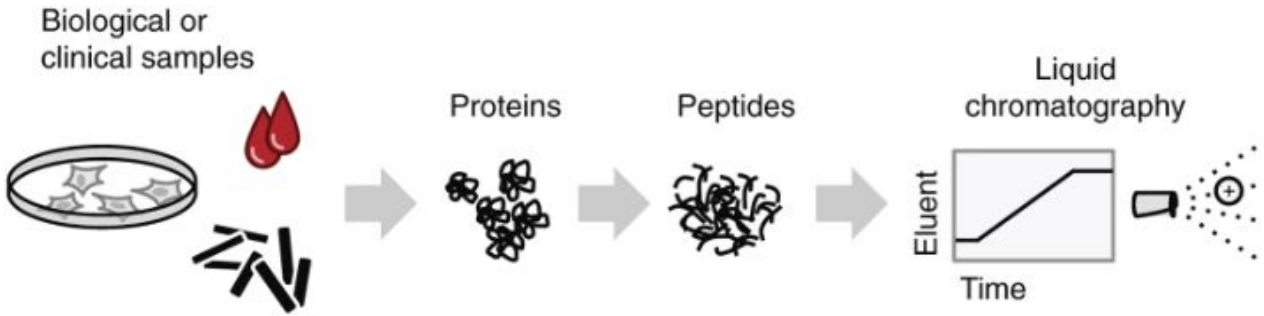


Image from Röst et al. Proteomics 289–307 (2017).

MS Allows For Better Characterization of the (Plasma) Proteome

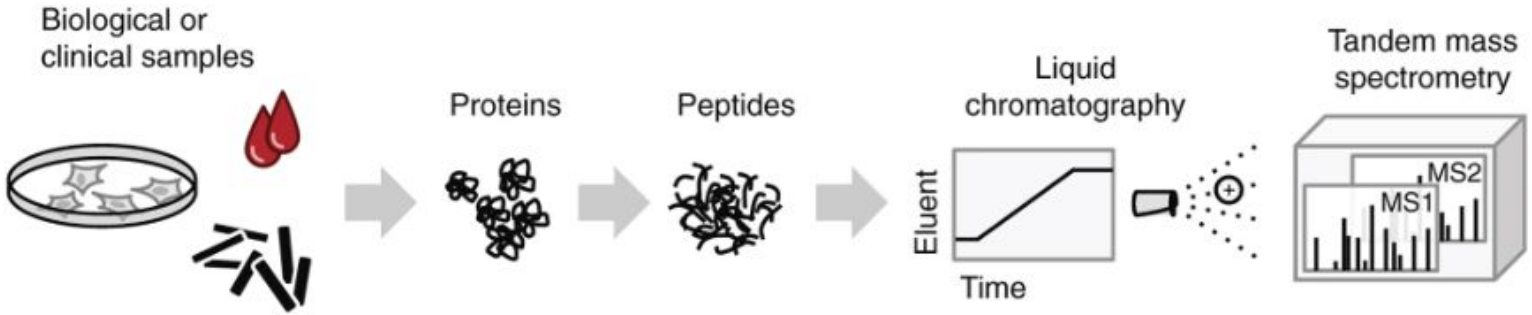
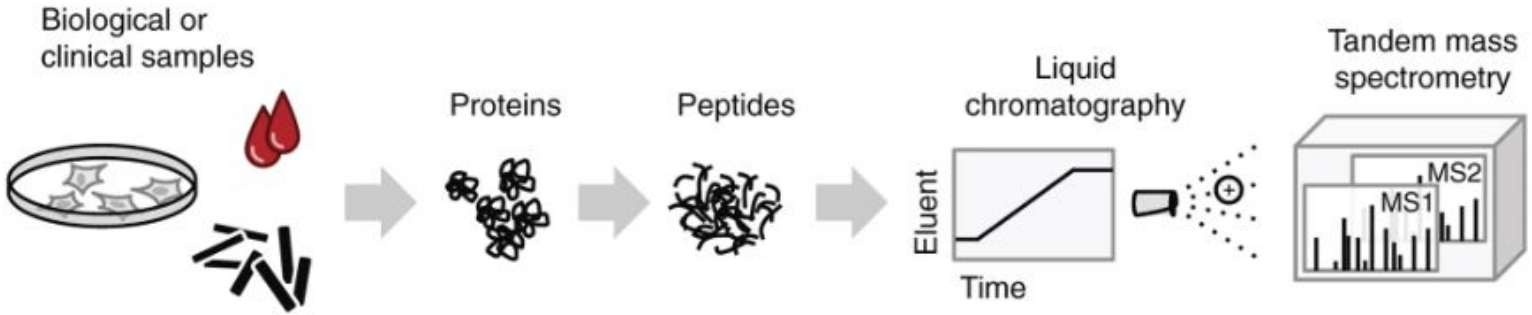


Image from Röst et al. Proteomics 289–307 (2017).

MS Allows For Better Characterization of the (Plasma) Proteome

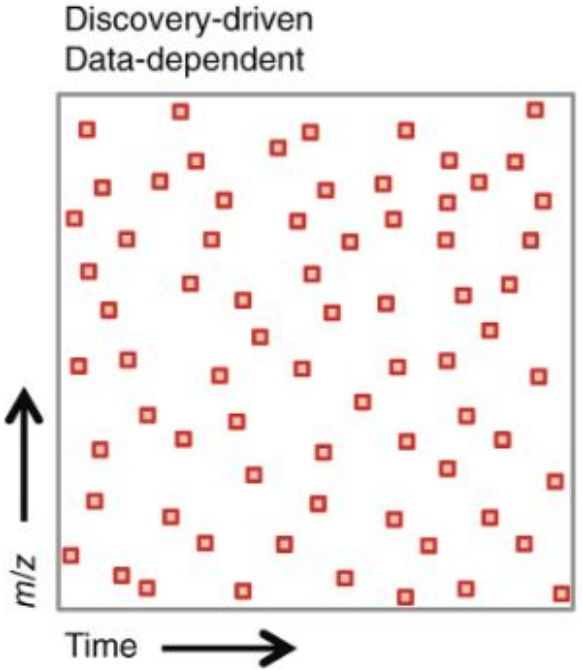


- Can be more analytically sensitive and specific than immunoassays
- Can provide a more direct and unbiased measurement
- Can process highly multiplexed samples

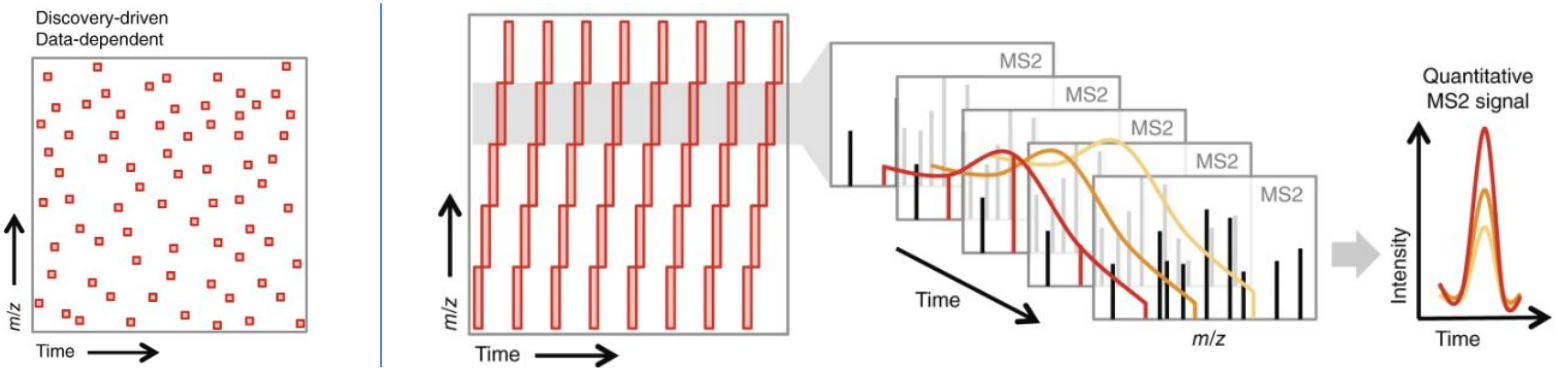
Image from Röst et al. Proteomics 289–307 (2017).

Many Challenges Remain Before Ready For Regular Clinical Use

- Large and dynamic range of expression levels
- High abundance proteins
- Improvements to sample processing has allowed for a greater depth of coverage, but results in lower throughput and lower reproducibility



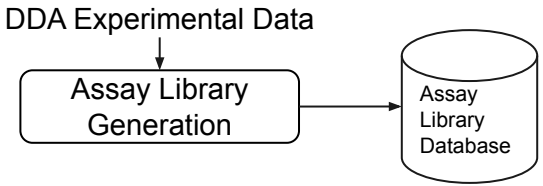
DIA-MS: Reproducibly Capture Proteome in High Throughput



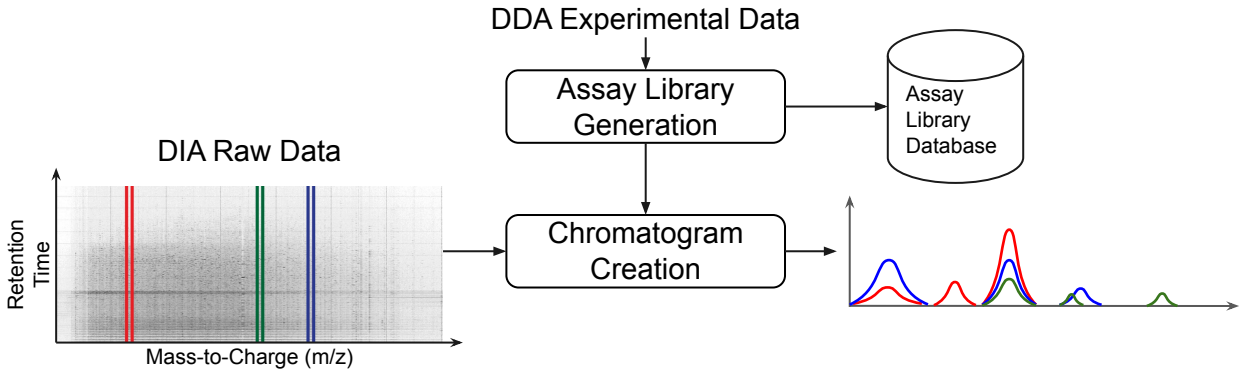
- High throughput!
- Highly reproducible and quantitative!
- Comprehensive record for future analysis!
- But also… Highly multiplexed and highly complex data! Requires sophisticated software analysis tools to deconvolute.

Image from Röst et al. Proteomics 289–307 (2017).

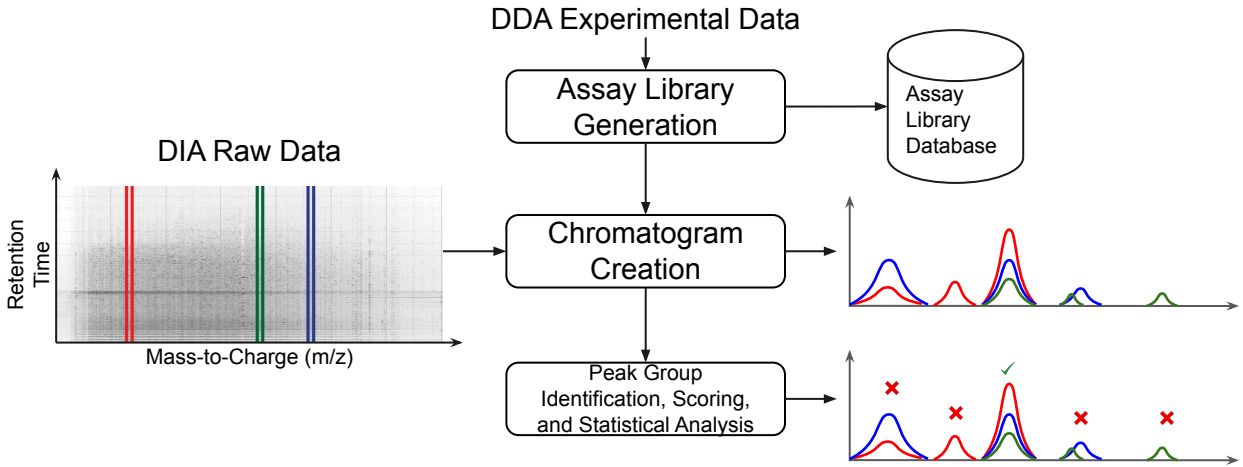
Current Targeted DIA Analysis Pipeline



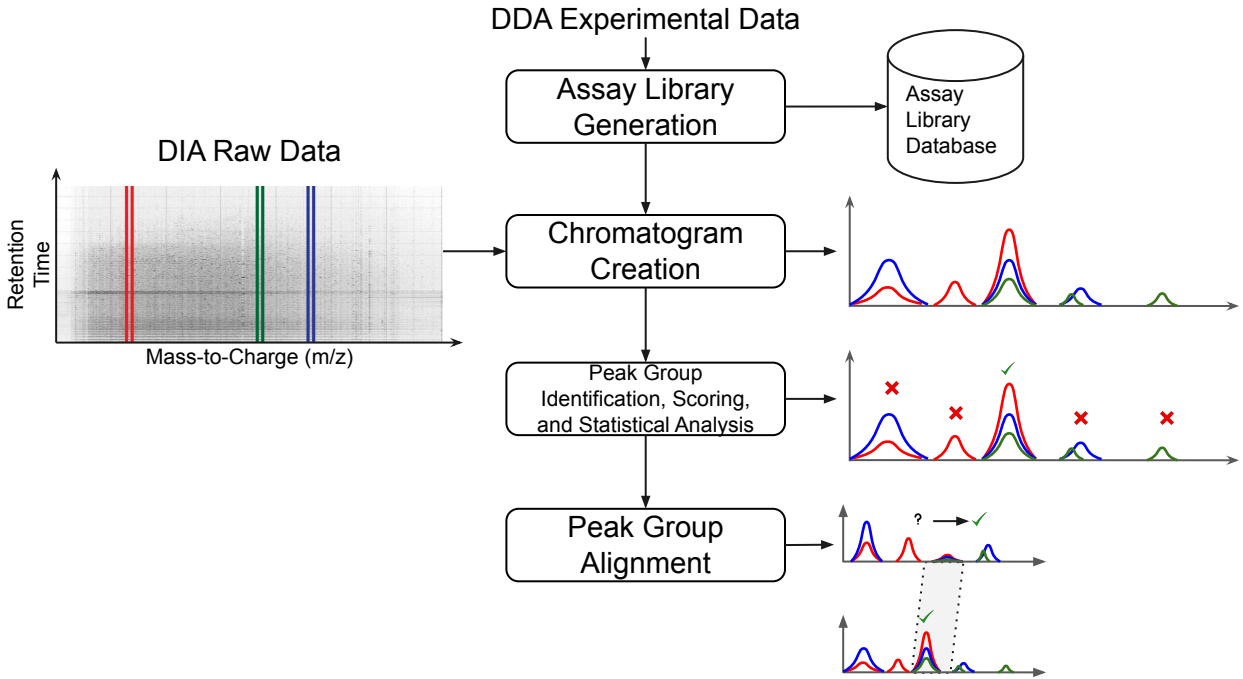
Current Targeted DIA Analysis Pipeline



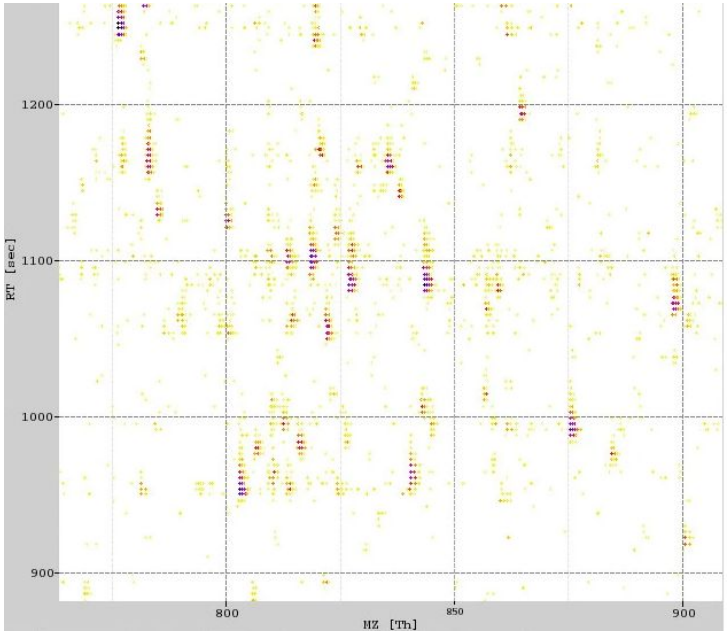
Current Targeted DIA Analysis Pipeline



Current Targeted DIA Analysis Pipeline



Increasing Data Complexity Due To More Sophisticated Methods



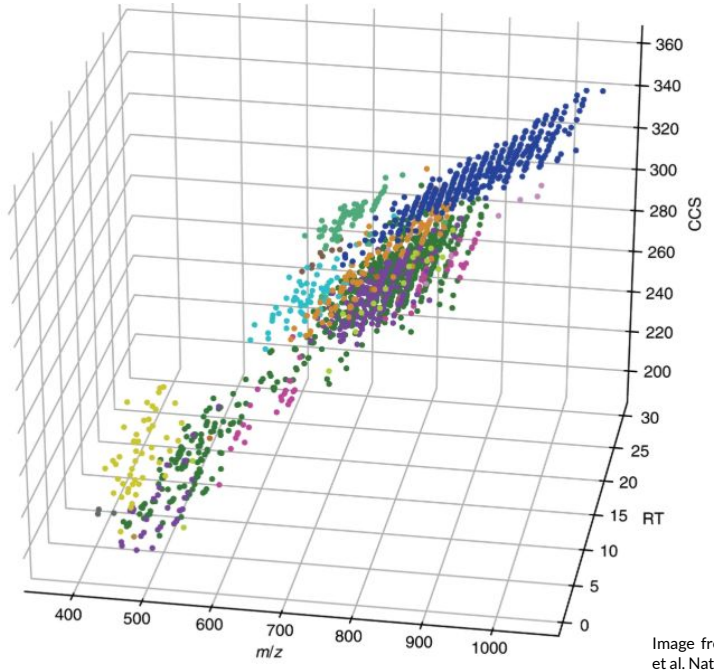
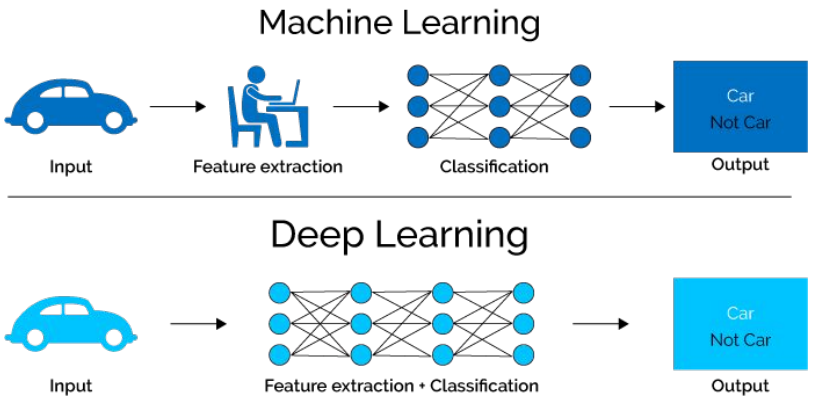
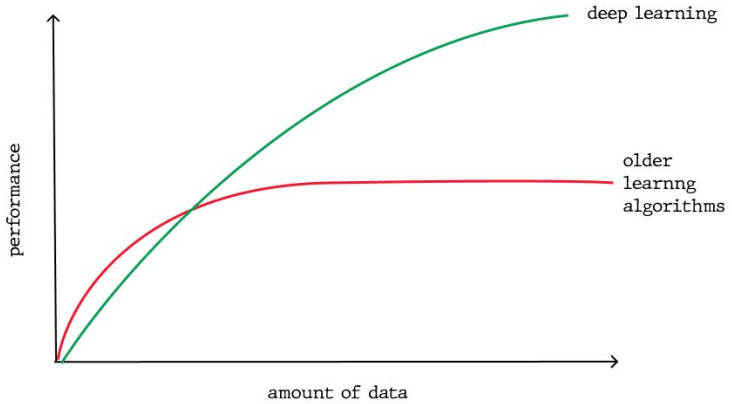


Image from Vasilopoulou et al. Nat. Comm. 331 (2020).

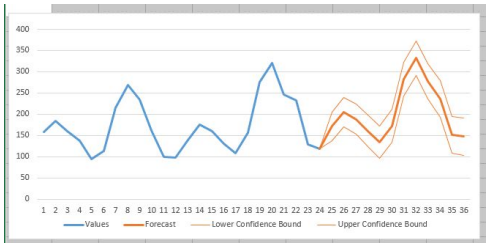
Deep Learning Is End to End and Extremely Scalable

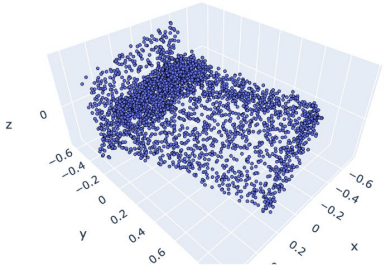




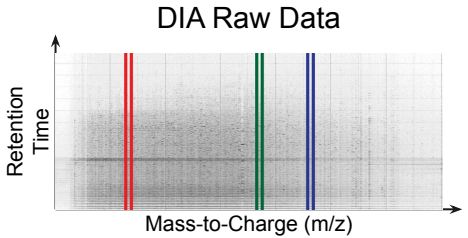
Images from towardsdatascience

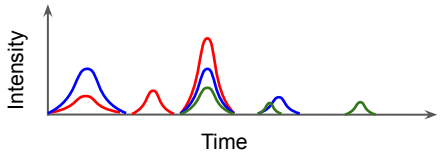
Applications of Deep Learning With Analogues In MS

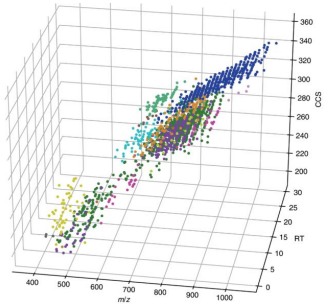




WORDS







PEPTIDES

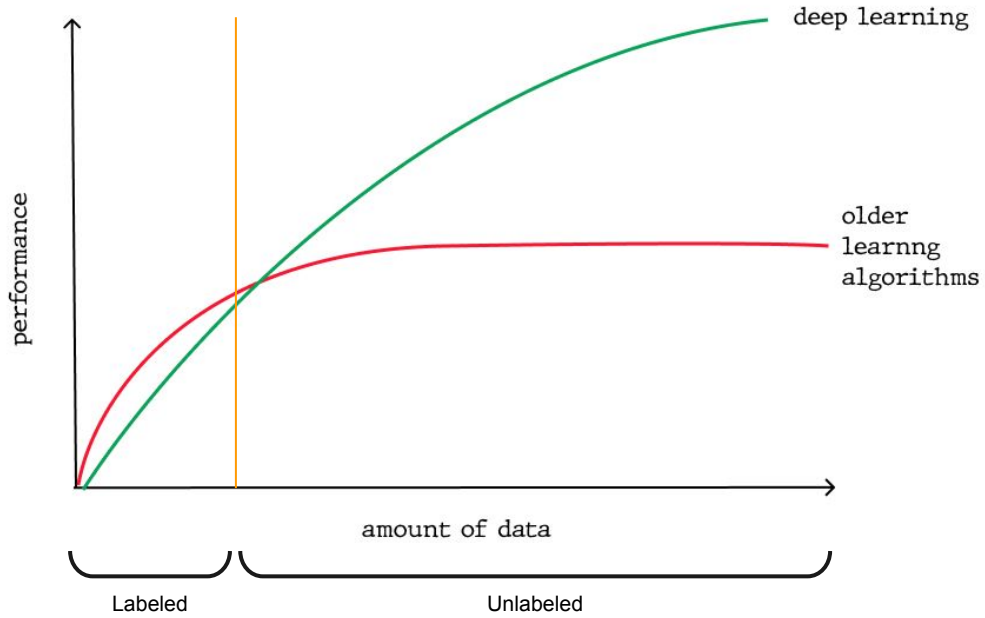
Aims

1. Semi-supervised identification of chromatographic Regions of Interest (ROI) for ***targeted*** DIA data analysis in a data-driven and scalable manner
2. Integration of information up- and downstream of current ROI identification process
3. Evaluation of method on a complex dataset: Detection of biomarkers involved in Type-II diabetes progression from plasma

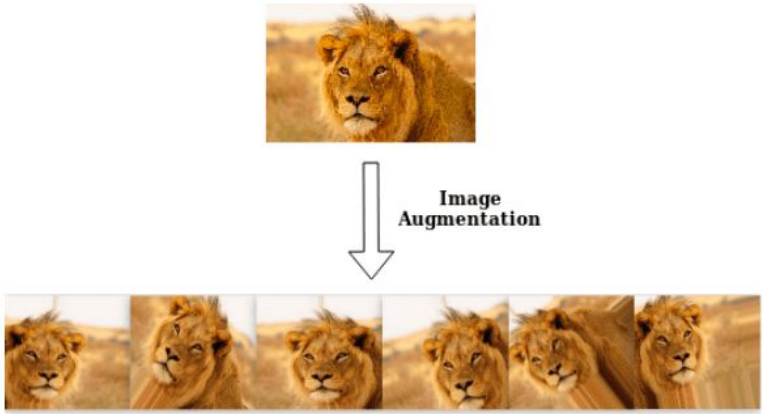
Aims

1. Semi-supervised identification of chromatographic Regions of Interest (ROI) for ***targeted*** DIA data analysis in a data-driven and scalable manner
2. Integration of information up- and downstream of current ROI identification process
3. Evaluation of method on a complex dataset: Detection of biomarkers involved in Type-II diabetes progression from plasma

DNNs Are Data Hungry, But Most Available Data is Unlabeled

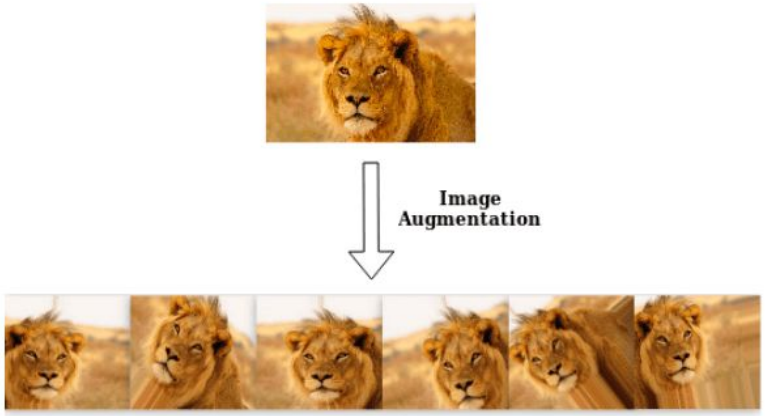


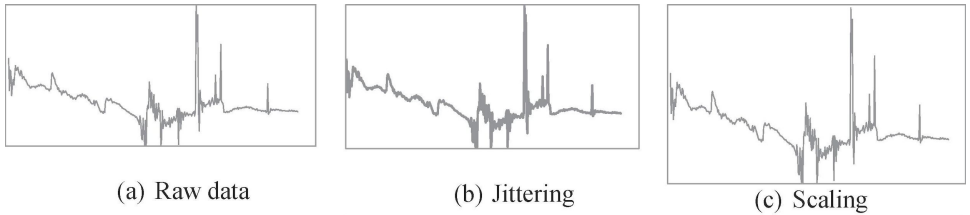
Data Augmentations Used For Consistency Regularization

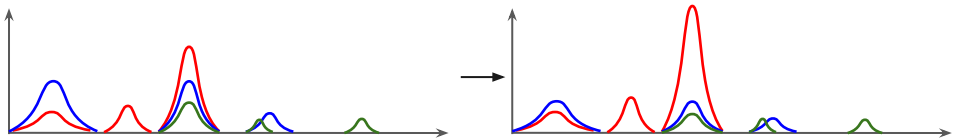


Images from towardsdatascience

Data Augmentations Used For Consistency Regularization



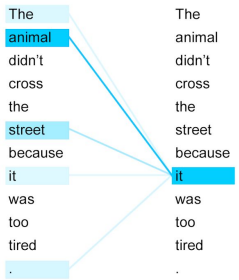


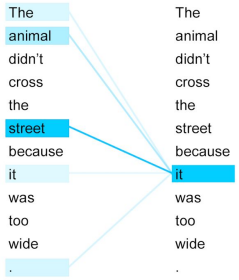


Scaling Chromatograms

Images from towardsdatascience, Rashid and Louis (2019)

To Train A Model Based On the Transformer Architecture





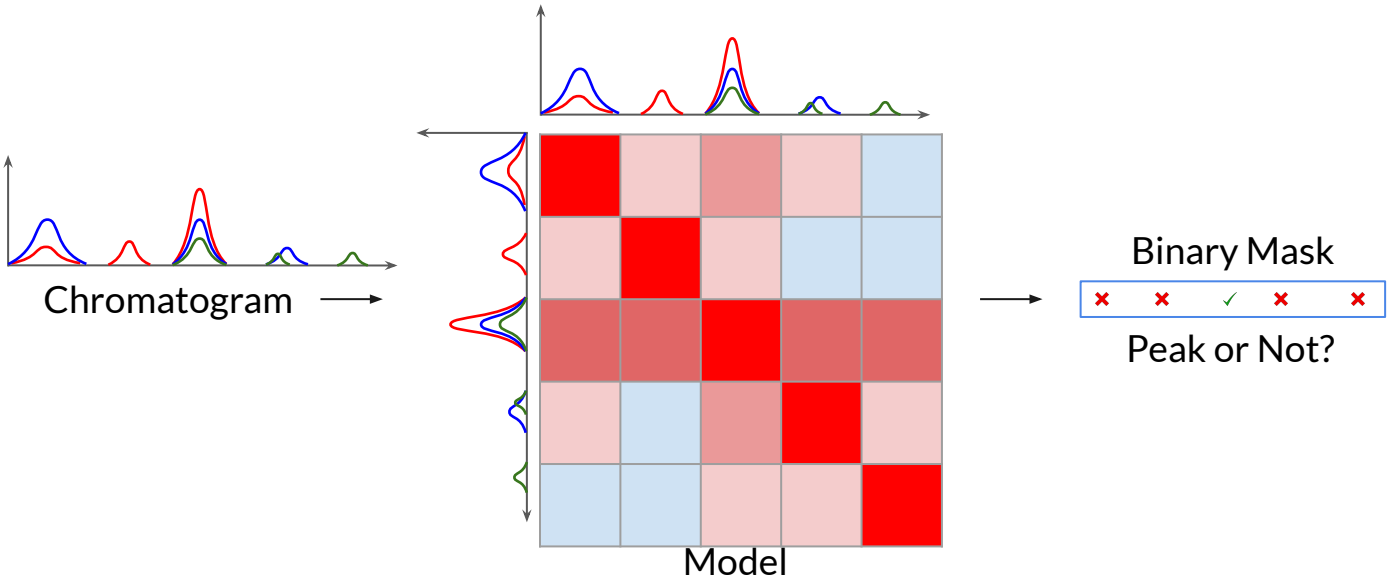
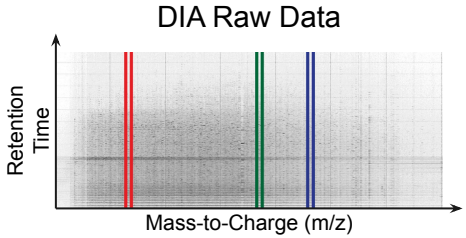
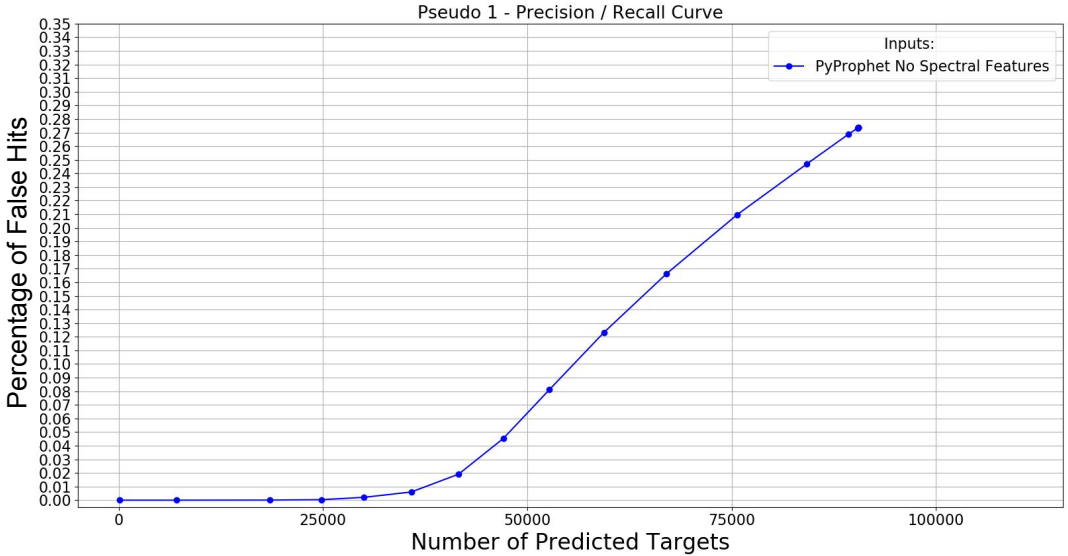


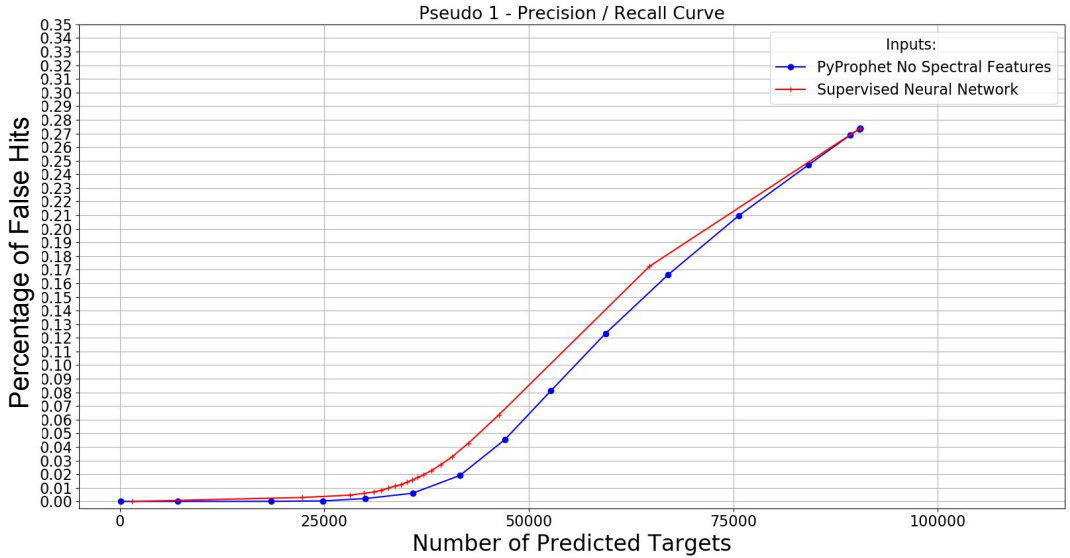
Image from
https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

Neural Network Matches State of the Art Performance

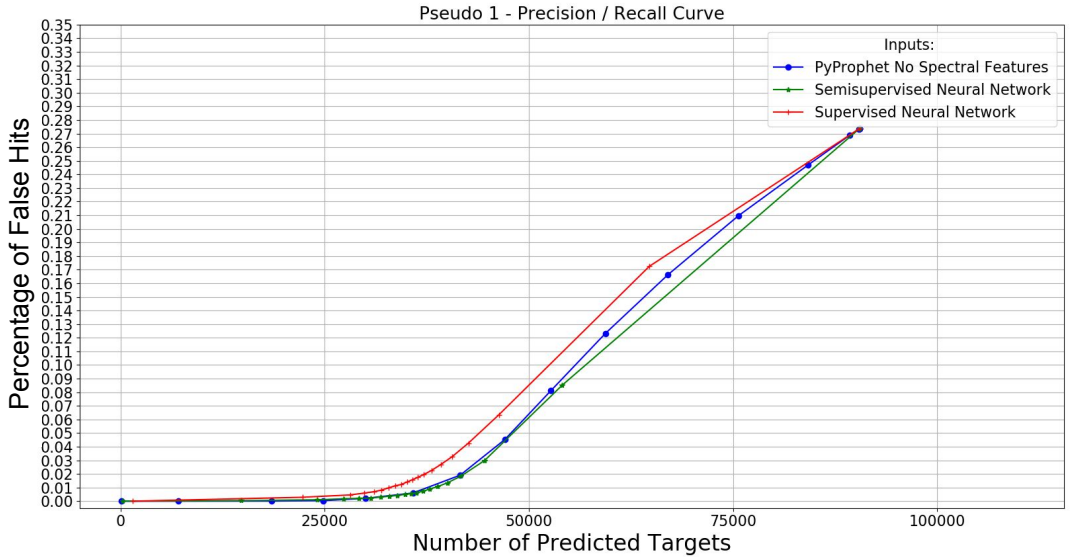




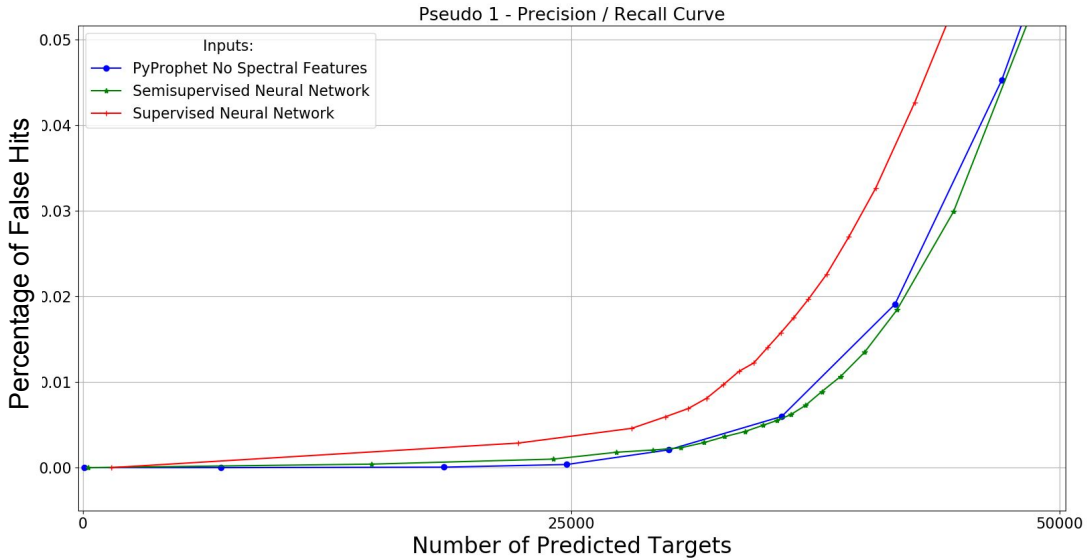
Neural Network Matches State of the Art Performance



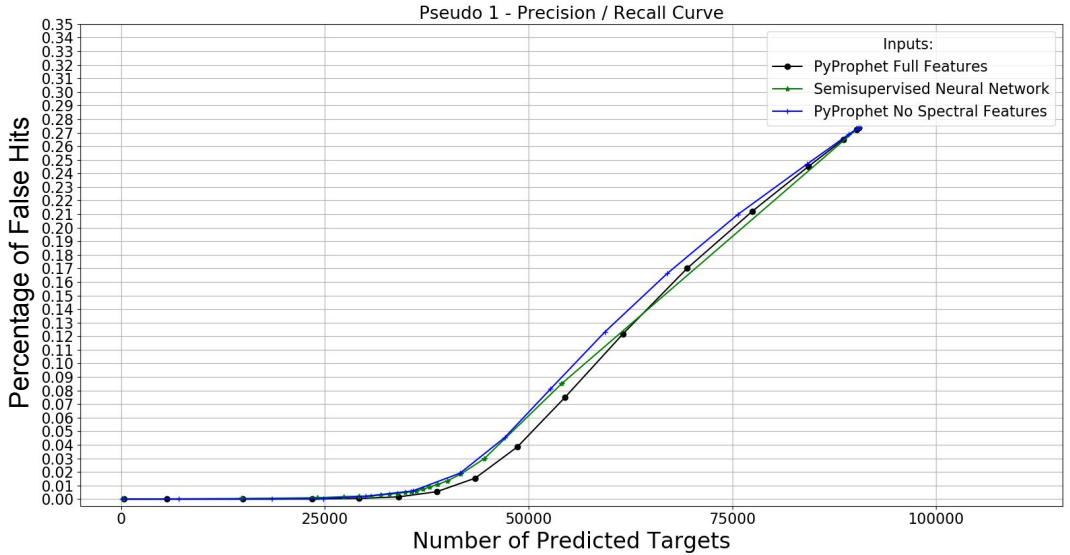
Neural Network Matches State of the Art Performance



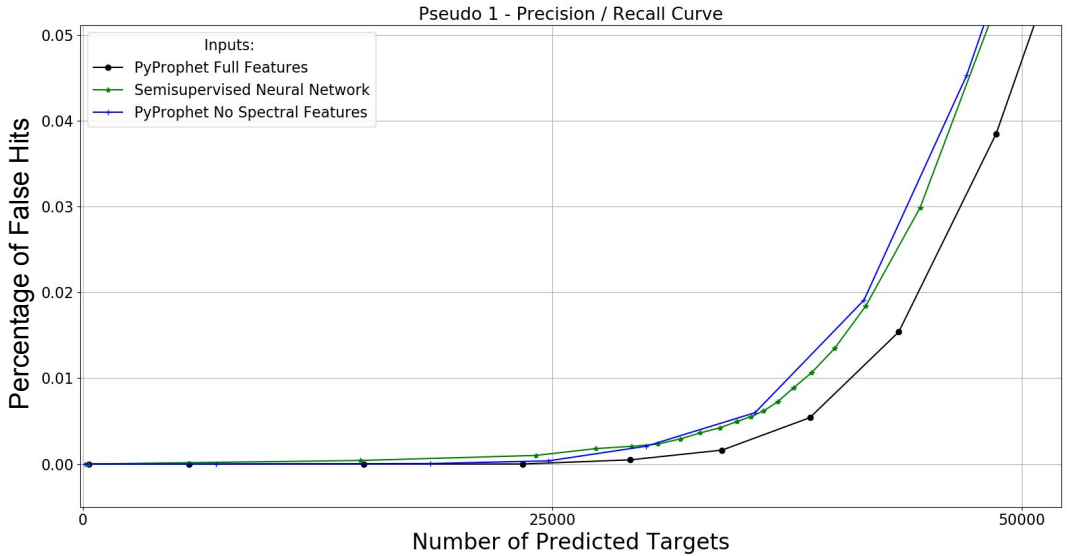
Neural Network Matches State of the Art Performance



Neural Network Matches State of the Art Performance



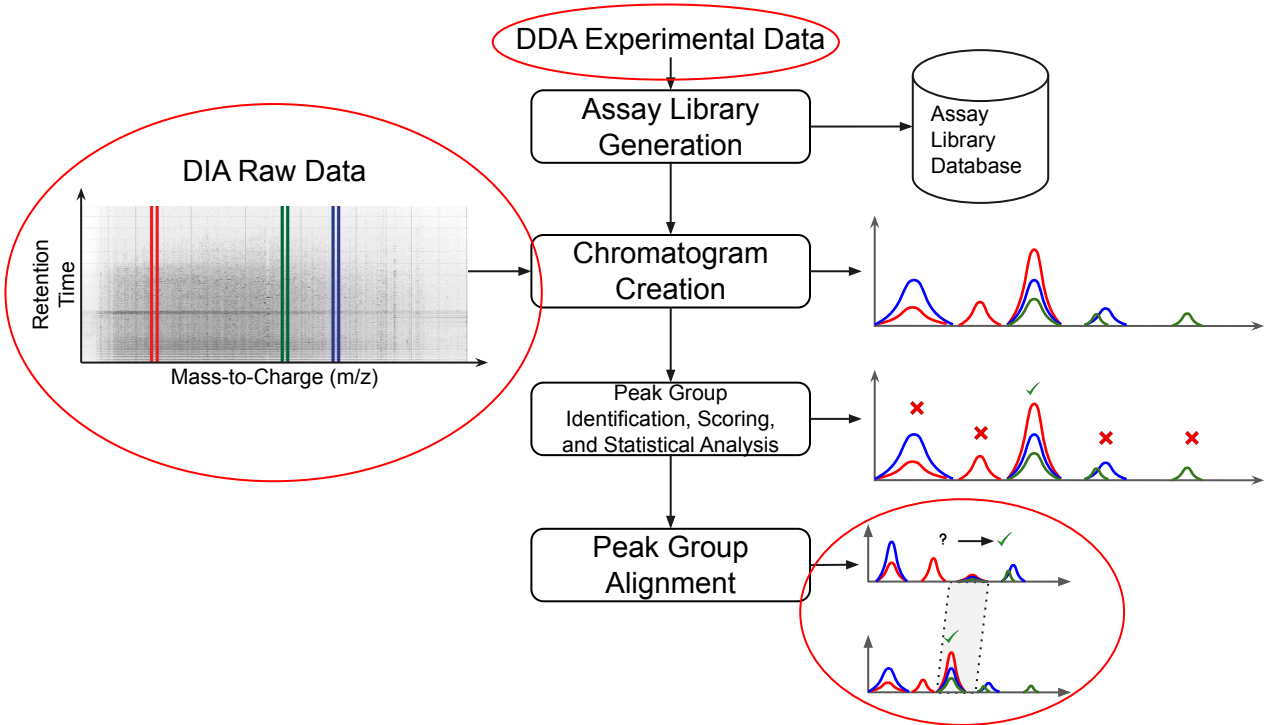
Neural Network Matches State of the Art Performance



Aims

1. Semi-supervised identification of chromatographic Regions of Interest (ROI) for ***targeted*** DIA data analysis in a data-driven and scalable manner
2. Integration of information up- and downstream of current ROI identification process
3. Evaluation of method on a complex dataset: Detection of biomarkers involved in Type-II diabetes progression from plasma

We Are Not Yet Taking Full Advantage of the Data… Why Not…



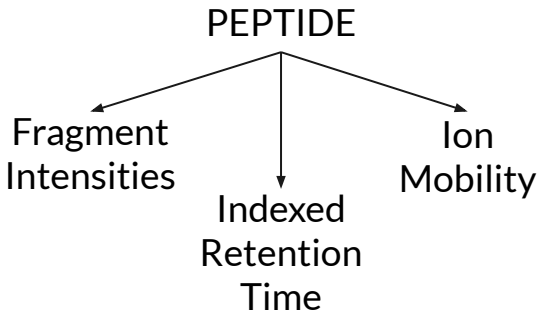
Generate Assay Libraries In Silico

nature methods

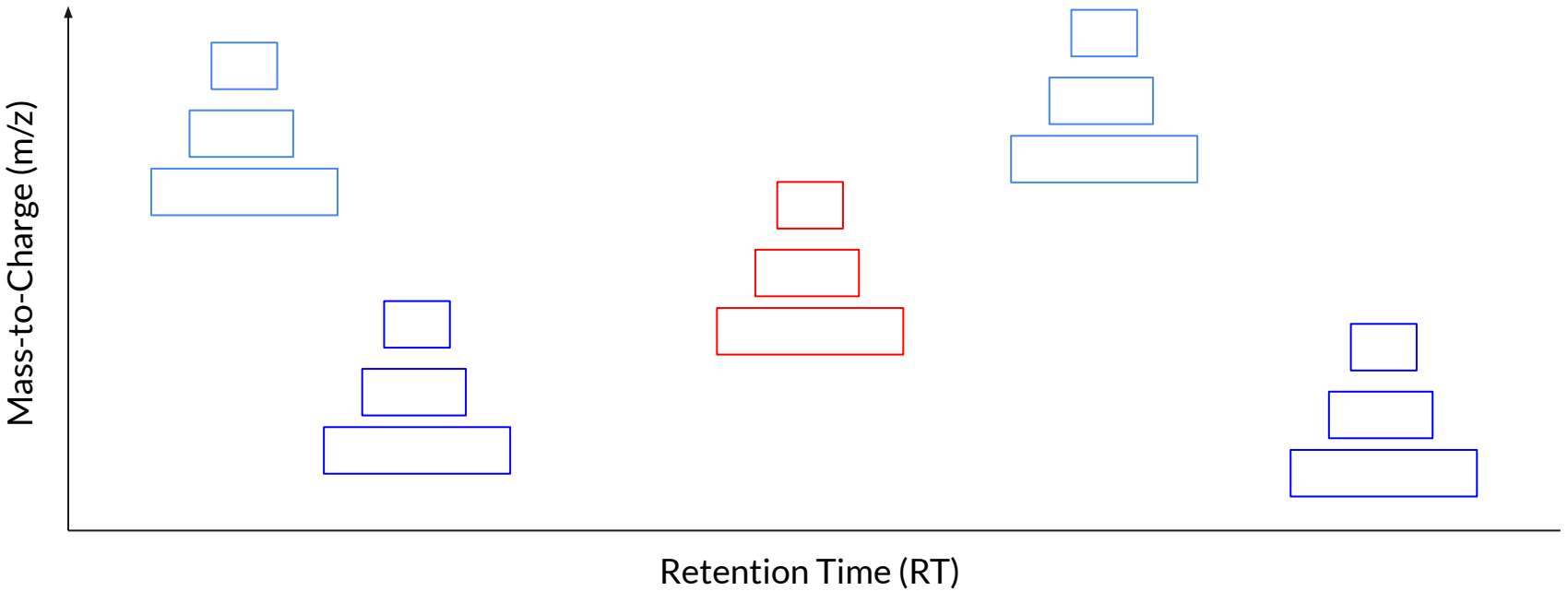
Article | Published: 27 May 2019

Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning

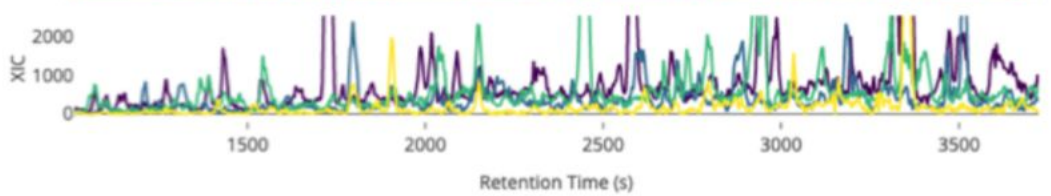
Siegfried Gessulat, Tobias Schmidt, Daniel Paul Zolg, Patroklos Samaras, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Julia Rechenberger, Bernard Delanghe, Andreas Huhmer, Ulf Reimer, Hans-Christian Ehrlich, Stephan Aiche, Bernhard Kuster & Mathias Wilhelm

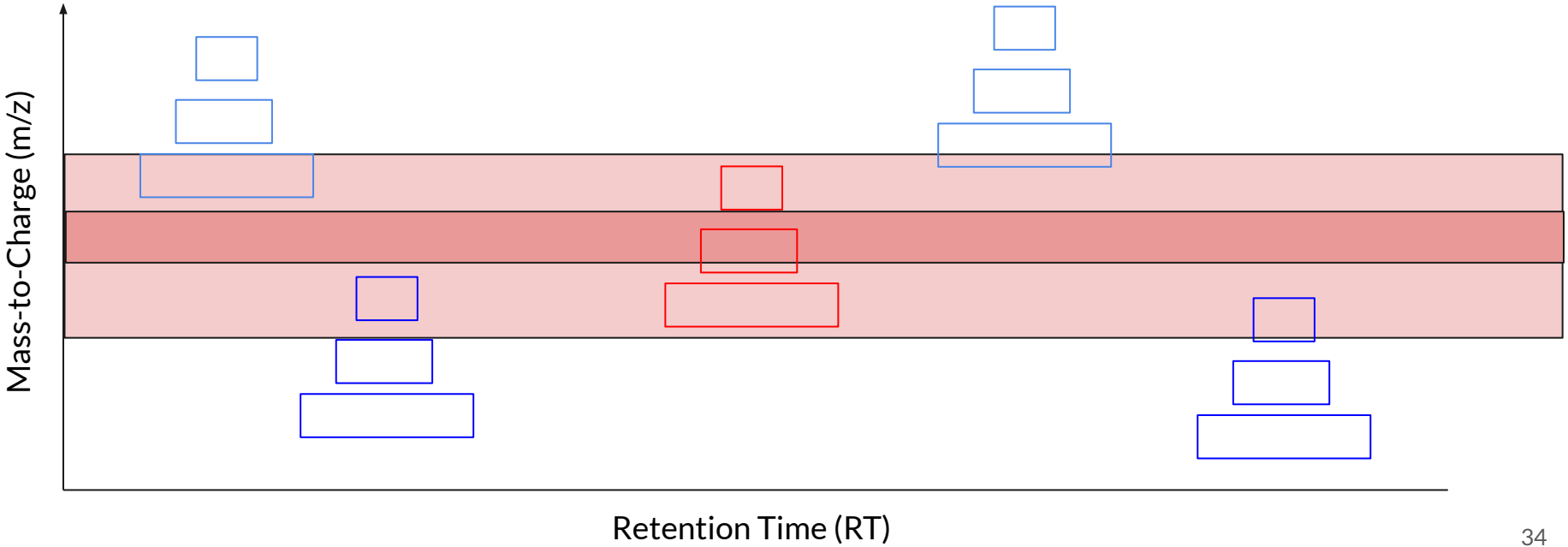


Encode Raw Spectra Maps

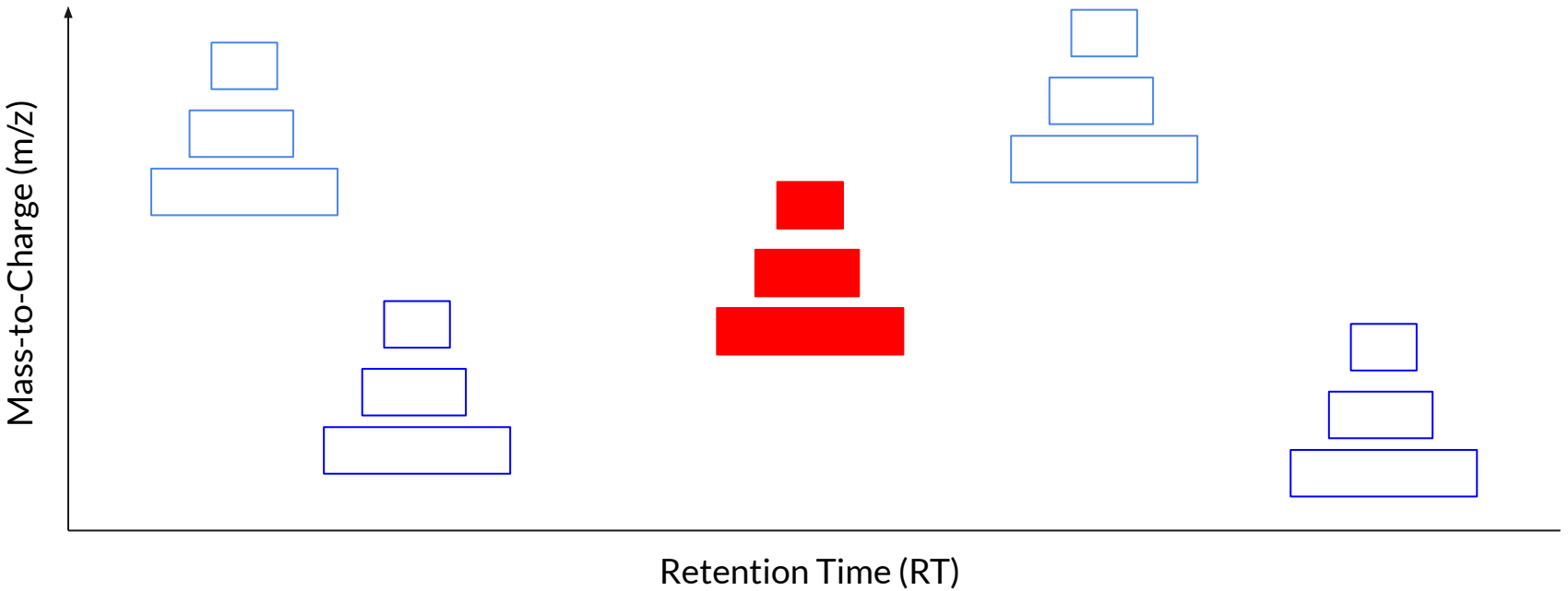


Encode Raw Spectra Maps

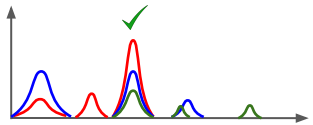




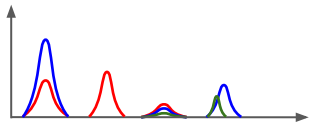
Encode Raw Spectra Maps (And Other Inputs Too)



Include Priors/Learn Alignment Function Across Runs

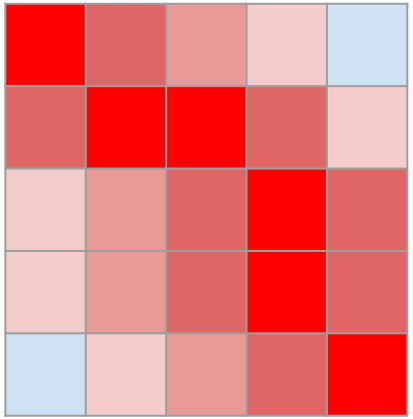


Input Template(s) (and Label(s))

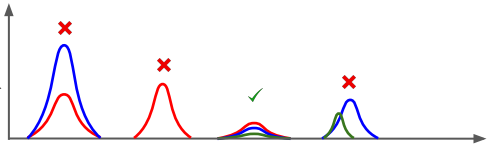


Input Query

Peak Group Deep Learning Model

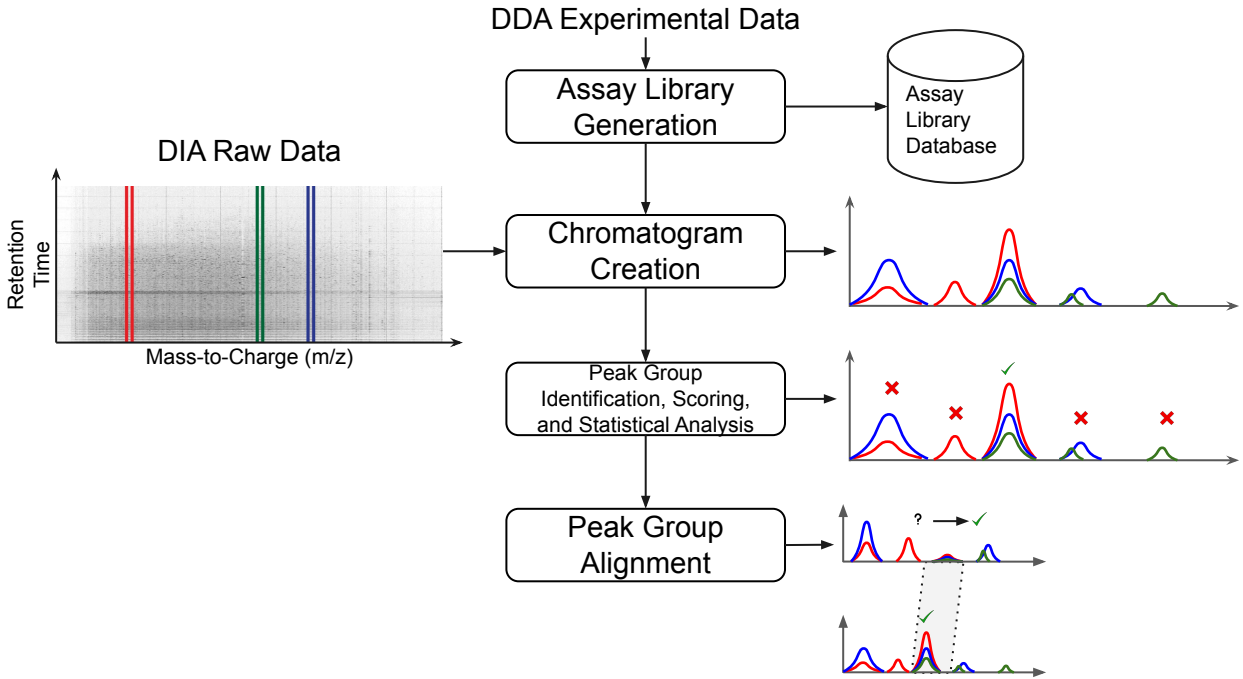


Learned “Alignment” Attention Matrix

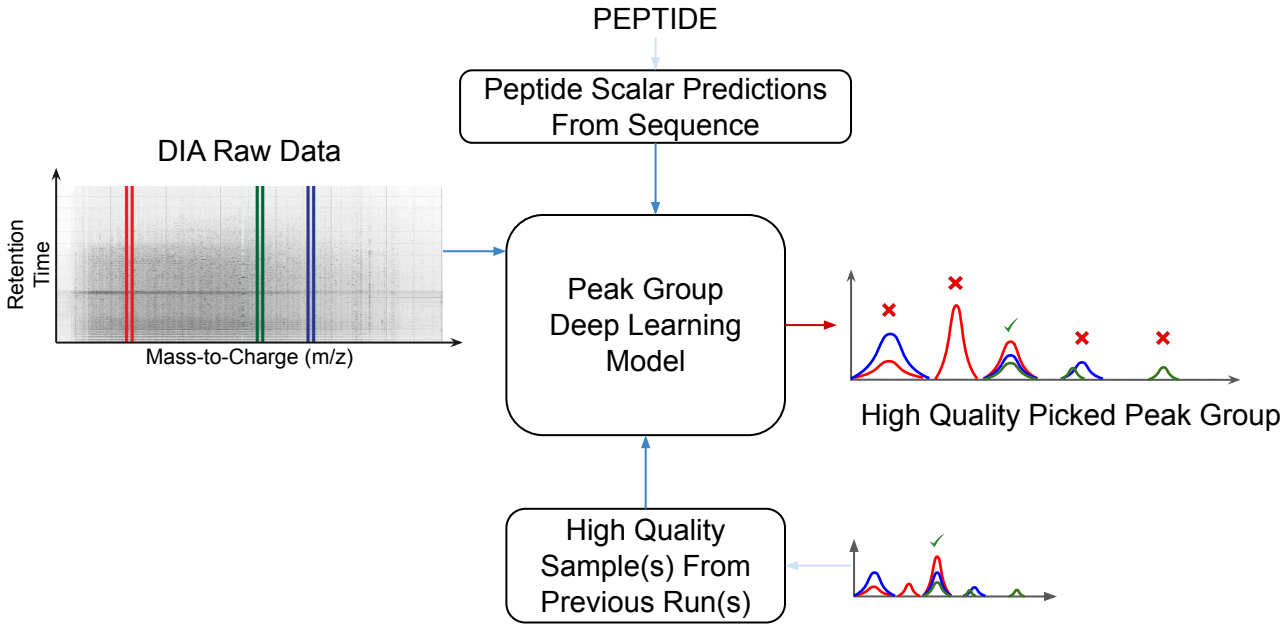


High Quality Picked Peak Group

From The Current Sequential Targeted DIA Analysis Pipeline



To An Integrated DIA Analysis Workflow



Aims

1. Semi-supervised identification of chromatographic Regions of Interest (ROI) for ***targeted*** DIA data analysis in a data-driven and scalable manner
2. Integration of information up- and downstream of current ROI identification process
3. Evaluation of method on a complex dataset: Detection of biomarkers involved in Type-II diabetes progression from plasma

Effectiveness on Difficult Longitudinal Plasma-Proteomics Dataset

Received initial data from Snyder Lab.

- 1044 blood samples collected over 4 years
- 105 individuals:
  - 55 female
  - 50 male
  - Age: 25-75 years
- Steady State Plasma Glucose used to measure insulin sensitivity
  - 32 insulin sensitive individuals
  - 30 insulin resistant individuals
  - 43 unknown
- Many issues (e.g. change of column) leading to difficult to analyze data
- Only 333 proteins quantified with reasonable coverage (detected in 67% of runs)
- **Can we do better?**

Summary

1. Semi-supervised identification of chromatographic Regions of Interest (ROI) for ***targeted*** DIA data analysis in a data-driven and scalable manner
2. Integration of information up- and downstream of current ROI identification process
3. Evaluation of method on a complex dataset: Detection of biomarkers involved in Type-II diabetes progression from plasma

Acknowledgements

Röst Lab
Dr. Hannes Röst
Annie Ha
Shubham Gupta
Justin Sing
Premy Shanthamoorthy

Röst Lab
Adamo Young
Dr. Mahmoud Ghaznavi
Emily Franklin
Audrina Zhou
Dr. Olga Zaslaver
Ron Blutrich

Röst Lab
Jimmy Chen
Charlotte Cai
Arshia Mahmoodi
Khaled Elemam
Jianyun Pan

Committee
Dr. Anne-Claude Gingras
Dr. Quaid Morris

External
Dr. Gary Bader
Dr. Frank Sicheri
Dr. Fritz Roth

Thank you.