STANFORD
UNIVERSITY

Multi-Instrument, Skyline-based Comparison of DIA Peptide Detection and Statistical Confidence Tools

Dario Amodei

Postdoctoral Scholar, Skyline Developer

Mallick Lab

Stanford University

6/15/2014

Acquisition Methods

Targeted

Selected Reaction Monitoring (SRM)

Few proteins
Good reproducibility
Good sensitivity

Discovery

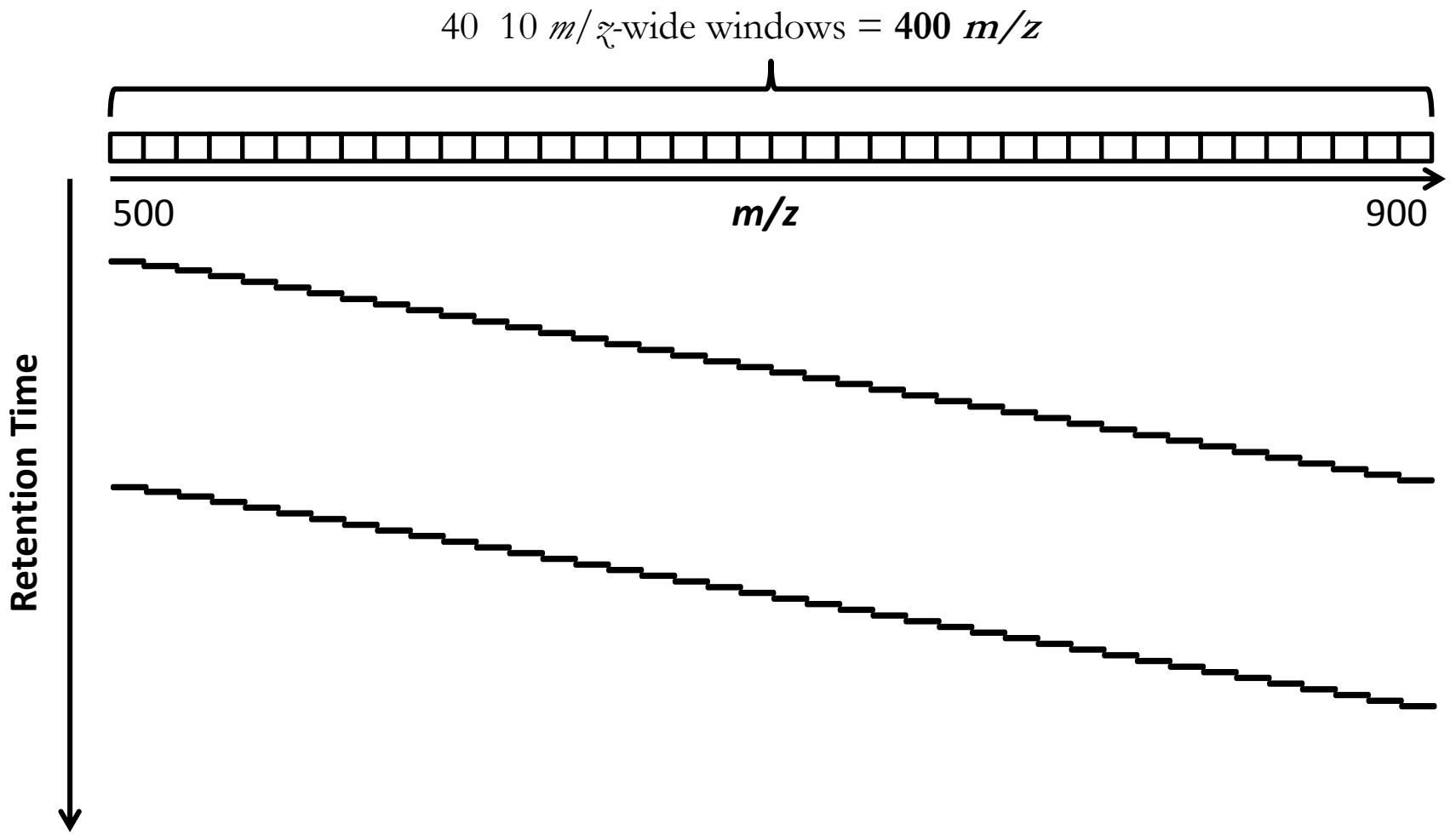
Data Dependent Acquisition (DDA)

Many proteins
Poor reproducibility
Poor sensitivity

Data Independent Acquisition (DIA)

e.g. SWATH-MS
"Best of both worlds?"

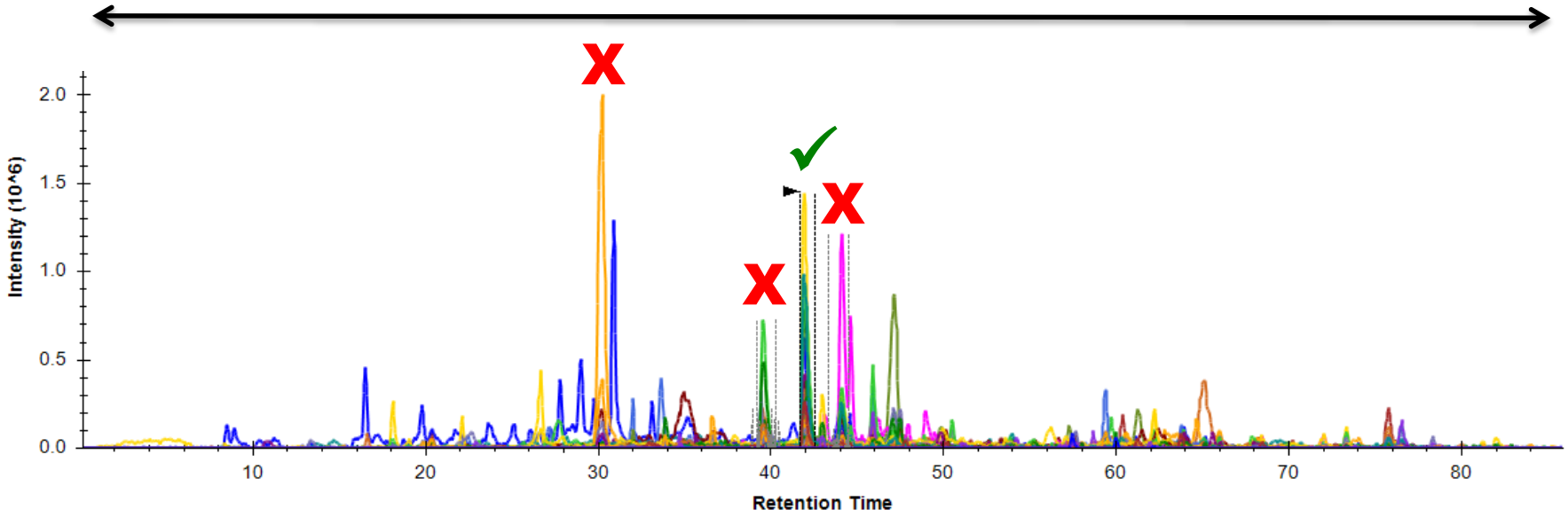
Data Independent Acquisition (DIA)



Peptide Detection in DIA Data

FEIELLSLDDDSIVNHEQDLPK *S. cerevisiae* lysate (soluble) 10 *m/z* wide window DIA (Q-Exactive)

Data is full chromatogram + can have lots of interferences



Which of these peaks is the actual peptide?

How sure are we?

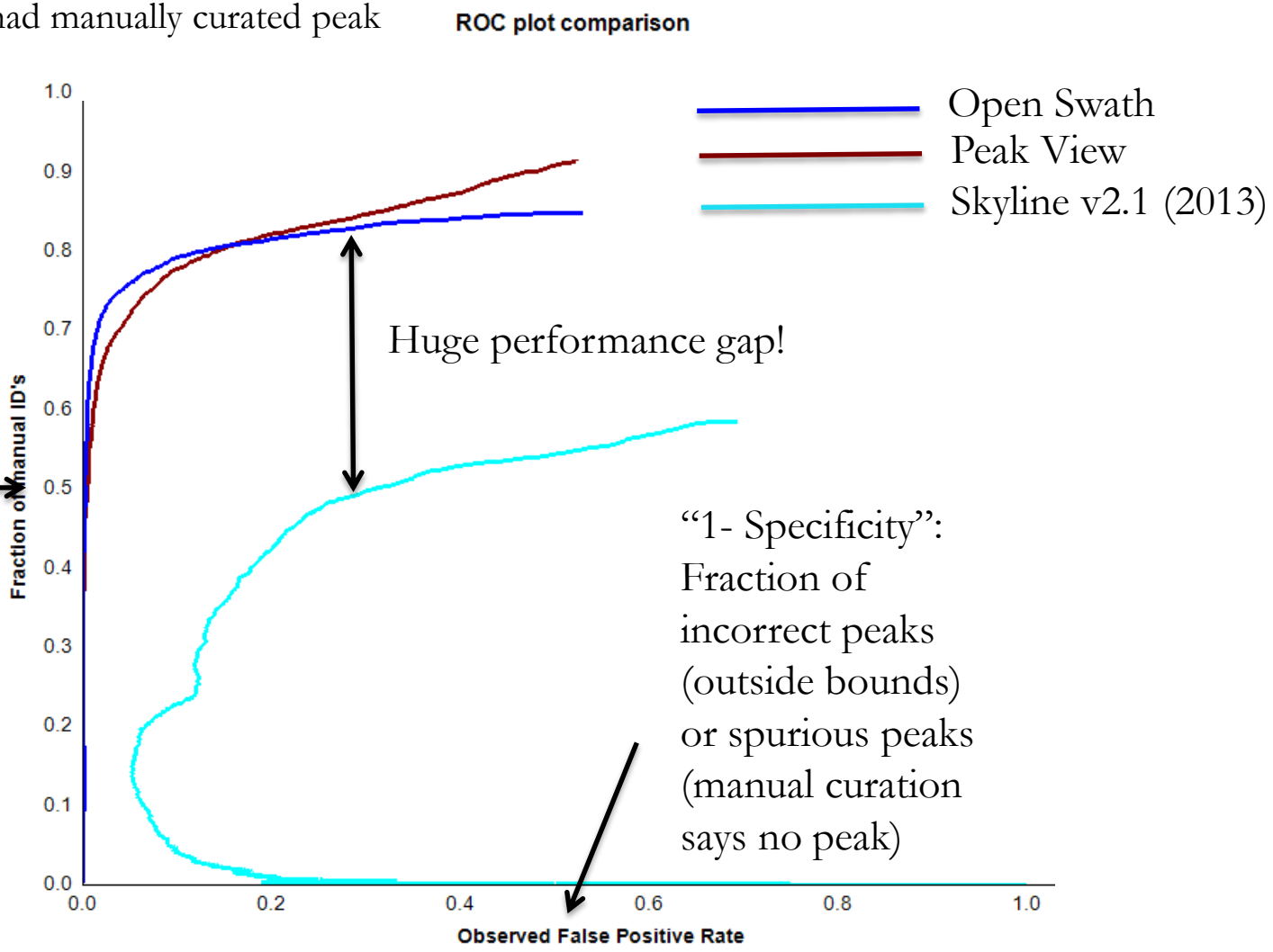
1000's of peptides to look at, need to automate the process

How do I process my DIA data?

- **My experience as a user (circa 2013)**:
  - Skyline could import DIA data but not good at peptide detection (peak picking), was still relatively SRM-geared
  - Other tools (Open Swath, Spectronaut, Peak View) were stronger on peak picking, but I preferred Skyline's rich environment / features / usability
- **Task (Brendan, Dario, Don Marsh)**:
  - Can we bring good peak picking to Skyline, using insights from these other tools as needed?

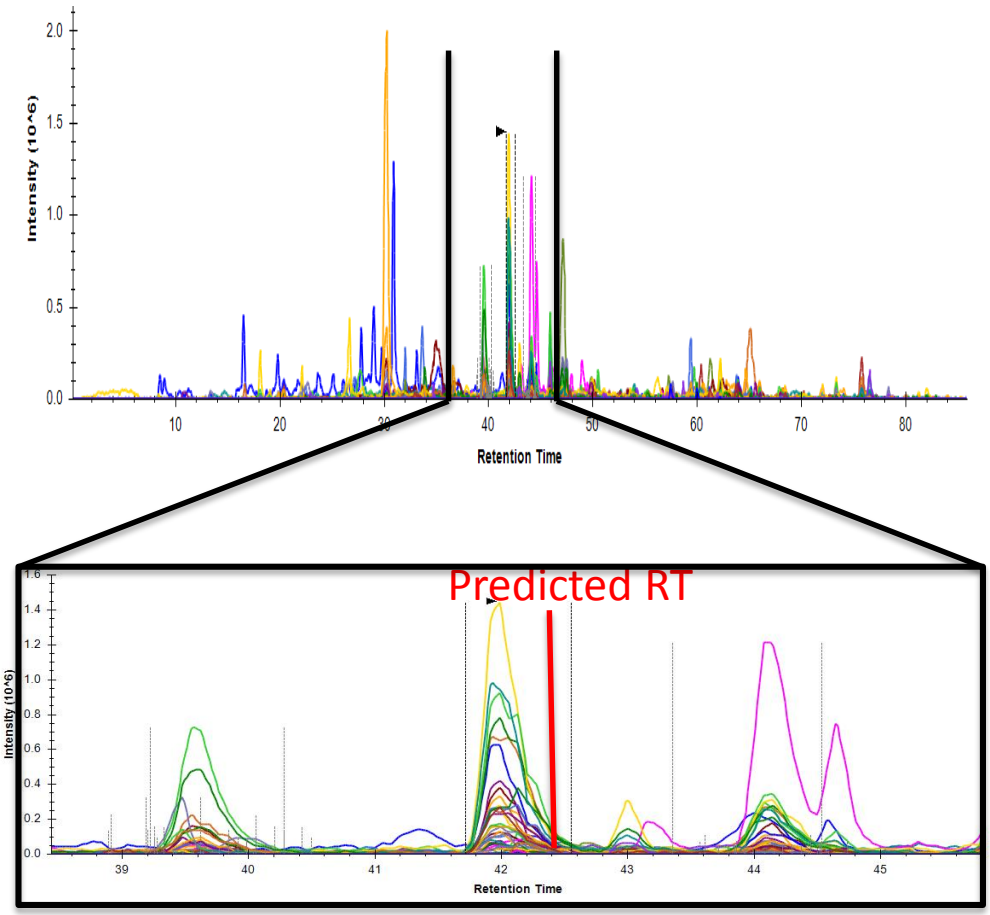
DIA in Skyline (circa mid-2013)

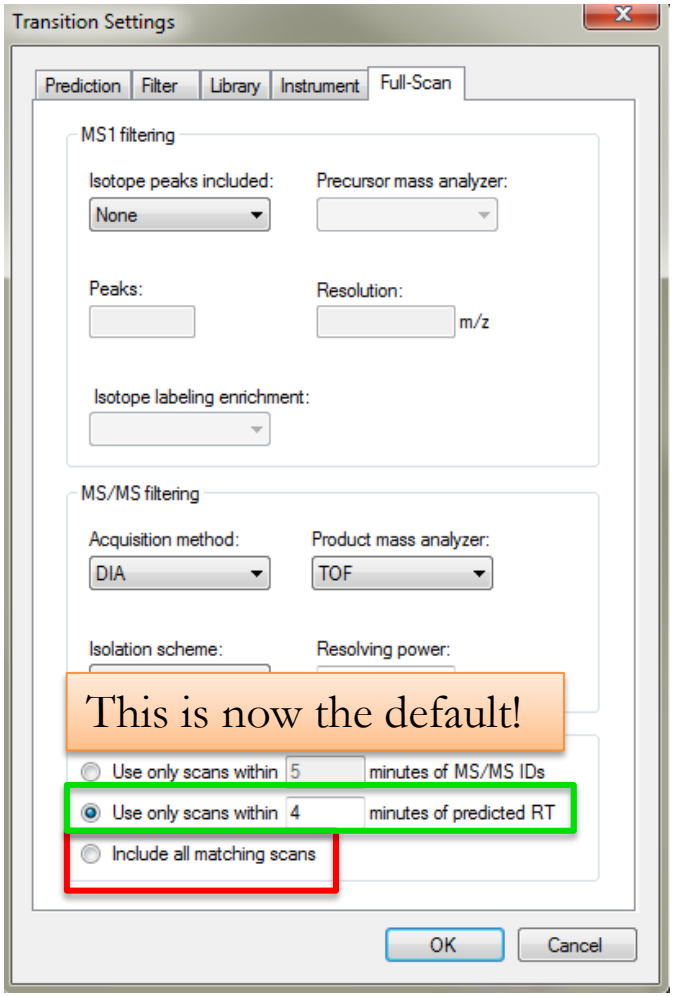
Dataset from Rost et al, *Nat. Biotech*, 2014 -- 344 heavy peptides spiked-in human background
10 dilution points x 3 replicates -- ABSCIEX 5600 SWATH 25 m/z windows
10350 total peptide-runs
80% of peptides had manually curated peak



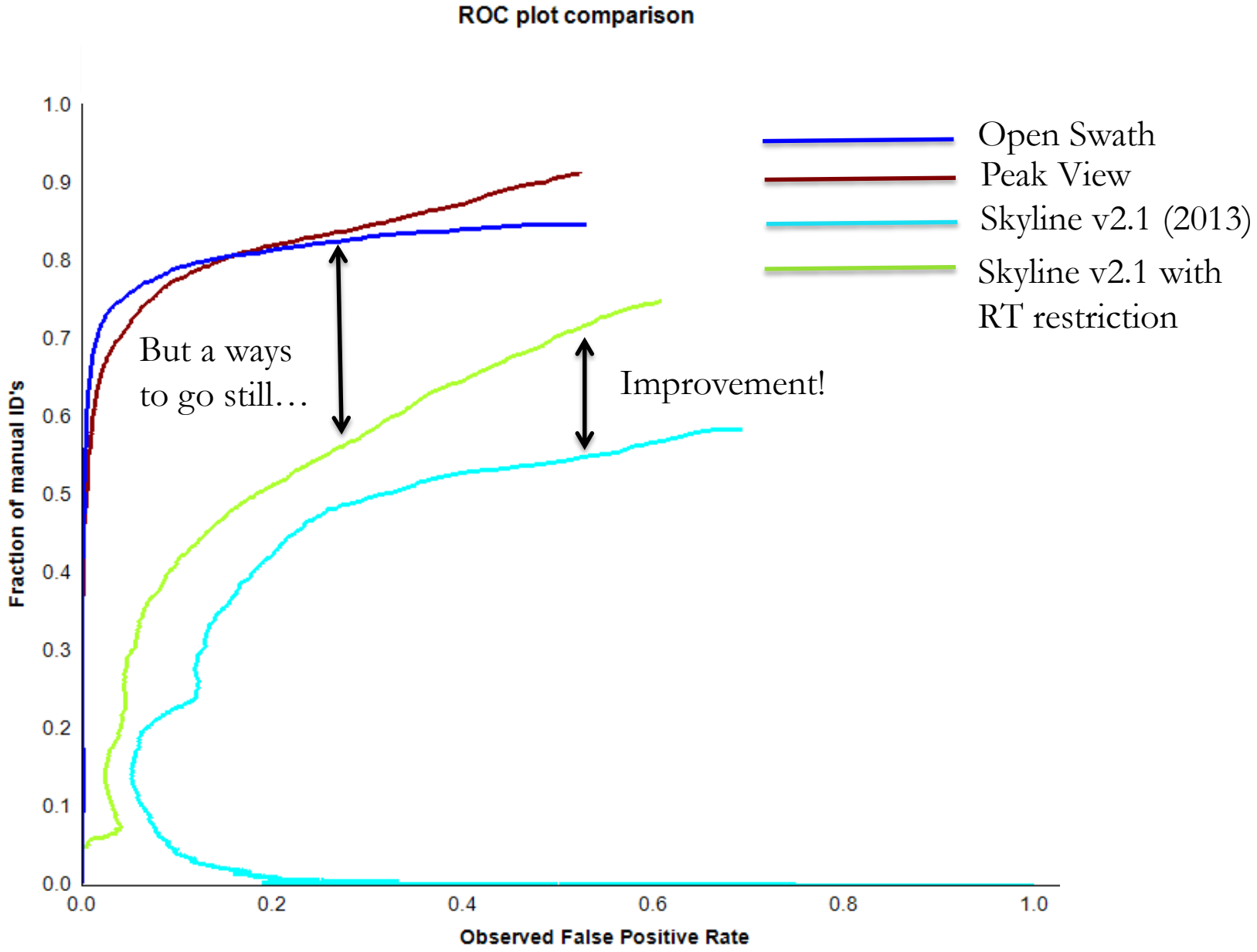
Problem 1: Full RT Range

- Initially, Skyline did not restrict the range to the predicted retention window by default
- Most users ended up importing the whole chromatogram making peak picking difficult



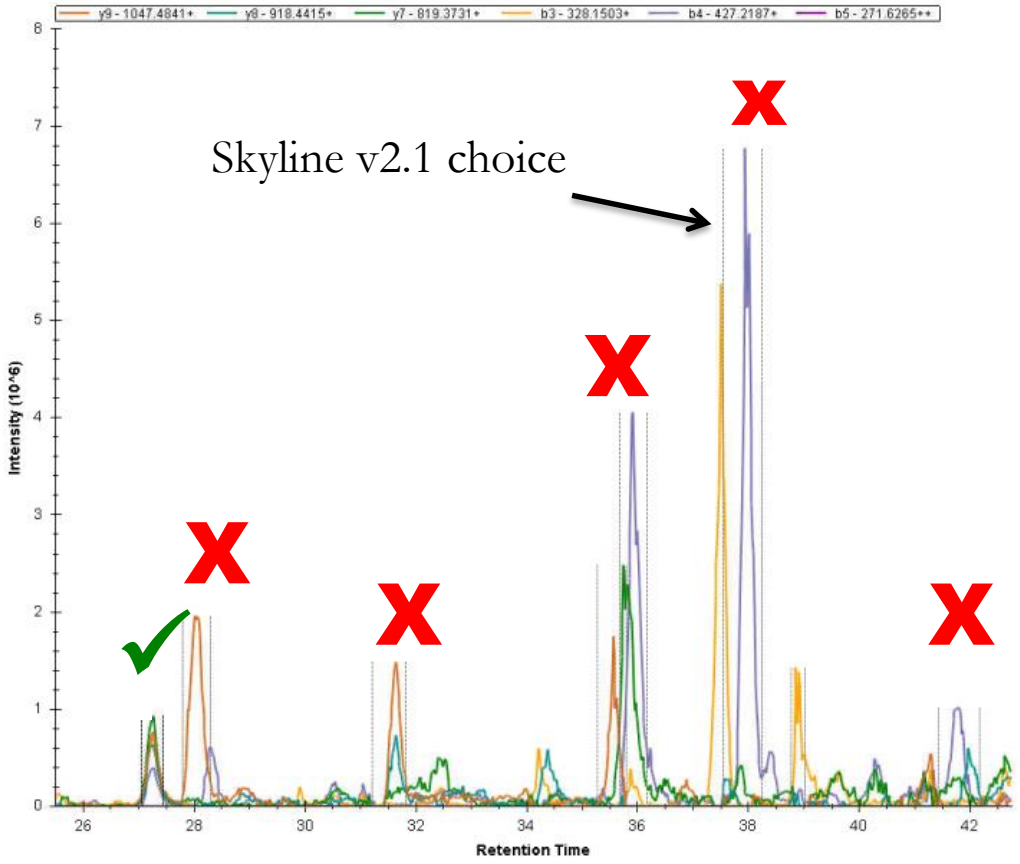


Impact of Chromatogram Restriction



Problem 2: Peak Scoring

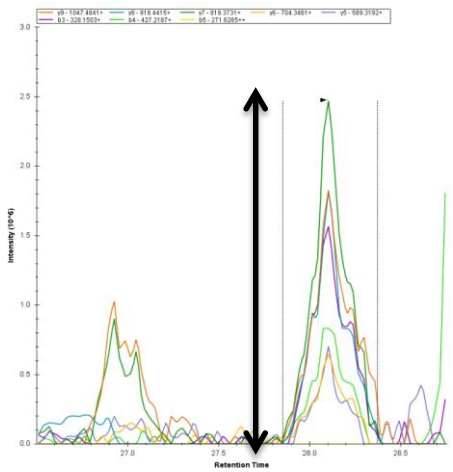
- In 2013, Skyline picked peaks based on intensity and crude co-elution measures
- Workflow geared towards SRM – Skyline suggests a peak, but you inspect manually, so OK if sometimes wrong



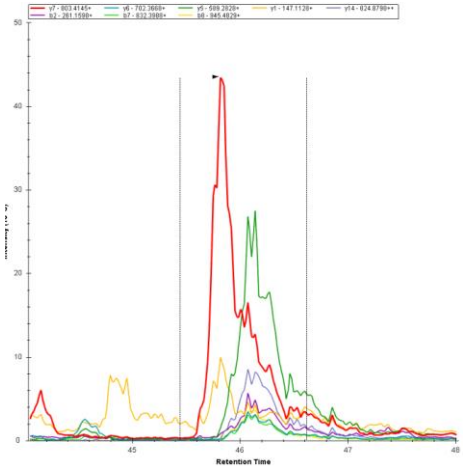
The mProphet Algorithm

Use machine learning (linear discriminant analysis) to form a composite score based on these (and other) indicators of peak quality, using decoy peptides to train (Reiter et al 2011)

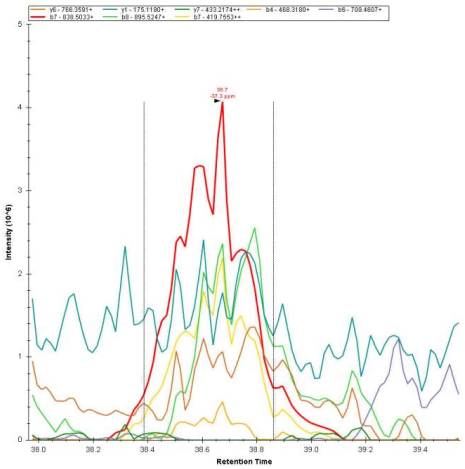
Intensity



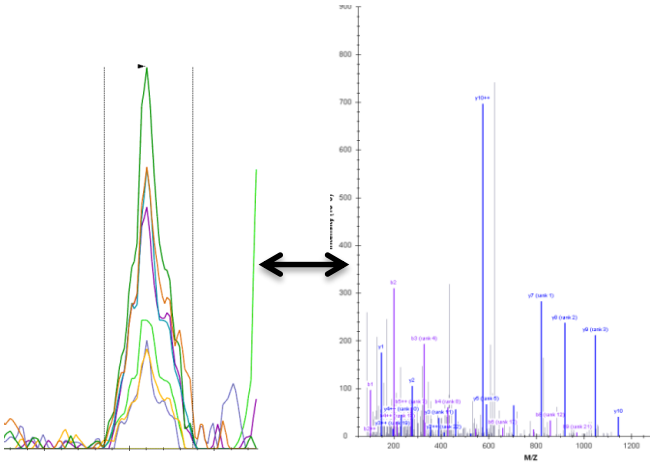
Co-elution



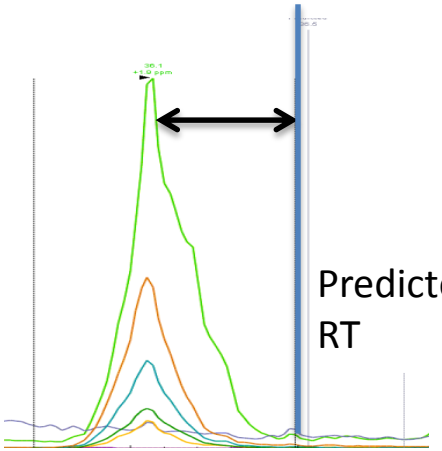
Shape



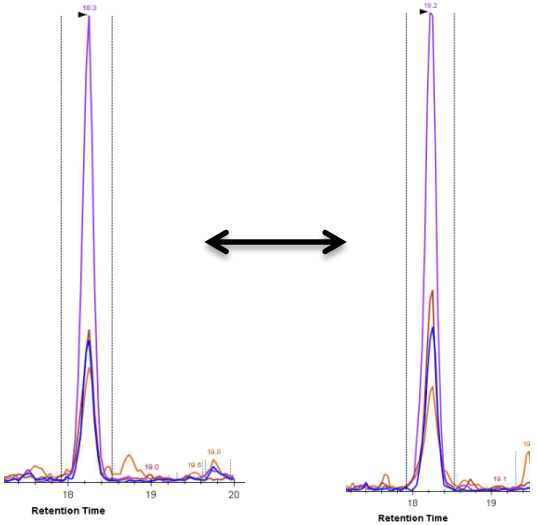
Spectral Library Correlation



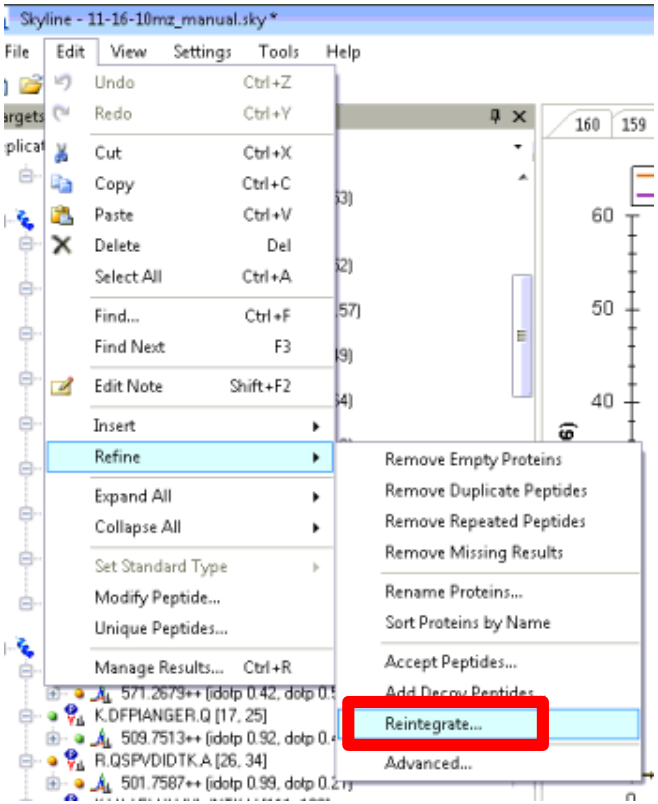
Retention Time

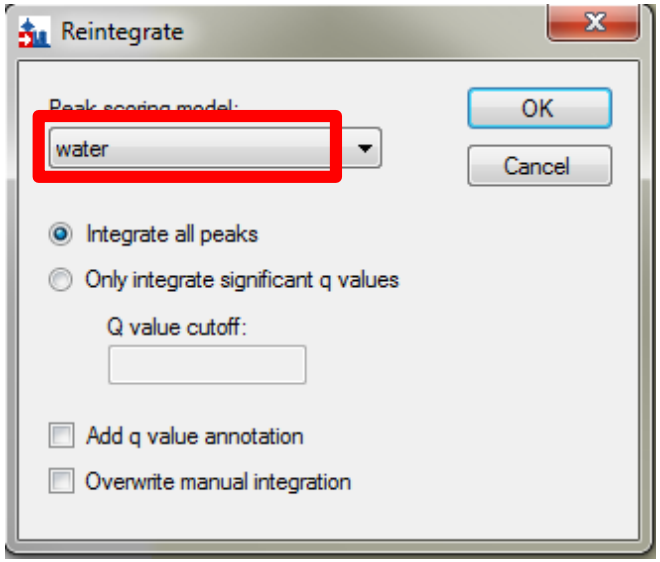


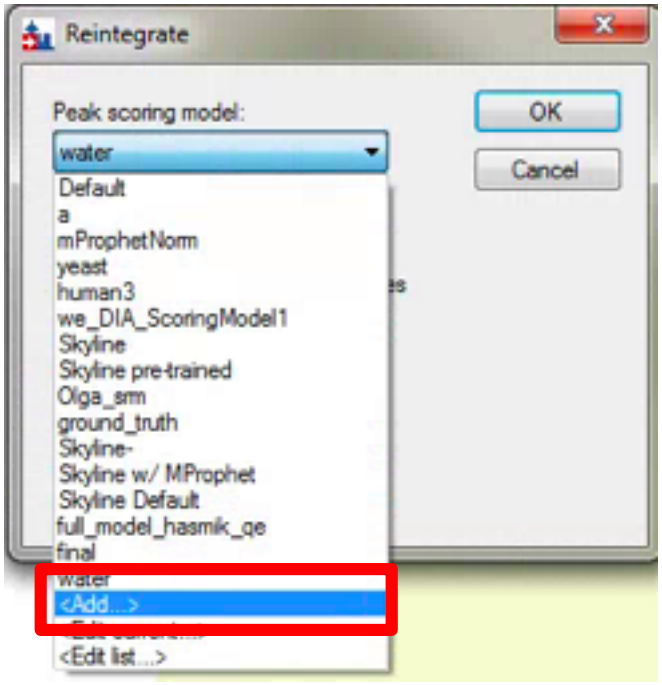
Comparisons w/ isotope standard



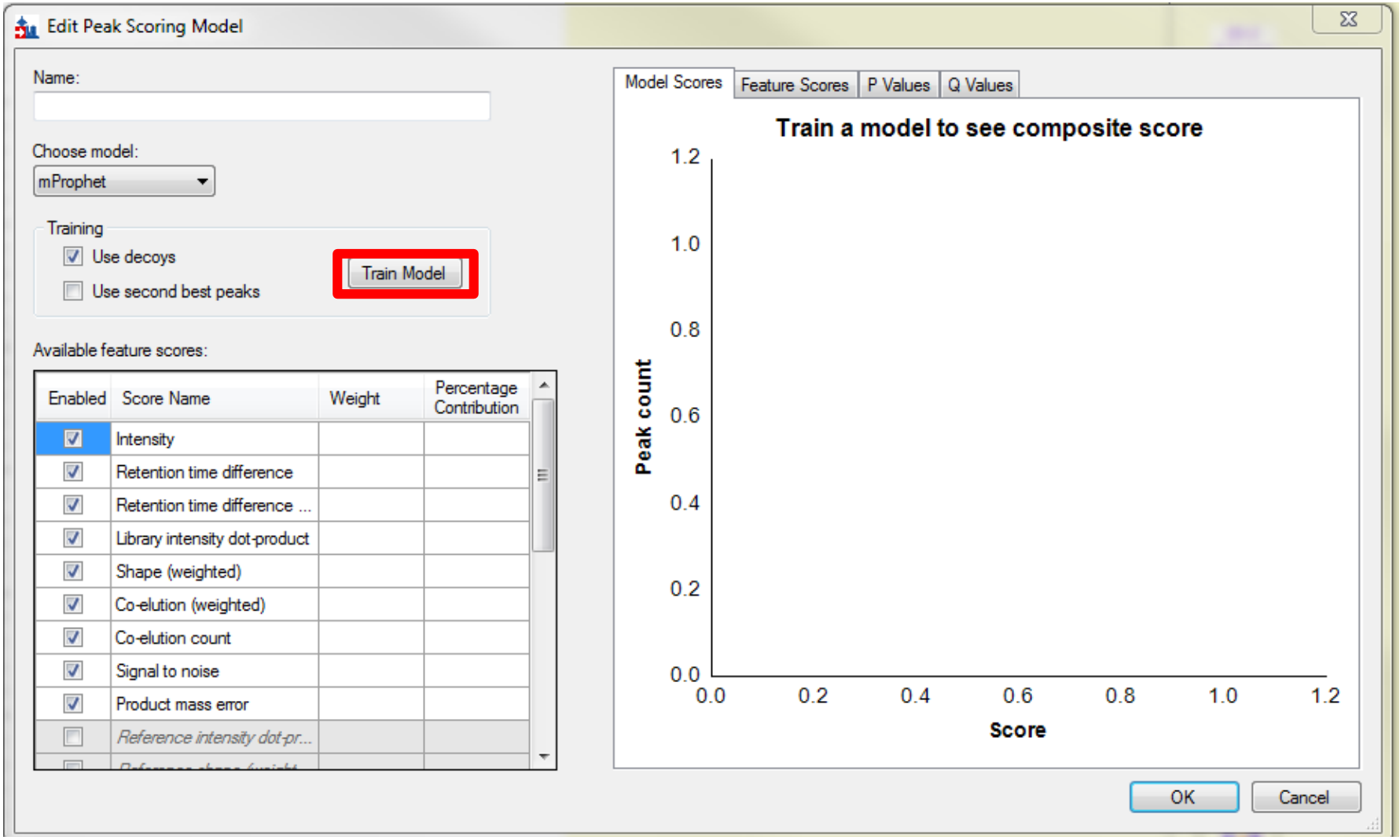
Using mProphet in Skyline



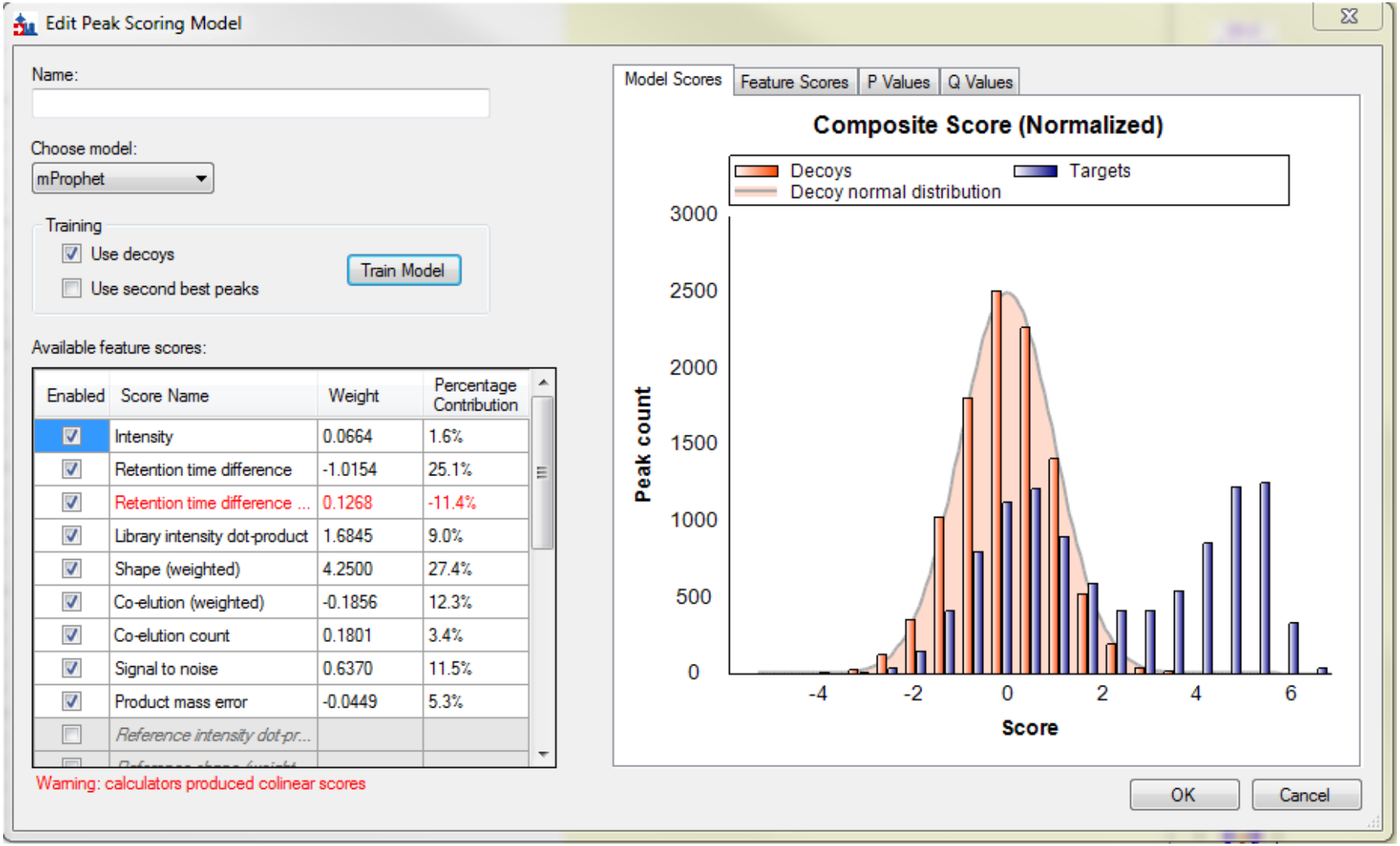




Using mProphet in Skyline

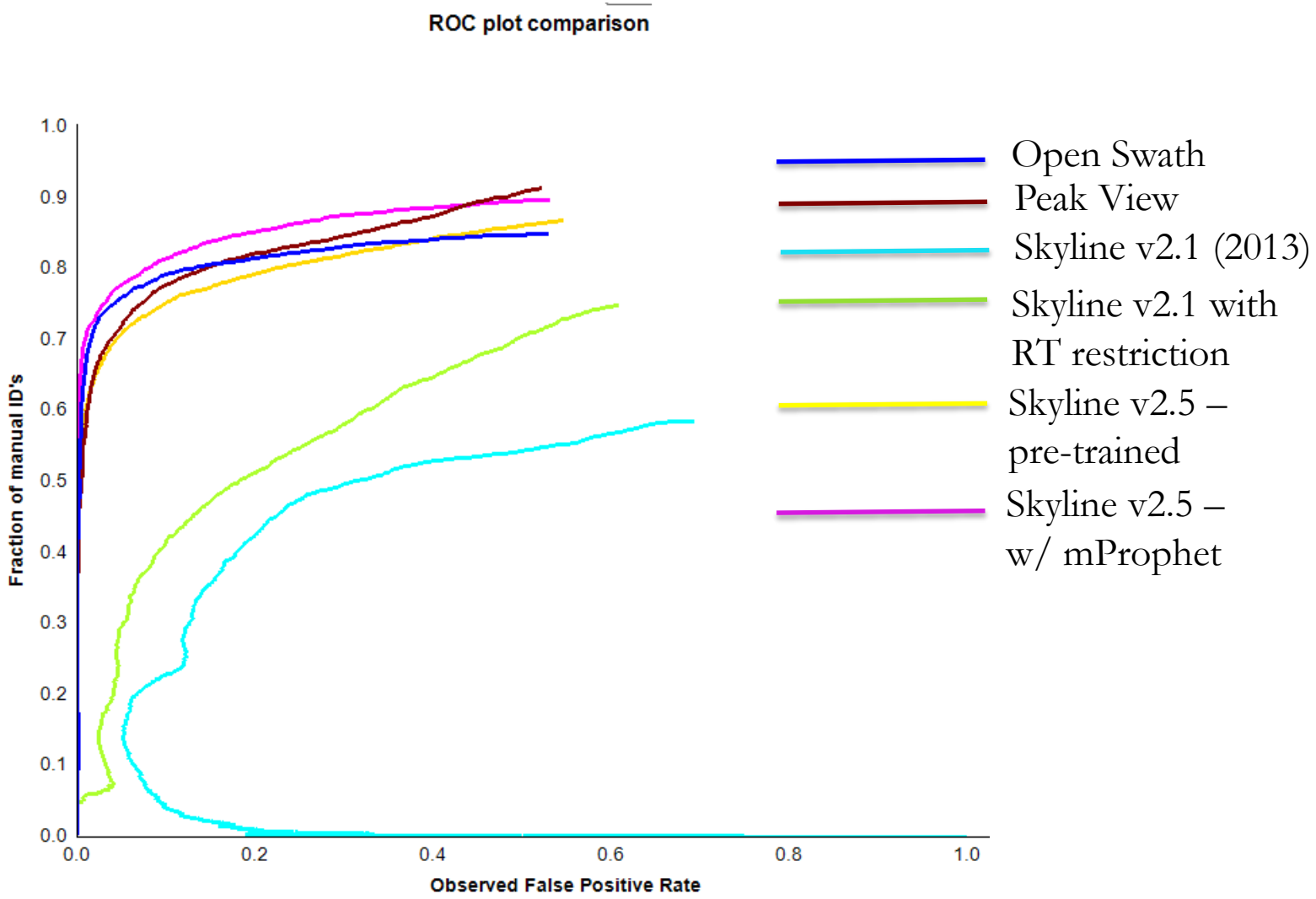


Using mProphet in Skyline



Swath Gold Standard Data (Human)

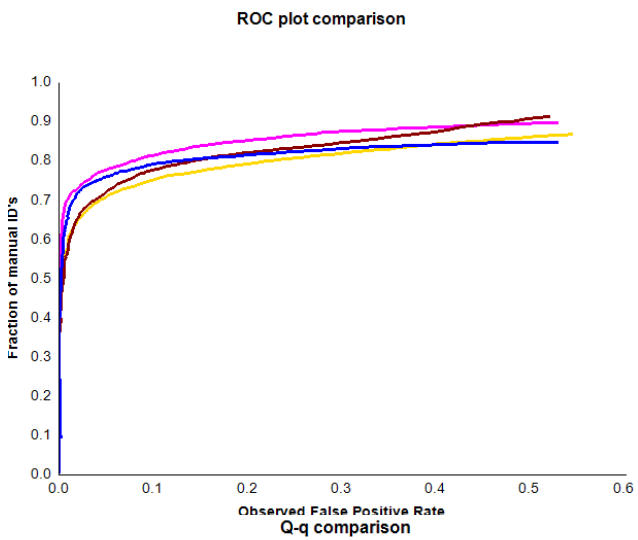
Skyline now matches or exceeds other tools.

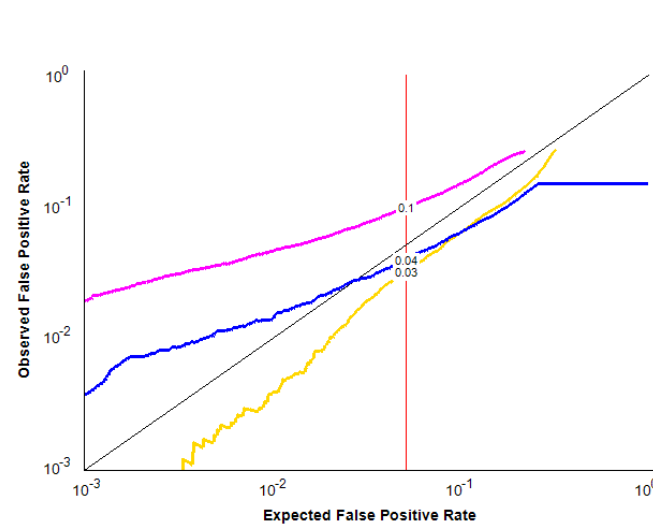


Results on Other Backgrounds

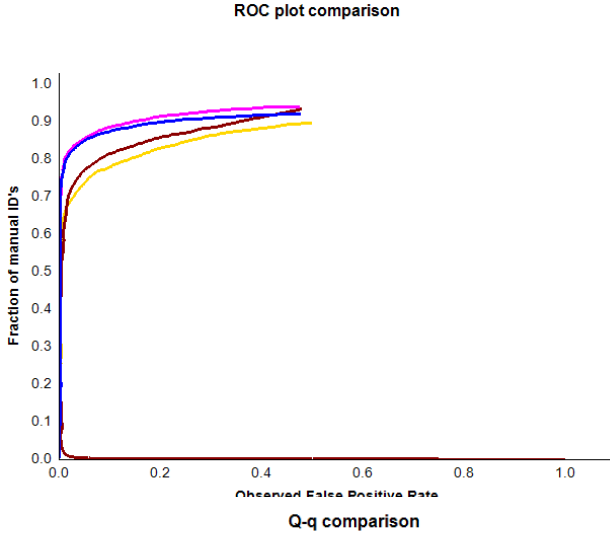
Open Swath
Peak View
Skyline v2.5 – pre-trained
Skyline v2.5 – w/ mProphet

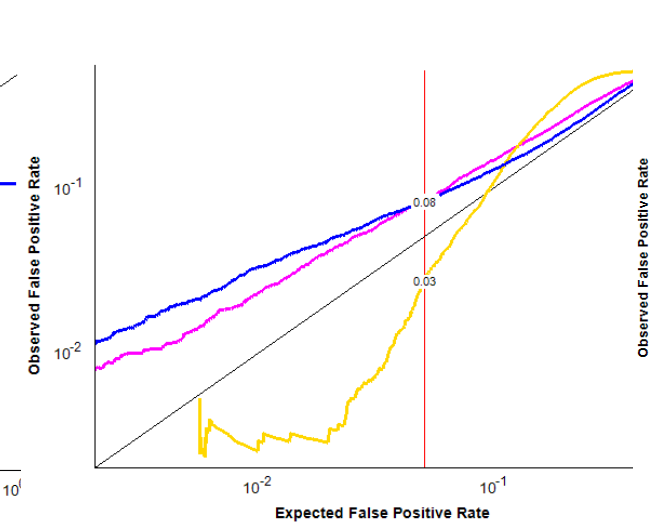
Human Background



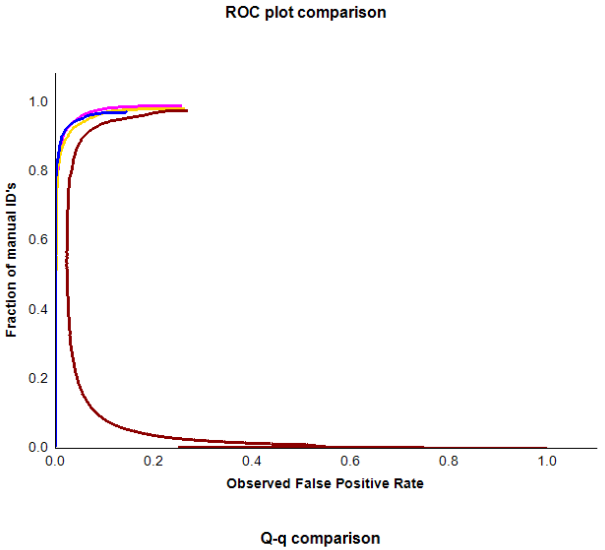


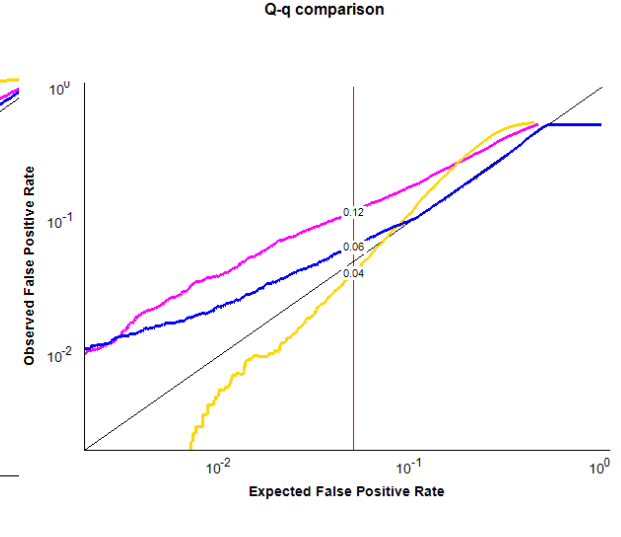
Yeast Background





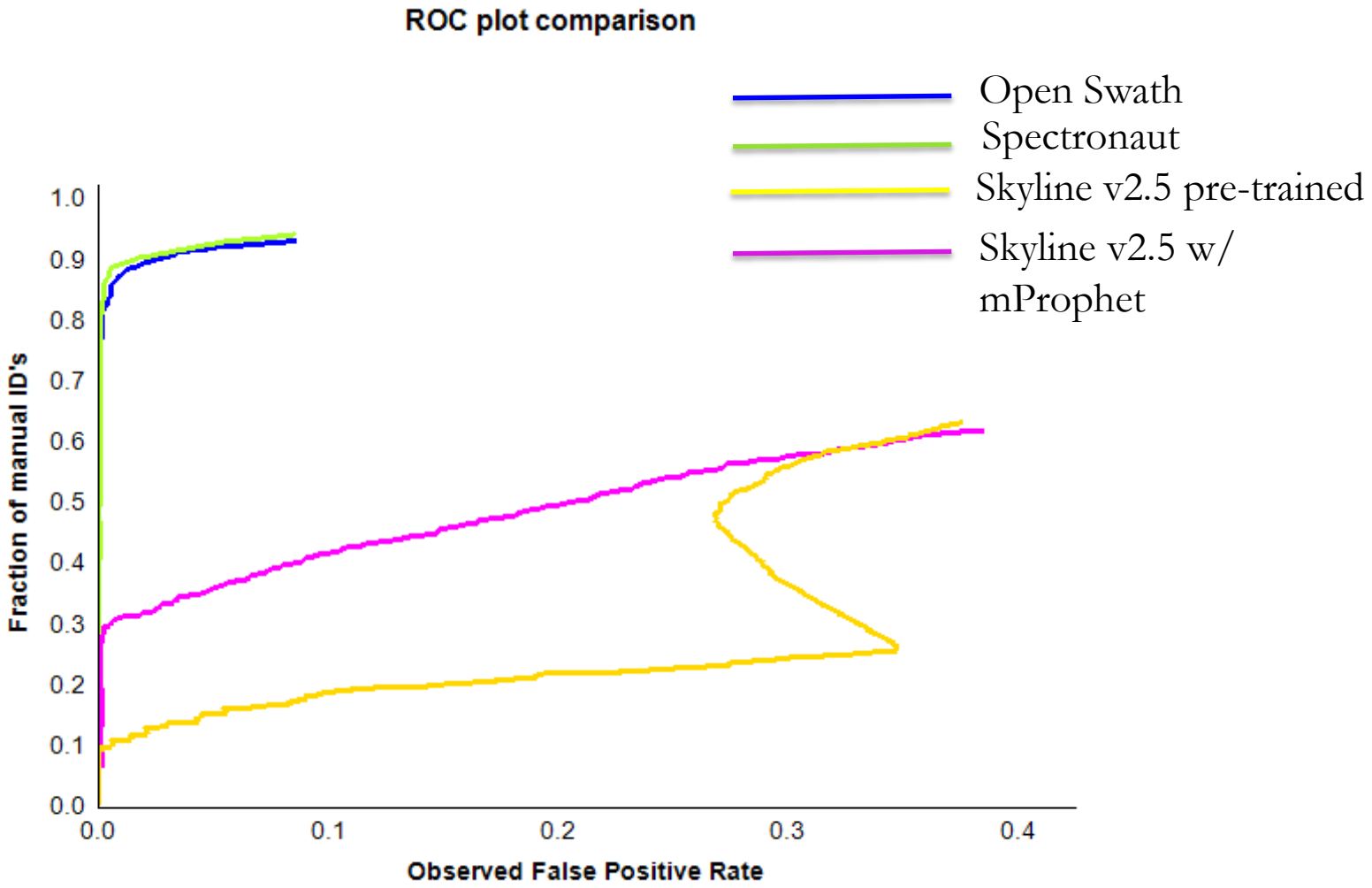
No Background





Problem 3: Isotopic Pairs

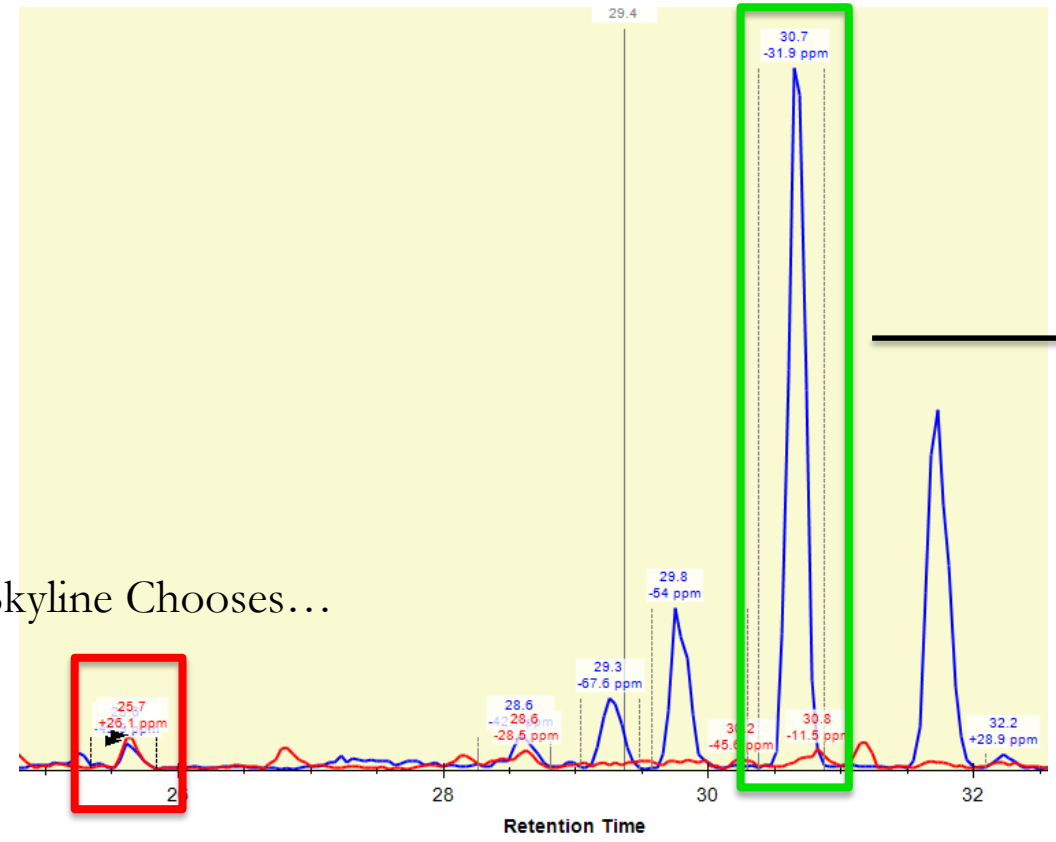
116 light/heavy synthetic peptide pairs spiked into depleted plasma background, 30 runs
Constant abundance heavy standards, light peptides in dilution series from 100 fm/ul to 0.01 fm/ul
4380 total peptide-file pairs
90% of peptide-runs had manually curated peak



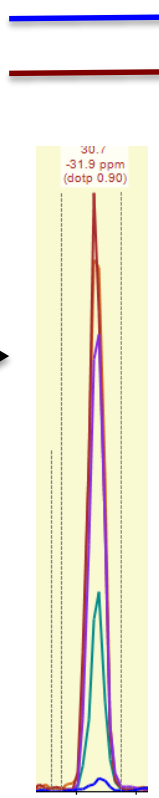
The cause of the problem

- mProphet only considers analyte-analyte and analyte-standard co-elution/shape/dotp
- If can't see analyte, peak will get a bad score even if standard is very easy to see

Correct peak is…



Skyline Chooses…



Standard (Heavy)

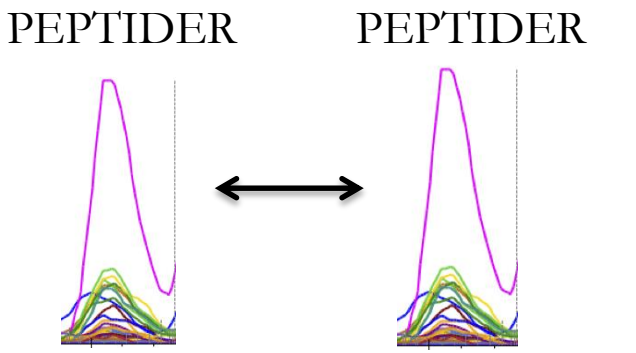
Analyte (Light)

This is very obviously the correct peak…

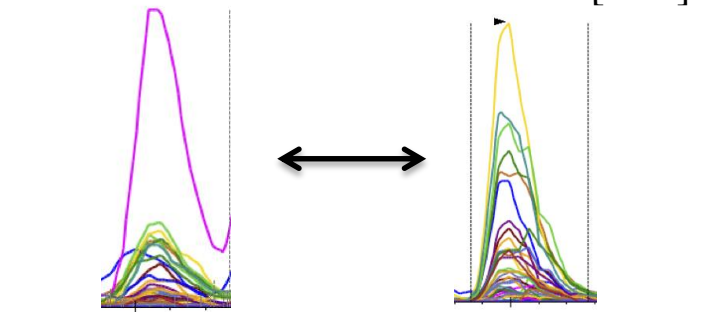
Adding Scores for Synthetic Peptides

Solution: add standard-standard co-elution/shape/dotp scores

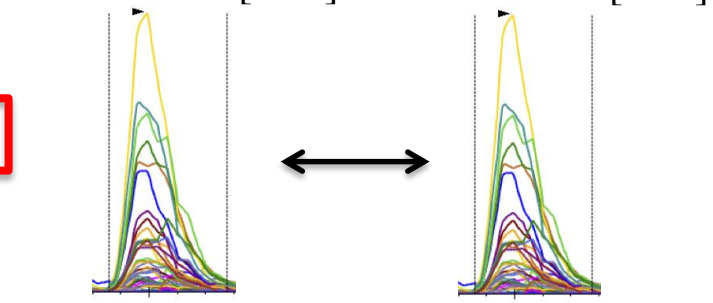
Analyte-Analyte

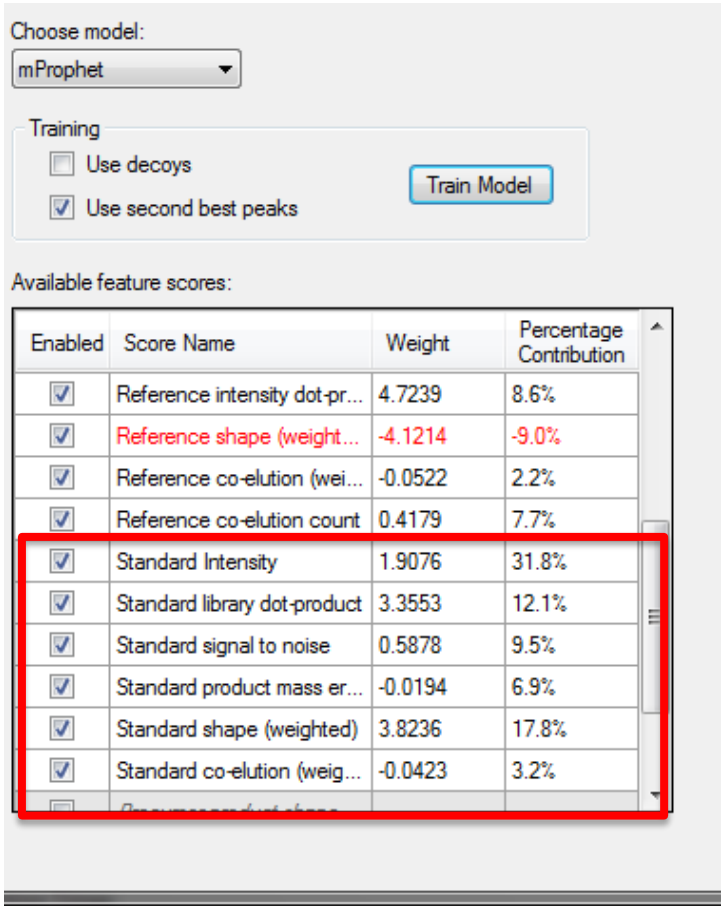


Analyte-Standard



Standard-Standard

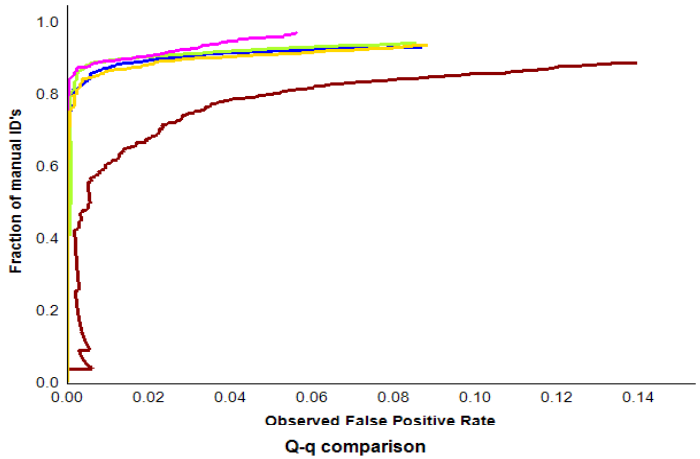


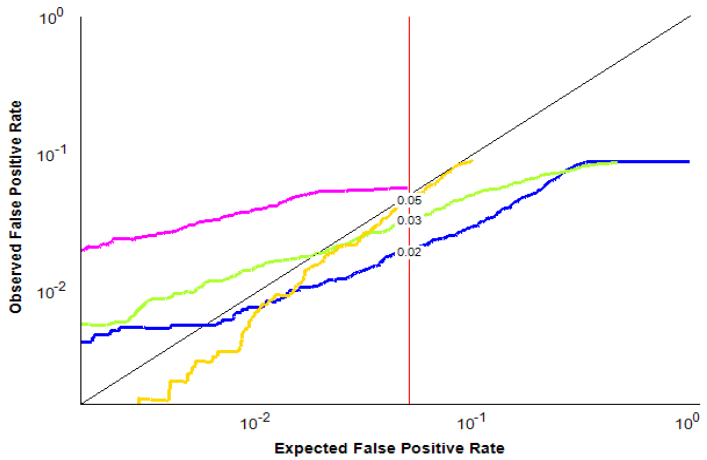


Pairs/Heavy Peptides

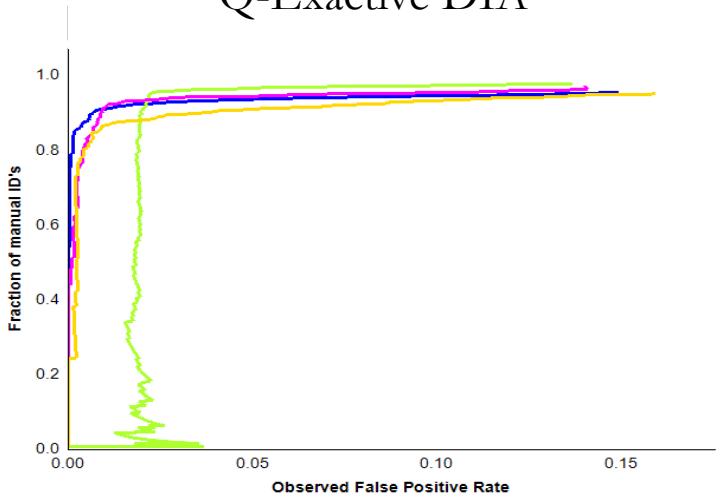
Open Swath
Peak View
Spectronaut
Skyline v2.5 – pre-trained
Skyline v2.5 – w/ mProphet

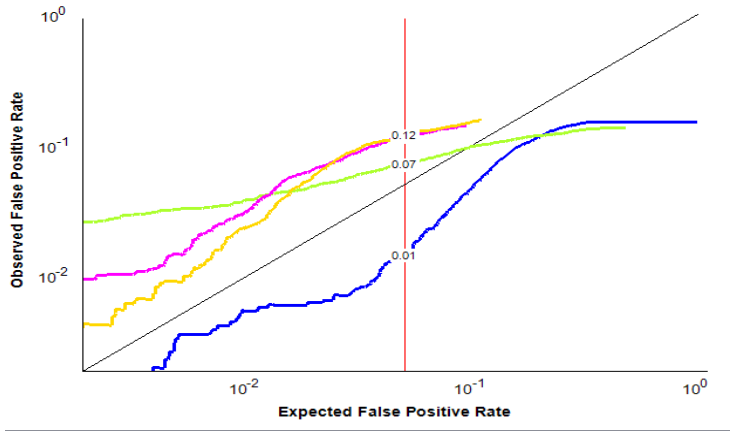
SWATH DIA





Q-Exactive DIA

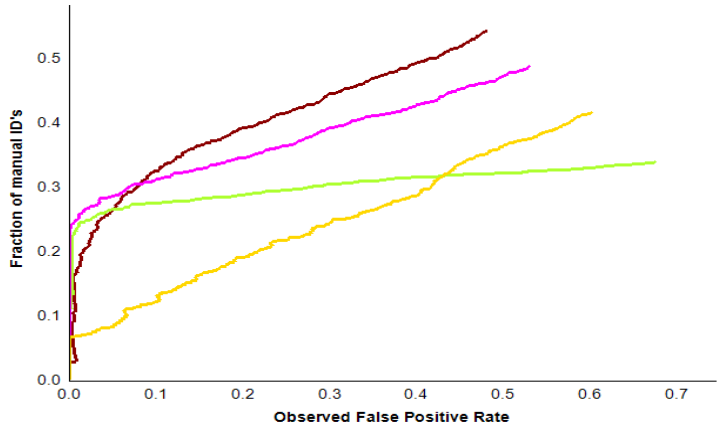




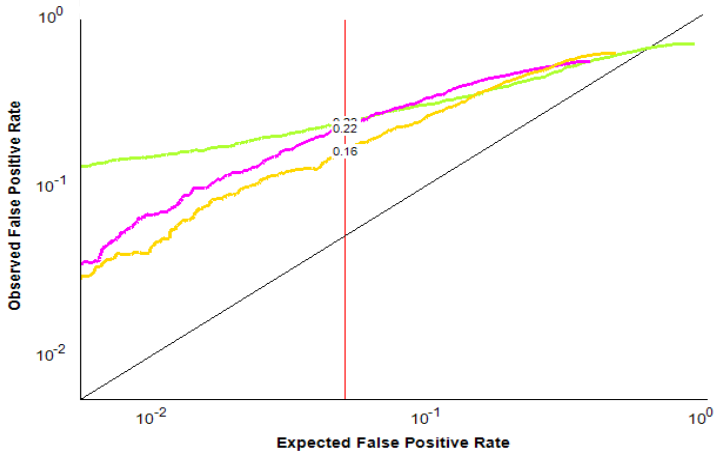
Light Peptides Only

Open Swath
Peak View
Spectronaut
Skyline v2.5 – pre-trained
Skyline v2.5 – w/ mProphet

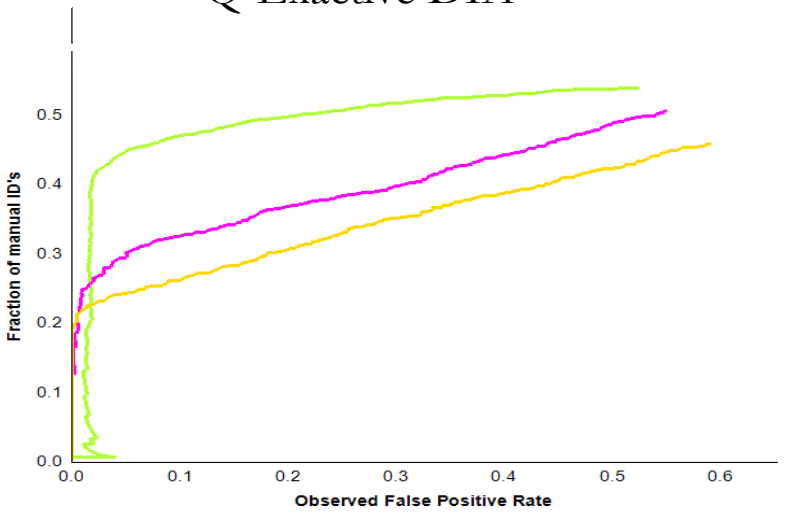
SWATH DIA



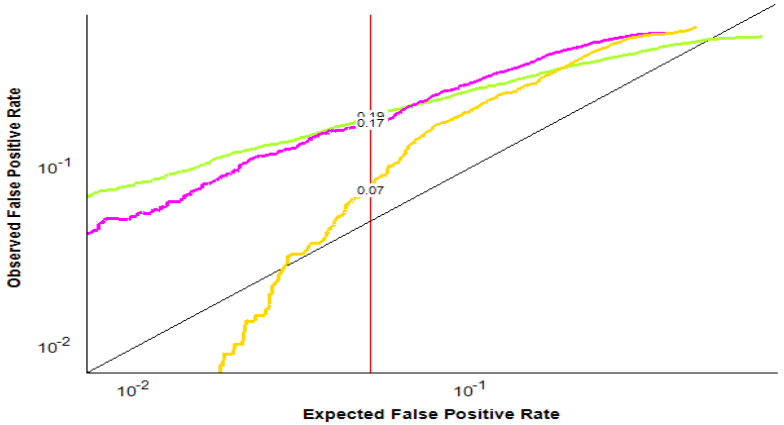
Q-q comparison



Q-Exactive DIA



Q-q comparison



Preliminary Conclusions

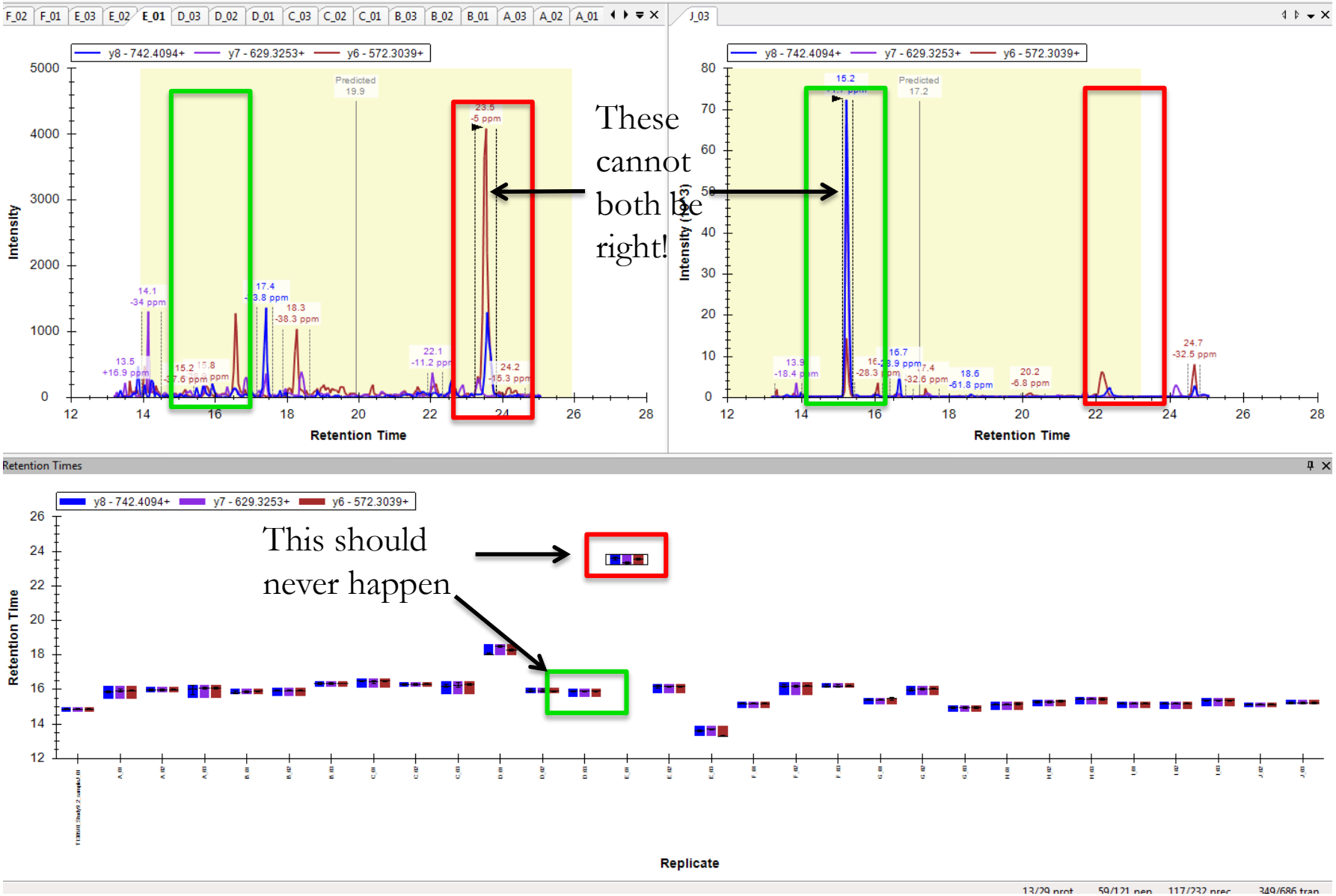
- **Skyline's peak picking is now roughly equivalent to other tools for processing DIA data**
- Proper treatment of isotope standards requires modification of published algorithms
- Most tools overstate statistical confidence

Help Us Make Skyline Better!

Send us your DIA data sets for testing Skyline against other tools and improving it.

- **We are interested in**:
  - Spiked-in synthetic peptides, especially dilution series
  - Biological data, especially if it has some ground truth (manually picked peaks, SILAC, isotope standards)
- **We will share anything we find with you**
- **We will keep your data confidential, aside from very high level peak picking statistics**

Future Plans – Cross Run Peak Picking



Future Peak Picking Plans (Winter Release)

- Multi-replicate peak picking
- Comparison of Skyline with other tools on larger number of DIA data sets
- Getting the statistics right

Acknowledgements

Stanford University
Parag Mallick
Seema Sharma

Purdue Univers
Olga Vitek

ETH Zurich
Hannes Rost
Reudi Aebersold
Lucia Espona Pernas

Steve Tate

University of Washington
Brendan MacLean
Don Marsh
l MacCoss
d Johnson
Ying Sonia Ting

Broad Institute
Hasmik Keshishian

Special thanks to **Brendan MacLean** for creating Skyline and reviewing all my code!

Dario's Poster: Thursday June 19th (#035) 10:30 AM – 2:30 PM