Harnessing the Power of Deep Learning with the CHIMERYS™ Intelligent Search Algorithm and Proteome Discoverer™ 3.0 Software

Philip Loziuk, Ph.D.
Product Manager

The world leader in serving science

Proteome Discoverer history

Software for identification and quantification of proteins in complex samples

1.0 – July 2008	Workflow-based solution for protein identification
1.1 – Nov 2009	Automation, batch processing, reporter ion quantification (e.g. TMT)
1.2 – Apr 2011	SILAC quantification
1.3 – Mar 2012	Validation (Percolator, PhosphoRS), biological annotation, 64-bit
1.4 – May 2013	Deep data mining (Sequest HT, library searching)
2.0 – Mar 2015	Architectural changes, study management, large data sets
2.1 – Oct 2015	Improved reporter ion quantification
2.2 – July 2017	Label Free Quantification, Statistics, Cross-Linking
2.3 – Feb 2019	Improved library search, heat maps, PTM site tables, cross-link quantification, annotation groups, ProSightPD™
2.4 – Oct 2019	Precursor detector/chimeric spectra, scripting node, TMTpro™, new licensing
2.5 – Oct 2020	**INFERYS® rescoring**, predicted spectral libraries, enrichment service, improvements in TMT and PTM analysis, SureQuant™ Method Development, ProSightPD 4.0, XlinkX improvements
3.0 – ASMS 2022	**CHIMERYS®, INFERYS® 2.0, Comet search engine (for TMT RTS)**

Proteome Discoverer 3.0

- **CHIMERYS**
  - Cloud and deep learning-based search engine
  - Up to 2x more unique peptide IDs for a given data file compared to PD 2.5
  - Includes INFERYS 2.0

- **INFERYS 2.0**
  - TMT, TMTpro and HLA peptides
  - HCD and CID MS/MS
  - 1-6+ precursors
  - Display of predicted spectrum in mirror plot (also available with CHIMERYS)

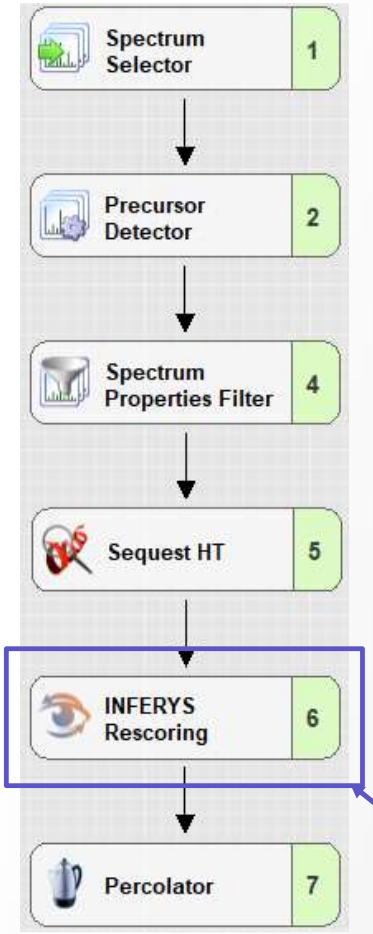
- **Other features**
  - Comet search engine for close match to OT Eclipse real time search (which also uses Comet)
  - SureQuant method exporter node
  - Updated installer
  - Faster Percolator FDR
  - Minora Improvement
  - Faster XlinkX search

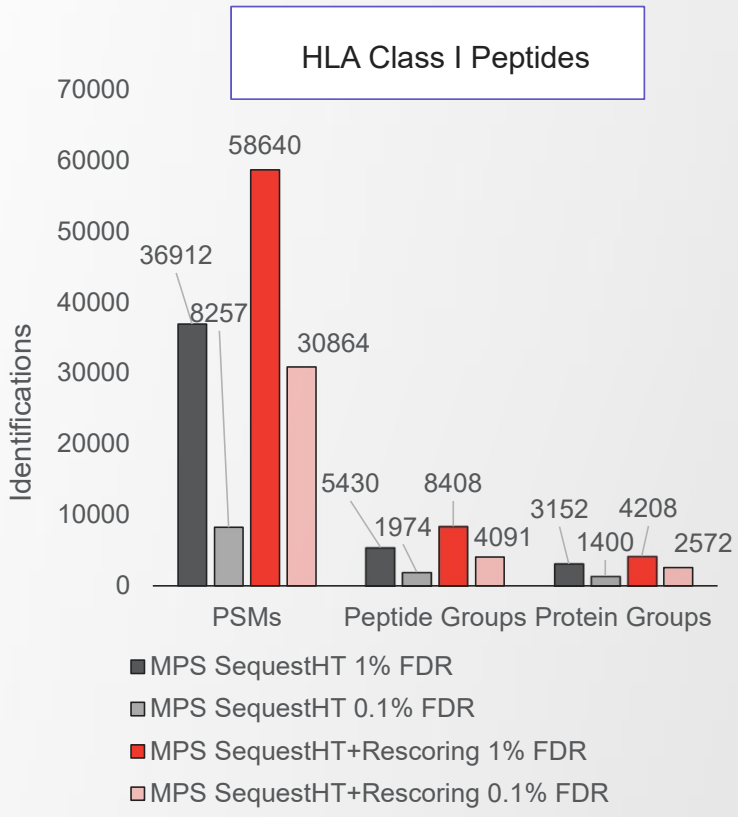
What is INFERYS®?

Deep learning-based model for peptide MS/MS spectral prediction

Prediction-based rescoring for improved confidence

- **INFERYS Rescoring** node predicts MS/MS spectra for PSMs identified by Sequest HT
- Updated **Percolator** node includes new features for improved peptide ID confidence
- Recent publication:
  - Zolg, DP *et al*, "INFERYS rescoring: boosting peptide identifications and scoring confidence of database search results", *Rapid Commun. Mass Spectrom.*, **2021,** https://doi.org/10.1002/rcm.9128



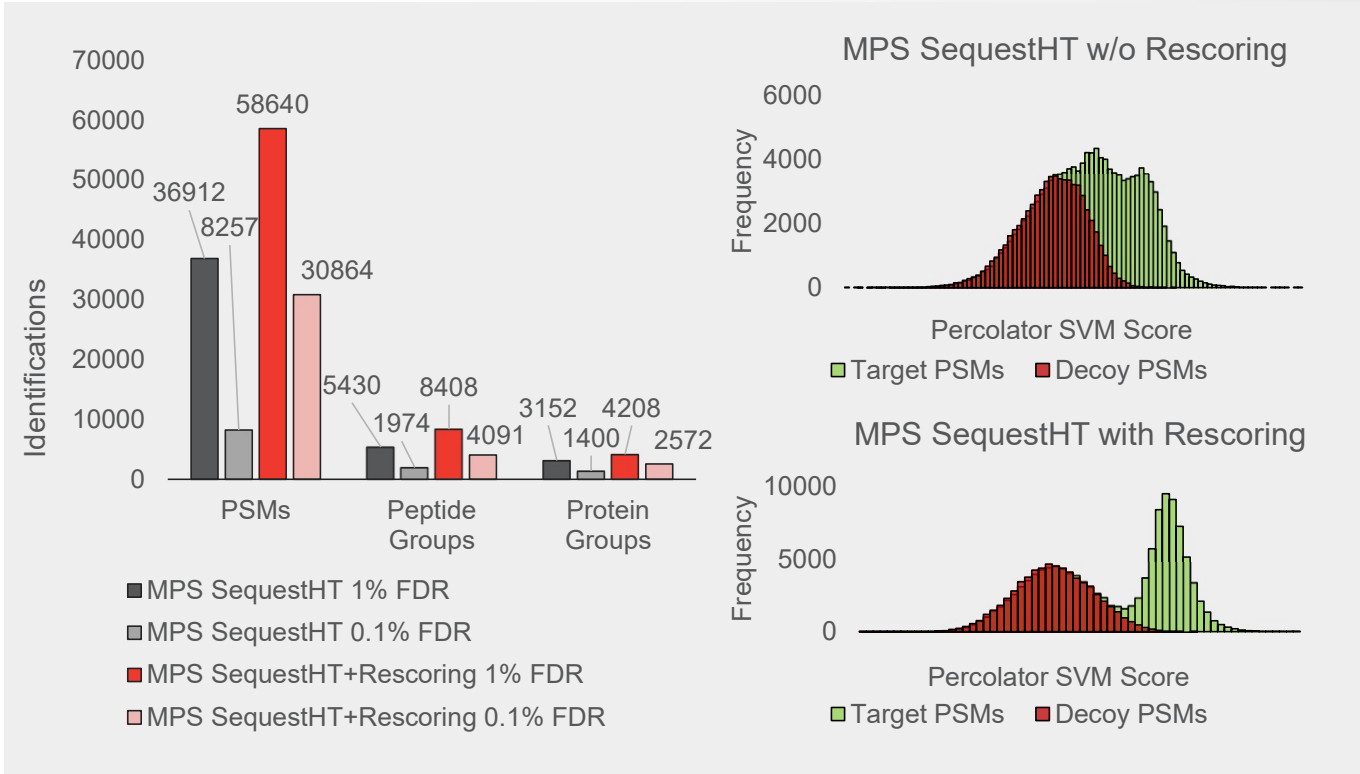


New INFERYS Rescoring node

Results – Immunopeptidomics data set

HLA Class I data set – Patient derived melanoma cell line (*Chong. et al, Nat. Com. 2020*)

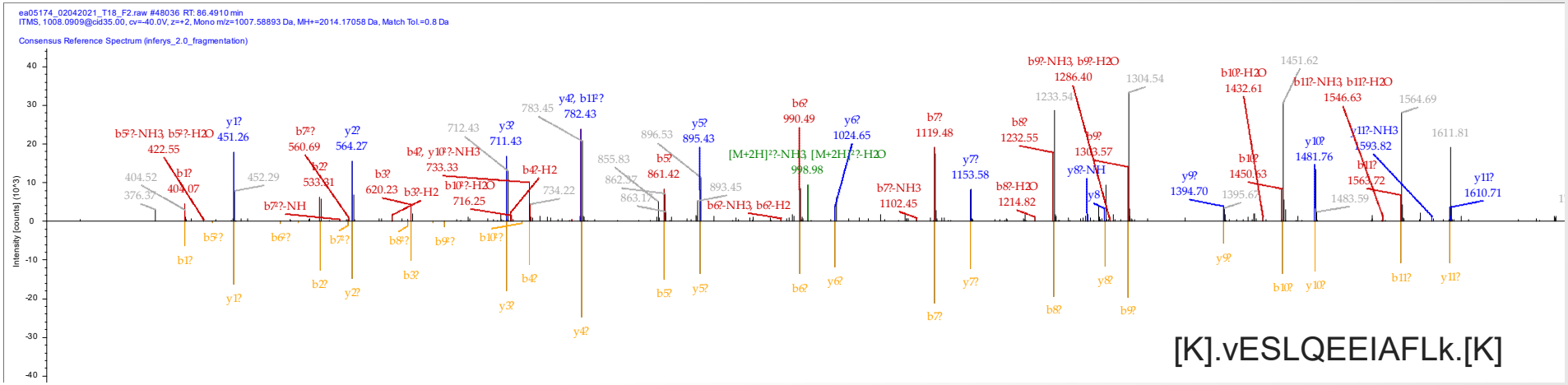
- HLA Class I peptide data set
- Huge search space and very similar peptide properties
- 59% increase in identifications at PSM, 55% at peptide and 34% at protein level
- Reduces peptide loss at 0.1% compared to 1% FDR from 64% to 25%



INFERYS 2.0 Rescoring for TMT Analysis

Prediction of TMT, TMTpro peptides now available in Proteome Discoverer 3.0 software

- INFERYS-predicted TMT-modified spectrum shown in the lower half of the mirror plot – **NEW in PD 3.0!**



Comet in PD 3.0

New Comet node available as a search engine to help increase IDs for TMT Real-Time Search data

Optional custom Comet params file to modify advanced parameters

PTM list provided by Proteome Discoverer

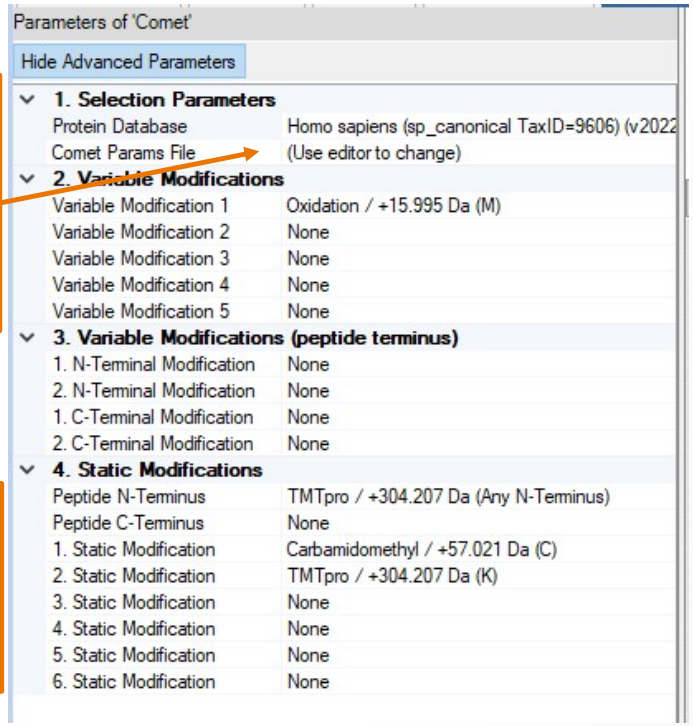
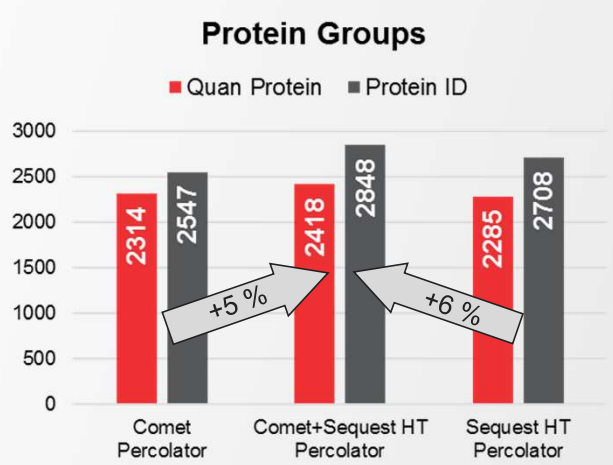


Figure 9. Protein groups identified and quantified TMT18plex Yeast digest standard using different PD analysis workflow

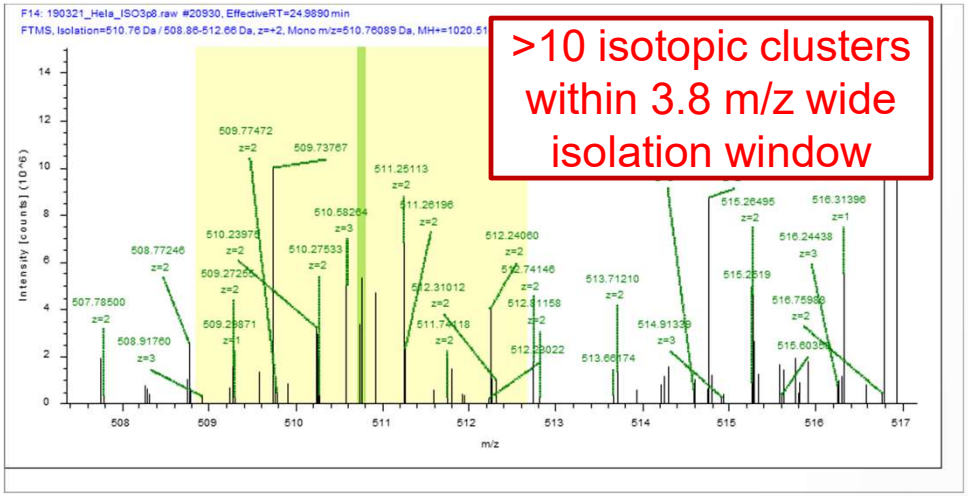


From Liu *et al,* ASMS 2021 poster

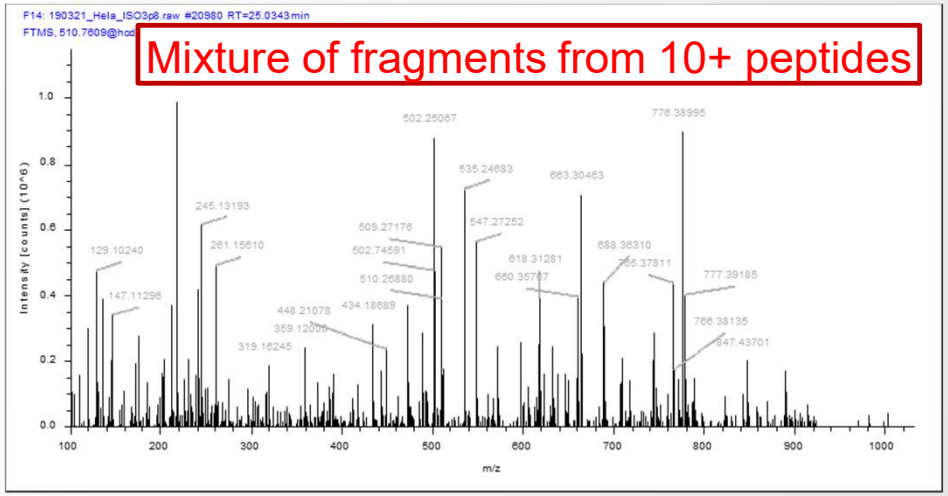
Chimeric spectra

Mass spectrometer can isolate and fragment many peptides for the same MS/MS spectrum

Precursor scan



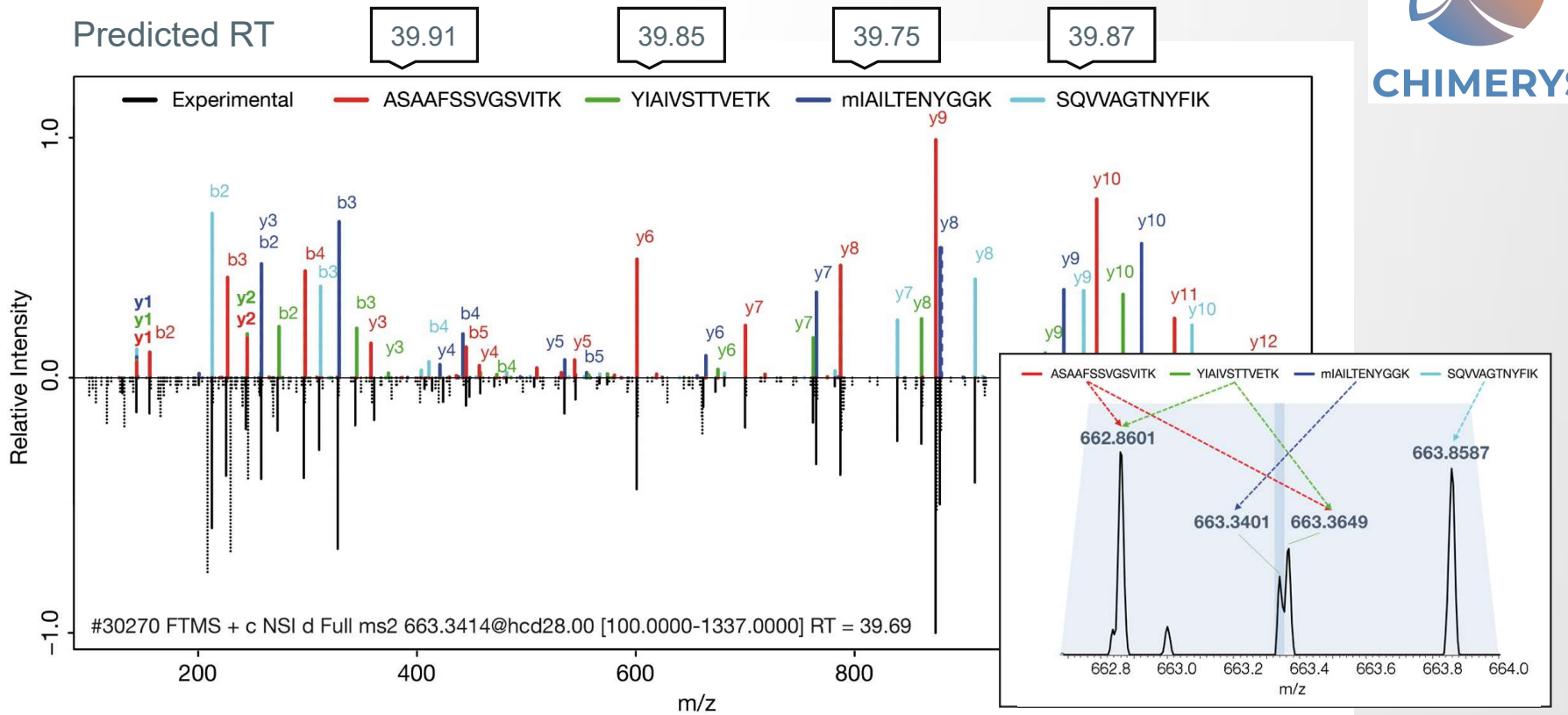
MS/MS spectrum



- No current search engine can identify more than a few of the peptides in this type of spectrum, leading to missing peptide IDs
- Use one or more extra dimensions of separation to attempt to simplify spectra in order to identify more of these peptides

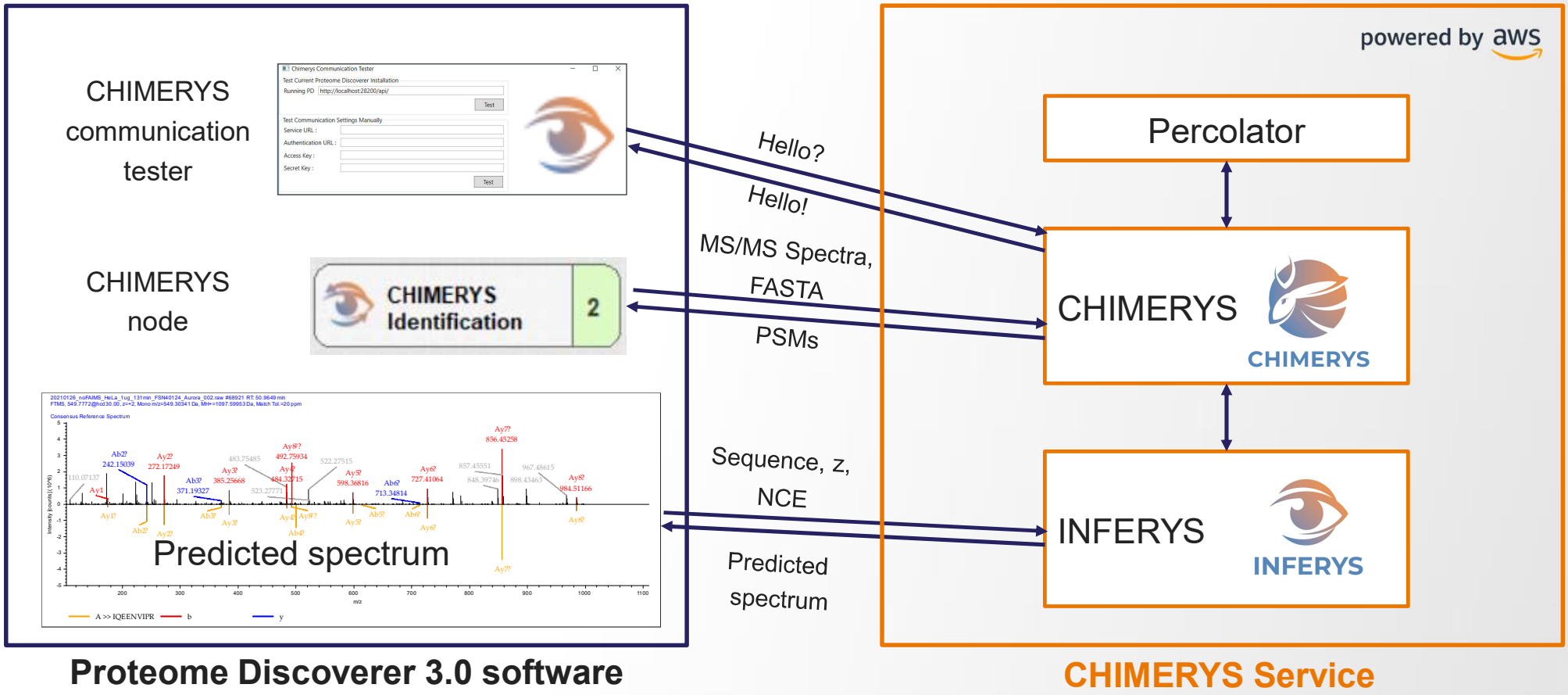
A novel, AI-based approach for chimeric spectrum ID

Intensity-based deconvolution of chimeric spectra based on accurate predictions.



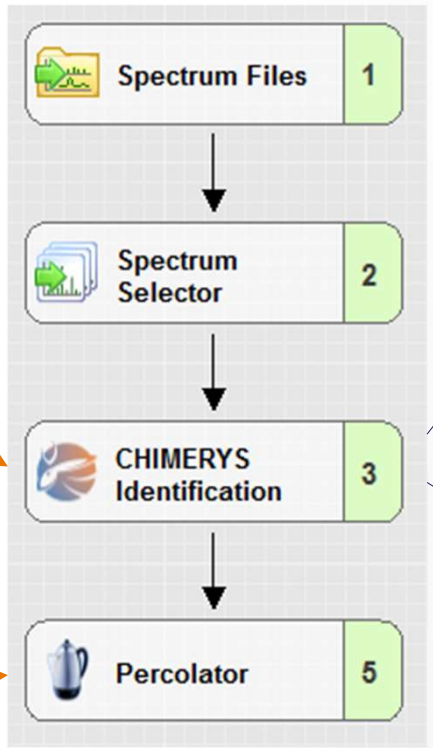
CHIMERYS service

Proteome Discoverer 3.0 software communicates via 3 different interfaces



Typical CHIMERYS workflow in Proteome Discoverer 3.0

Substitute CHIMERYS for Sequest HT



All CHIMERYS searches are on the cloud

For CHIMERYS, Percolator is also run on the cloud

Select FASTA database, enzyme specificity, tolerance, PTM

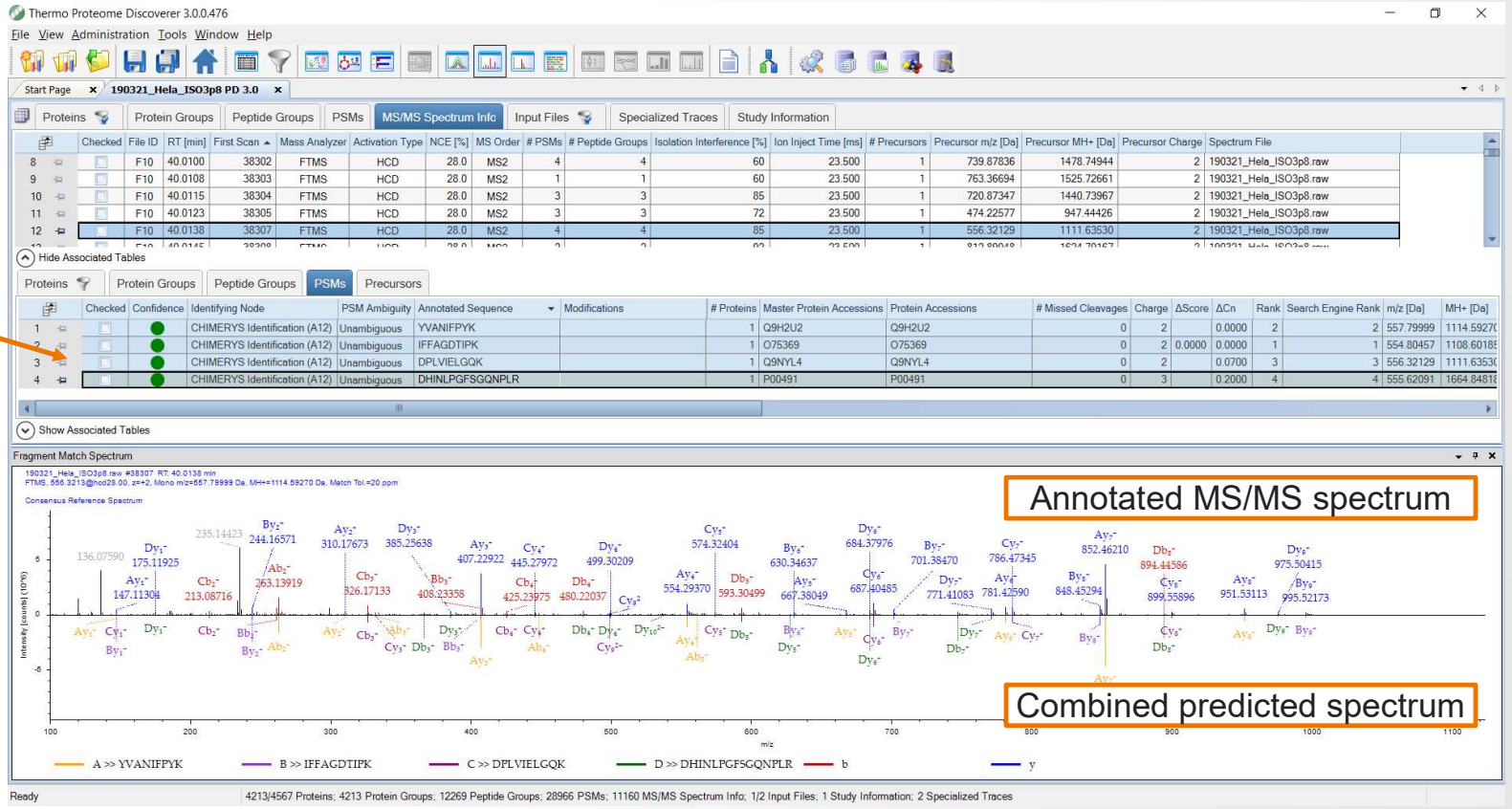
1. Input Data	
Protein Database	uniprot.human.reviewed.20211102.fas
Prediction Model	inferys_2.0_fragmentation
Enzyme	Trypsin
Max. Missed Cleavage Sites	2
Min. Peptide Length	7
Max. Peptide Length	30
Max. Number Of Modification:	3
Min. Peptide Charge	2
Max. Peptide Charge	4
2. Tolerances	
Fragment Mass Tolerance	20 ppm
3. Dynamic Modifications	
Max. Equal Modifications Per	3
1. Dynamic Modification	Oxidation (M)
2. Dynamic Modification	
1. N-Terminal Modification	
4. Static Modifications	
Peptide N-Terminus	
1. Static Modification	Carbamidomethyl (C)
2. Static Modification	TMTpro 16plex Tandem Mass Tag (
5. Validation	
Percolator Mode	Auto

CHIMERYS results presentation

Predicted spectrum for CHIMERYS and INFERYS Rescoring new for PD 3.0 software

Select up to 5 PSMs from same MS/MS spectrum

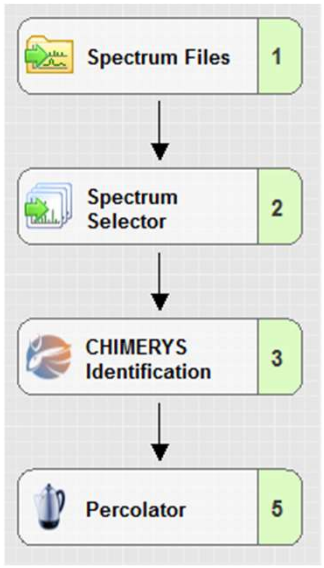
Mirror plot automatically updates



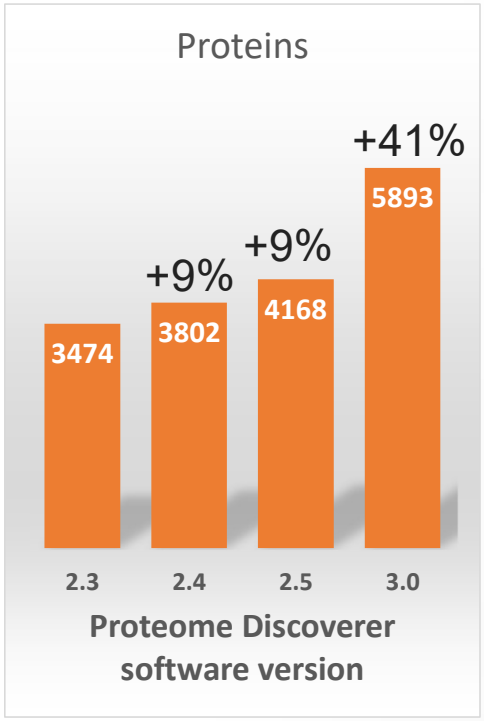
Improvement in peptide ID by release

HeLa dataset, 90-minute run, ID counts by Proteome Discoverer software release version

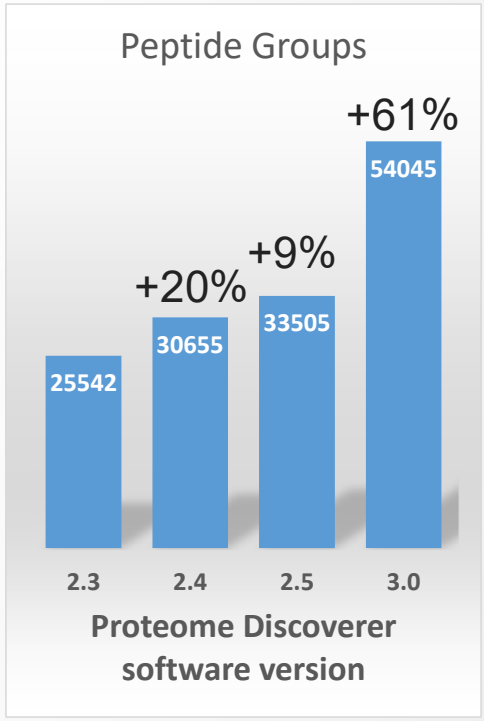
PD 3.0



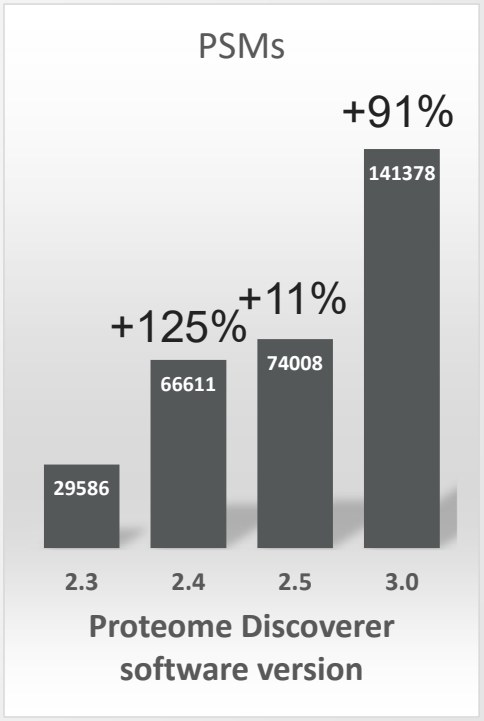
New CHIMERYS intelligent search algorithm!



CHIMERYS produces 70% more proteins than PD 2.3



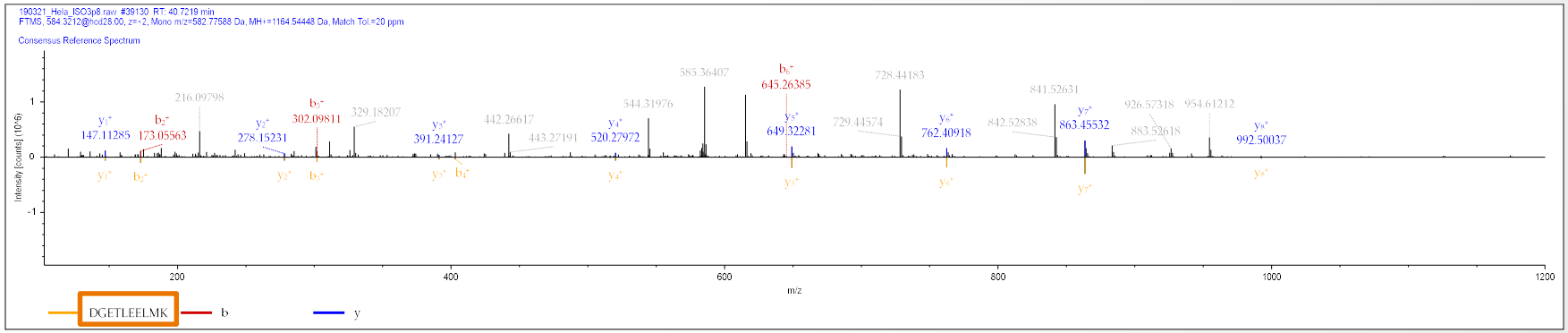
CHIMERYS produces more than 2x more unique peptides than PD 2.3

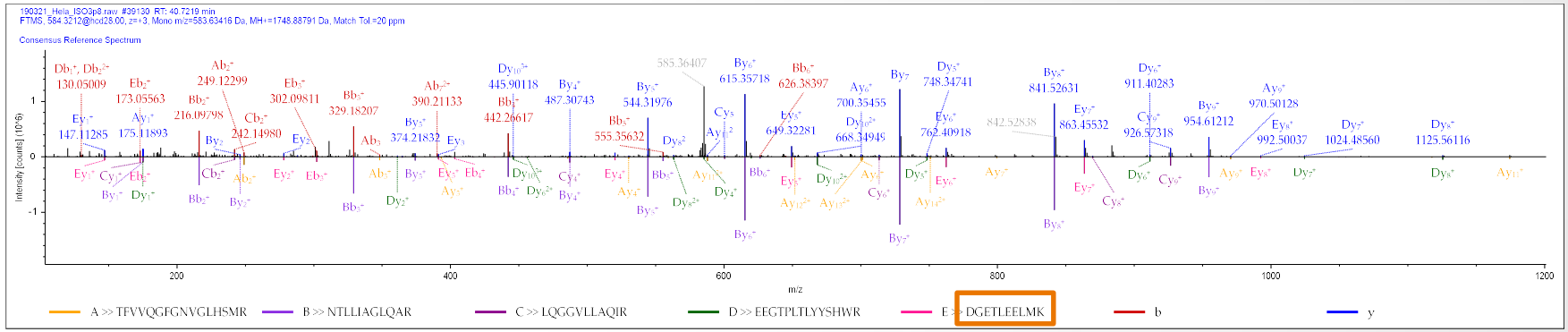


CHIMERYS produces almost 5x as many PSMs as PD 2.3

Proteome Discoverer 2.5 vs 3.0 software

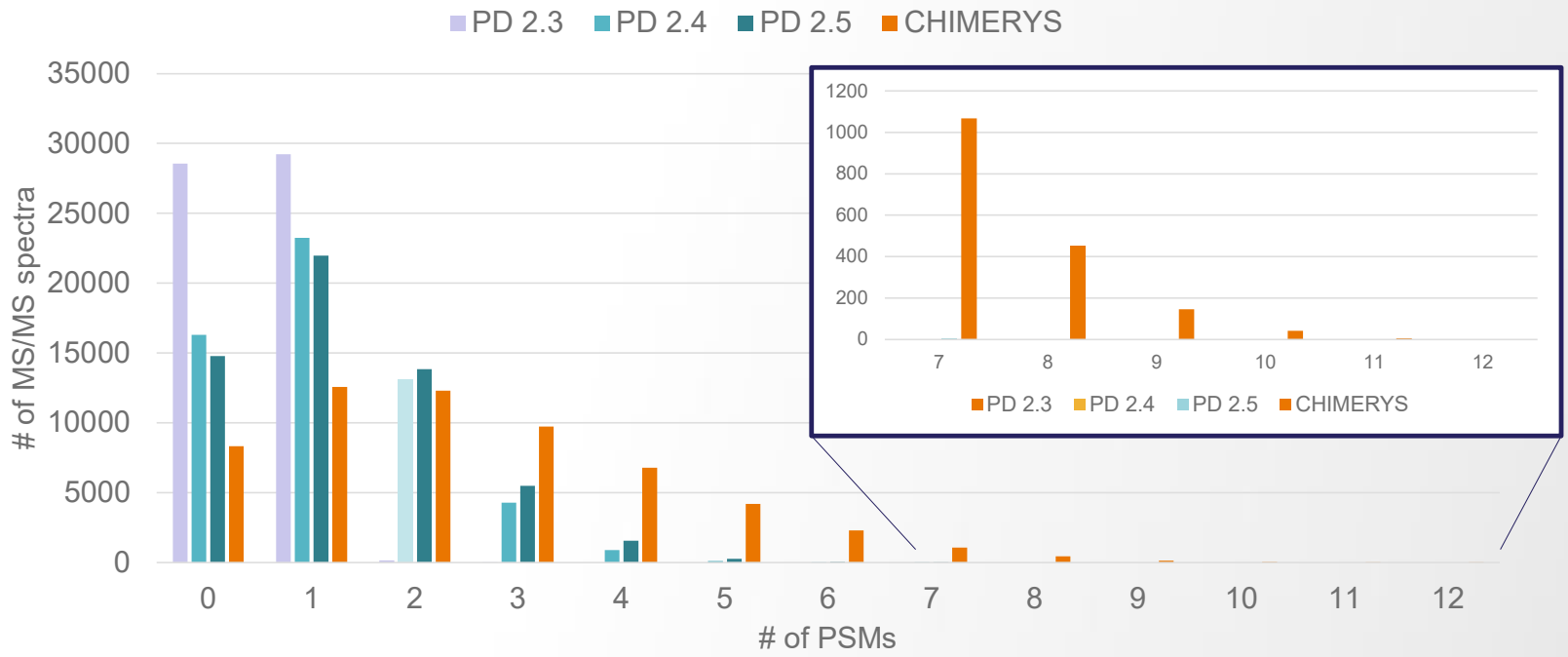
CHIMERYS identifies 5 peptides vs INFERYS Rescoring identifies 1





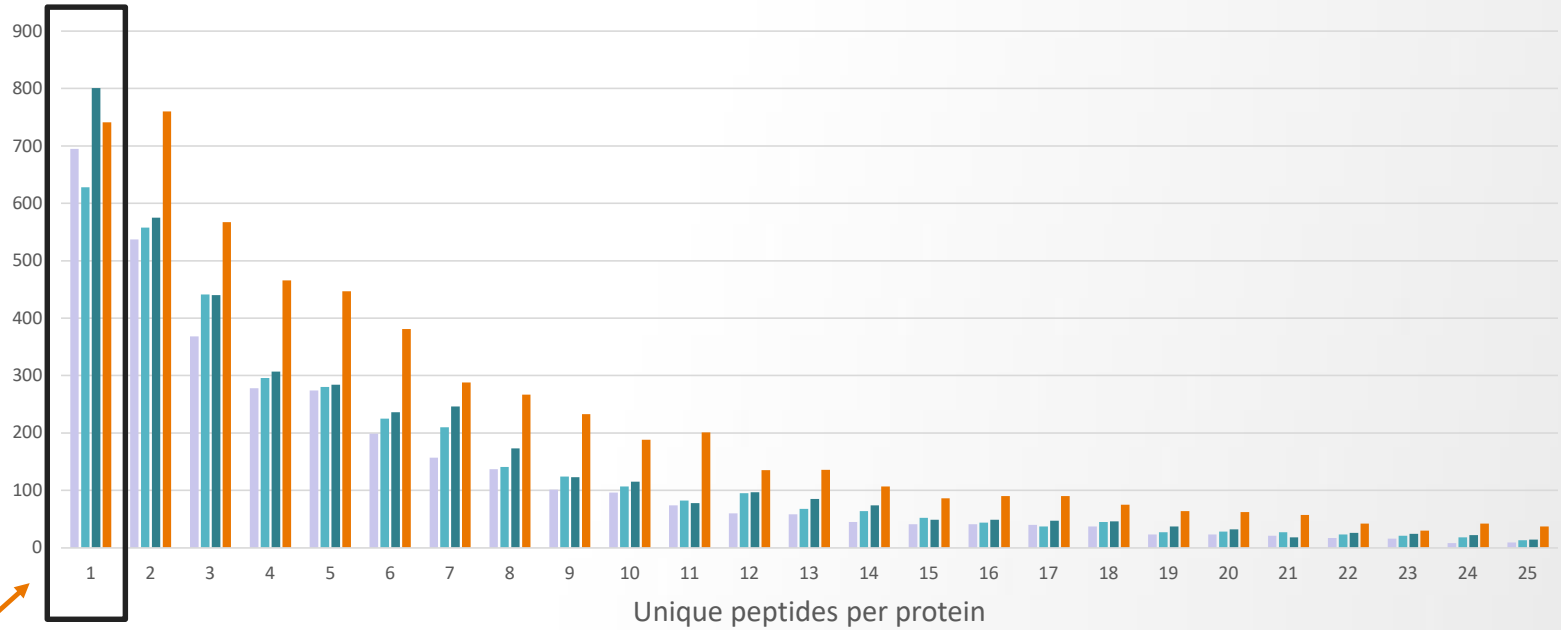
CHIMERYS identifies up to 12 peptides per spectrum

For 3.8 m/z isolation window data, over 8000 spectra have 5 or more PSMs



No increase in “one hit wonders”

CHIMERYS produces more sequence coverage than previous versions



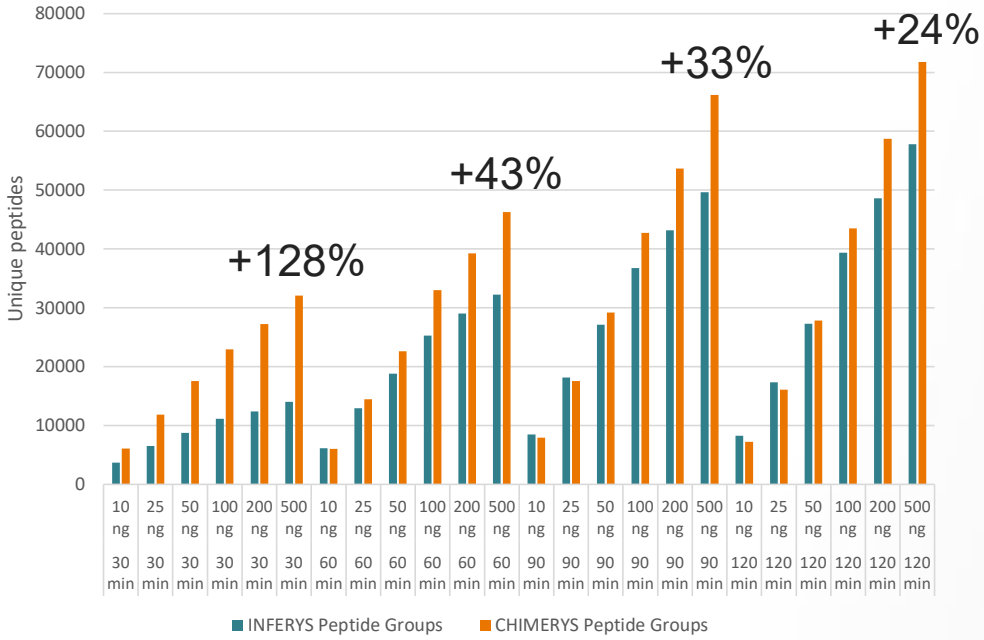
Proteins with one unique peptide

PD 2.3 PD 2.4 PD 2.5 CHIMERYS

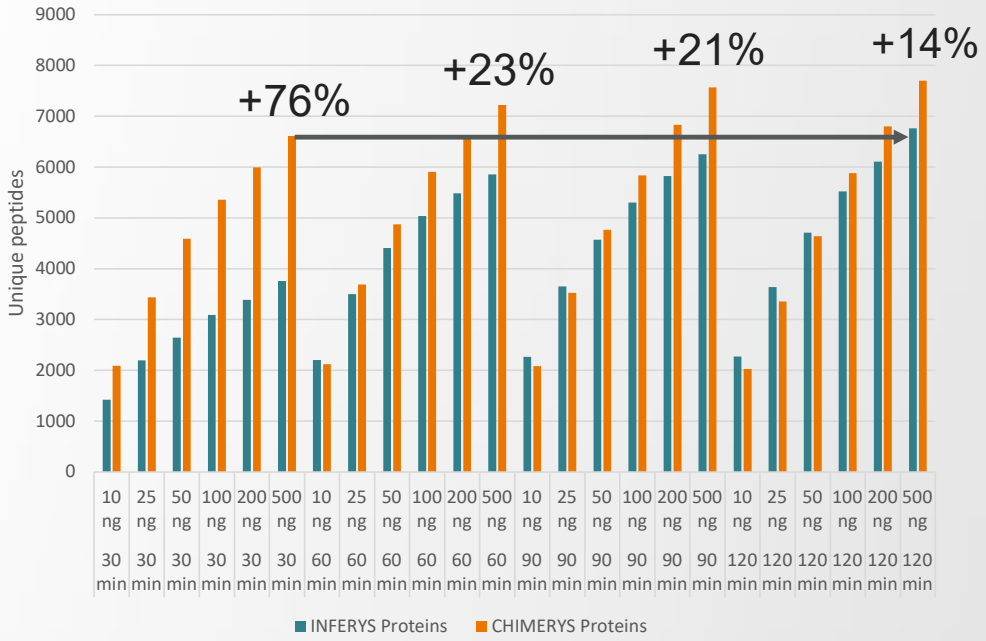
Avg peptides/protein:	6.8	7.4	7.4	**8.6**

CHIMERYS vs Sequest HT + INFERYS

Thermo Scientific Orbitrap Exploris™ 480 MS with FAIMS Pro™ Interface “golden data”



Unique peptides



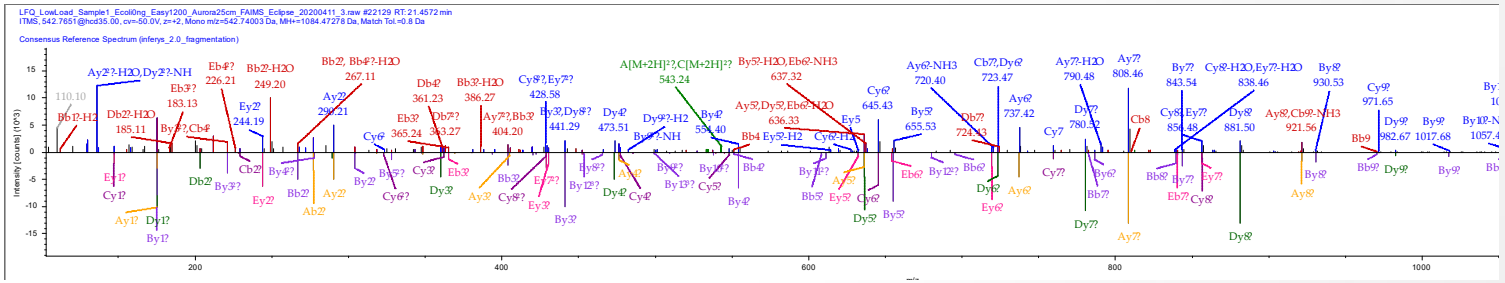
Proteins

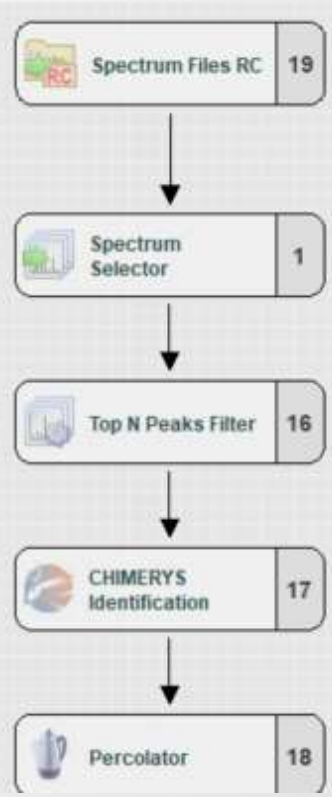
OT/IT support

CHIMERYS supports ion trap MS/MS data

- “Top N” Peaks Filter is used to remove background noise prior to CHIMERYS search

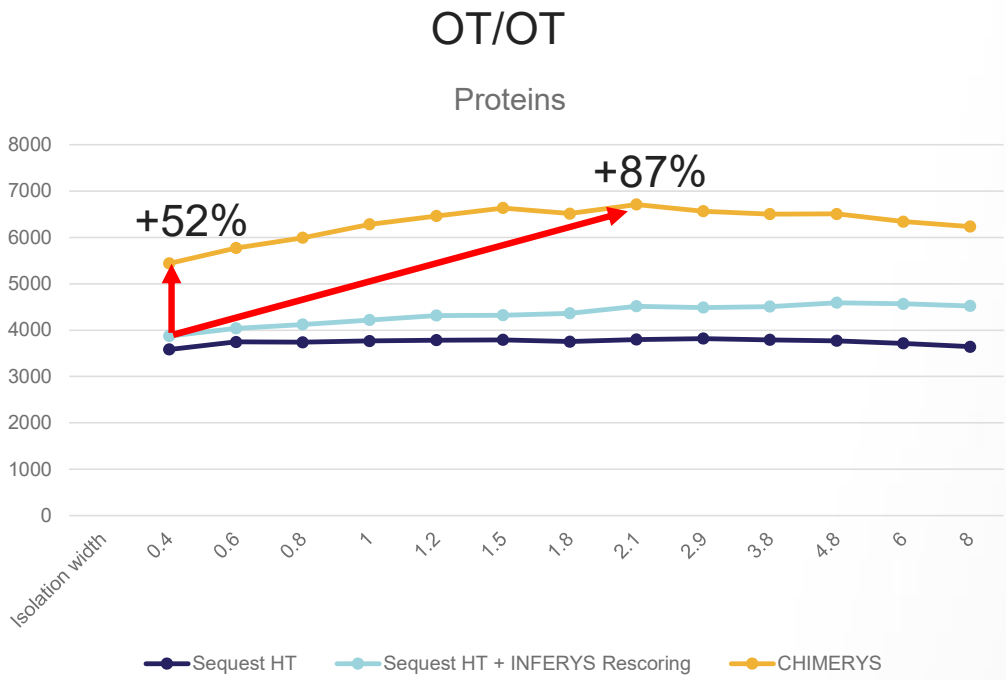
- 1 ug 60min HeLa dataset from OT Ascend Tribrid:
  - Sequest HT INFERYS: 5496 protein groups, 37351 peptides
  - CHIMERYS: 7141 protein groups (+30%), 59132 peptides (+58%)
  - CHIMERY quantified 42973 peptides (+32%) vs 32533 for Sequest+INFERYS
  - Sequest HT INFERYS+MSPepSearch: 5714 protein groups (+4%), 47490 peptides (+27%)

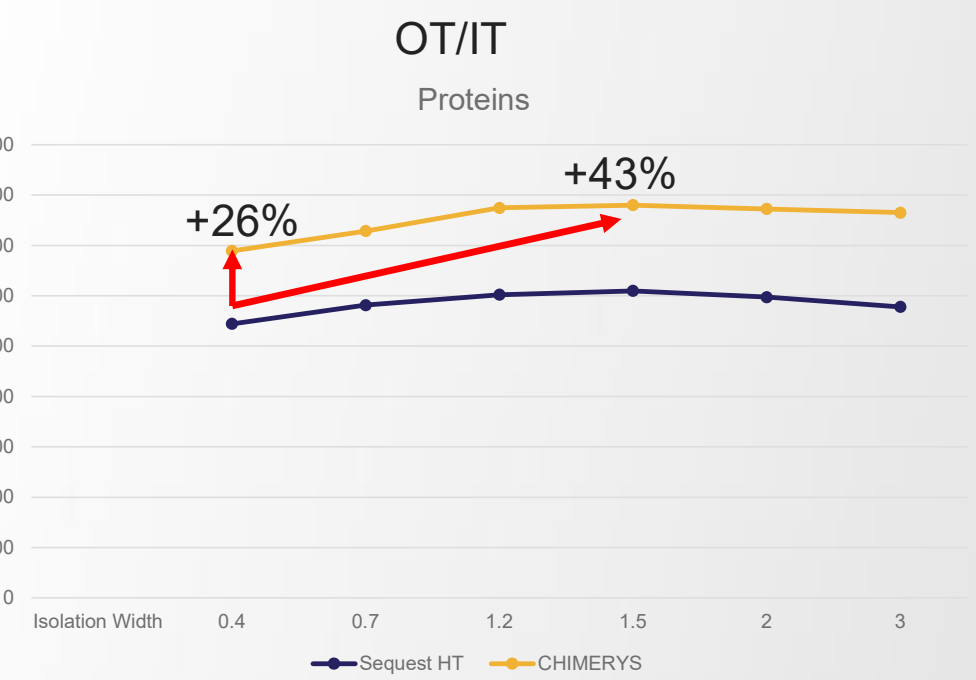




Not just software – Wide Window Acquisition

Optimize data-dependent acquisition methods with Proteome Discoverer 3.0 software and CHIMERYS



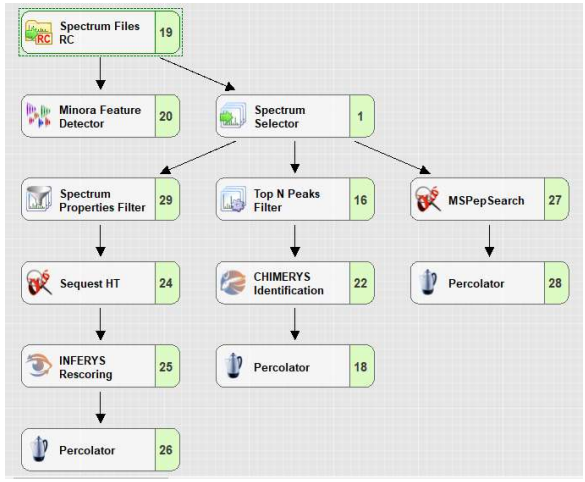


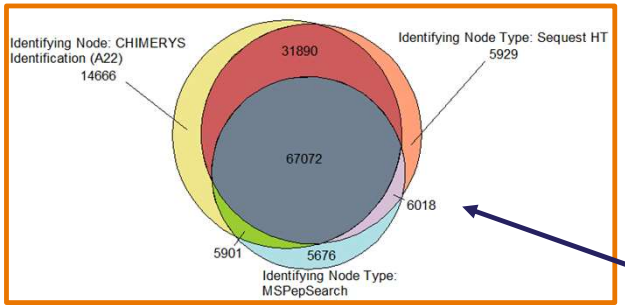
See poster from ASMS

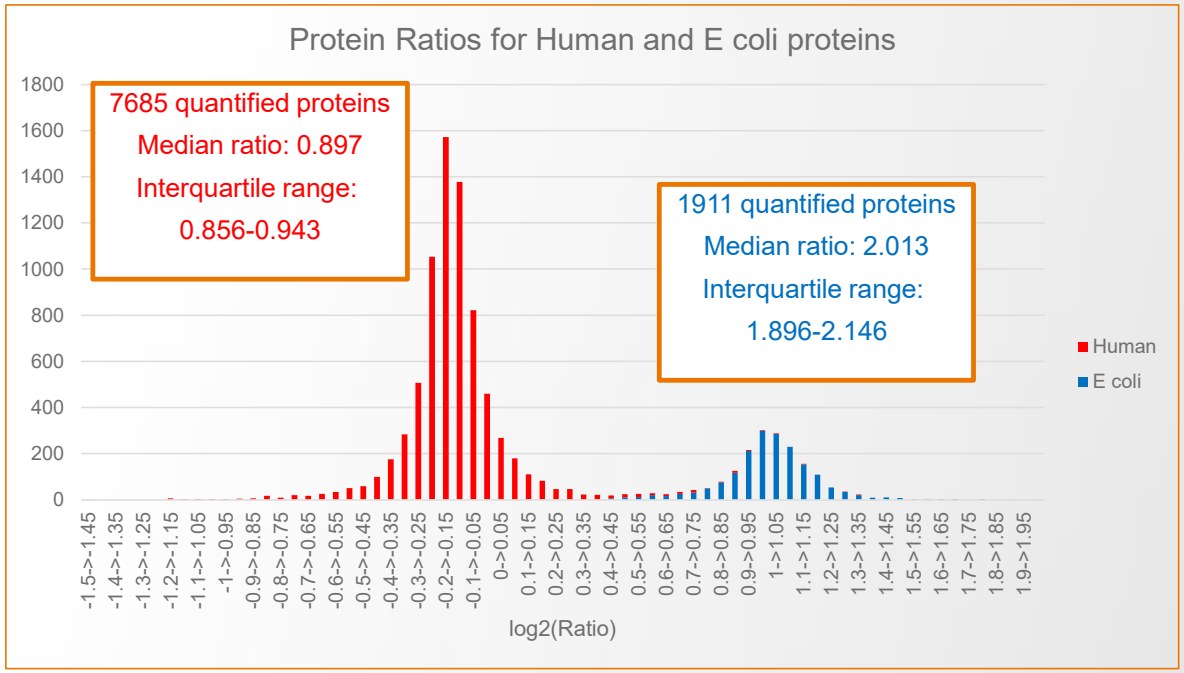
HeLa cell lysate digest acquired on a Thermo Scientific™ Orbitrap Eclipse™ Tribrid™ mass spectrometer using a 1-hour gradient and MS2 isolation windows between 0.4 Th and 8 Th

Label-free quantification of Human and E coli Mixture

OT/IT HCD MS/MS, 1 ug total protein per sample, 3 search engine strategy



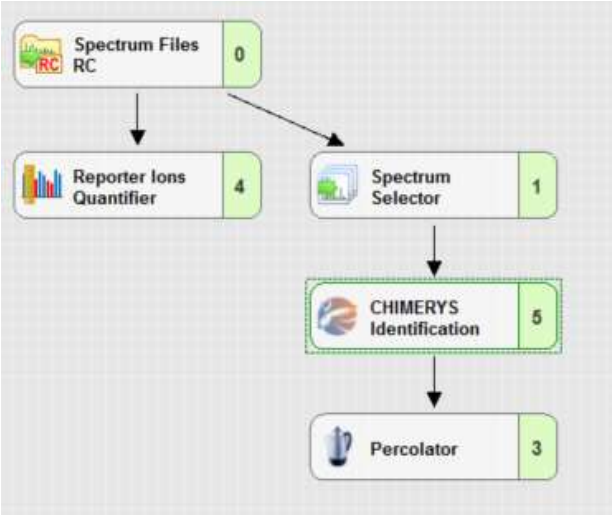




146069 identified peptides

CHIMERYS TMT MS2 Processing workflow

Can CHIMERYS be used to improve TMT quantification results?

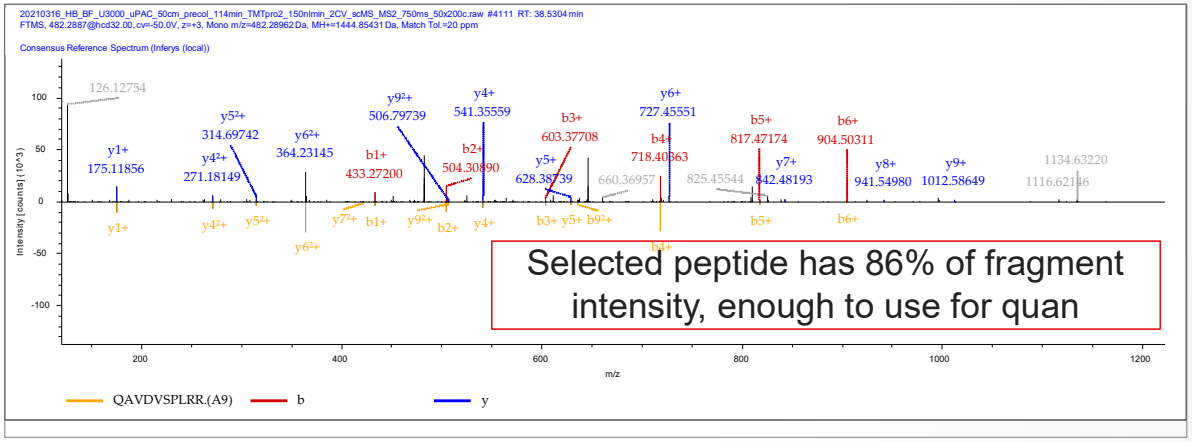


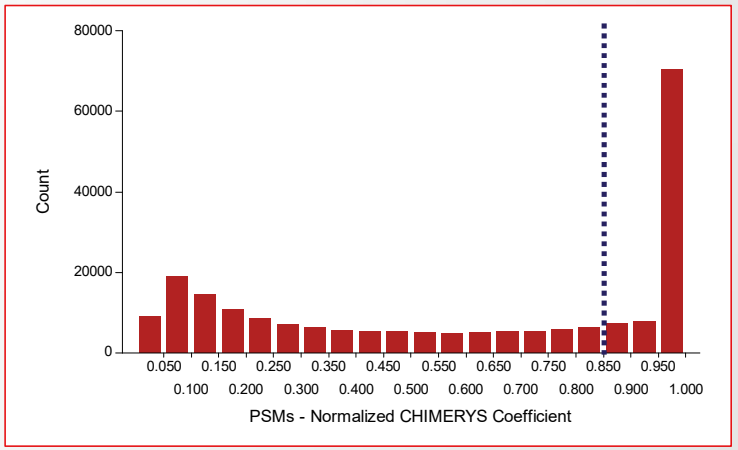
Show Advanced Parameters	
1. Input Data	
Protein Database	Saccharomyces cerevisiae (sp_canonical
Prediction Model	inferys_2.1_fragmentation
Enzyme	Trypsin
Max. Missed Cleavage Sites	2
Min. Peptide Length	7
Max. Peptide Length	30
Max. Number of Modifications per	3
Min. Peptide Charge	2
Max. Peptide Charge	4
2. Tolerances	
Fragment Mass Tolerance	20 ppm
3. Dynamic Modifications	
1. Dynamic Modification	Oxidation (M)
2. Dynamic Modification	
3. Dynamic Modification	
Peptide N-Terminus	
4. Static Modifications	
Peptide N-Terminus	TMT6/10/11 Tandem Mass Tag (N-term)
1. Static Modification	Carbamidomethyl (C)
2. Static Modification	TMT6/10/11 Tandem Mass Tag (K)
3. Static Modification	
5. Validation	
Percolator Mode	Auto

- User selects
  - Protein database
  - PTMs
  - Mass tolerance (20 ppm for OT MS/MS, 0.3 Da for IT MS/MS)
  - “Top N Peak Filter” required for optimal OT/IT analysis

New Normalized CHIMERYS Coefficient parameter

- Normalized CHIMERYS coefficient parameter in Reporter Ions Quantifier in Consensus workflow
  - Parameter measures percentage of fragment abundance for selected PSM compared to all in same MS/MS spectrum
  - Coefficient >0.8 considered acceptable





Example: SCP TMT paper

- “Real-Time Search-Assisted Acquisition on a Tribrid Mass Spectrometer Improves Coverage in Multiplexed Single-Cell Proteomics”, Fürtwangler *et al*, MCP, **2022**.

- Samples were labeled with TMTpro 16plex
- Acquired data via multiple acquisition methods:
  - MS2-based TMT
  - SPS MS3
  - RETICLE – real-time search-triggered MS2 ID and quan

- Compare workflows from PD 2.5 to new approaches in PD 3.0 for all 3 acquisition methods

INFERYS Rescoring results

Comparison of PD 2.5 and PD 3.0 results

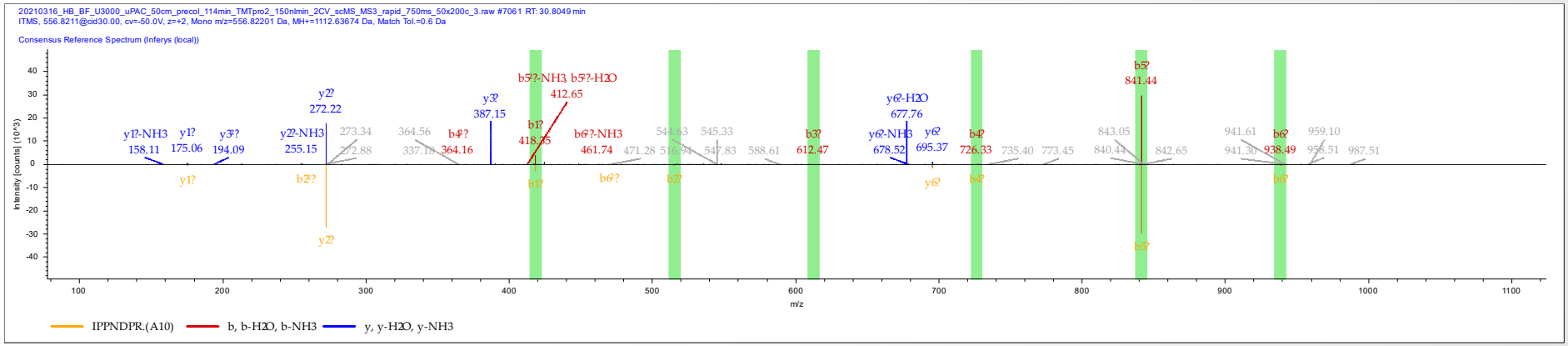
Acquisition Method	PD 2.5 Sequest HT					PD 3.0 Sequest HT + INFERYS Rescoring (+ Comet for RETICLE)				
	Protein Groups	Quantified Proteins	Peptide Groups	Quantified Peptides	PSMs	Protein Groups	Quantified Proteins	Peptide Groups	Quantified Peptides	PSMs
HCD MS2	1135	1081	**4073**	**3854**	9124	**1150**	**1085**	4008	3769	8985
SPS MS3	1347	1111	4863	3490	10461	**1667 (+23%)**	**1229 (+10%)**	**6164 (+27%)**	**4086 (+17%)**	**13205**
RETICLE	1539	1417	5307	4832	10577	**1591 (+3%)**	**1472 (+4%)**	**5437 (+2%)**	**4901 (+2%)**	**19347**

- **Incremental improvement** in results for high resolution HCD
- **Substantial improvement** in results for SPS MS3 data.
- **Incremental improvement** for RETICLE results.

Example of a peptide unique to INFERYS Rescoring

No high confidence identification for Sequest HT alone

- (TMTpro)-IPPNDPR, 2+
- Xcorr of 1.29 due to not many matched fragments
- Predicted spectrum in mirror plot correlates experimental abundances to predicted ones
- All fragments selected for MS3 are from identified peptide, ensuring accurate quantification



MS2-based TMT results

CHIMERYS produces the most differentially expressed proteins for single cell data

Search strategy	Protein groups	Peptide groups	PSMs	Quantified peptides	Quantified proteins
Sequest HT	1135	4073	9124	**3854**	1081
Sequest HT + INFERYS Rescoring	1150	4008	8985	3769	1085
CHIMERYS	**1658 (+44%)**	**5037 (24%)**	**11021**	3518	**1168 (+8%)**

- CHIMERYS produces a substantial gain in the number of protein and peptide identifications.
- Although many identified peptides are rejected due to interferences, CHIMERYS is still able to quantify proteins not identified by other approaches.
- **CHIMERYS is the best approach for MS2-based TMT quantification.**

Single cell SPS MS3 results

Sequest HT + INFERYS Rescoring is still the best approach for ion trap MS/MS-based methods

Search strategy	Protein groups	Peptide groups	PSMs	Quantified peptides	Quantified proteins
Sequest HT	1347	4863	10461	3490	1111
Sequest HT + INFERYS Rescoring	**1667**	**6164**	**13205**	**4086**	**1229**
CHIMERYS	1632	5824	12144	3662	1169

- **INFERYS Rescoring is the best approach for SPS MS3 acquisition.**

Single cell RETICLE results

RETICLE results behave similarly to MS2-based TMT

Search strategy	Protein groups	Peptide groups	PSMs	Quantified peptides	Quantified proteins
Sequest HT	1539	5307	10577	4832	1417
Sequest HT + INFERYS Rescoring + Comet	1591	5437	19347	**4901**	**1472**
CHIMERYS	**2025**	**5892**	11078	3869	1322

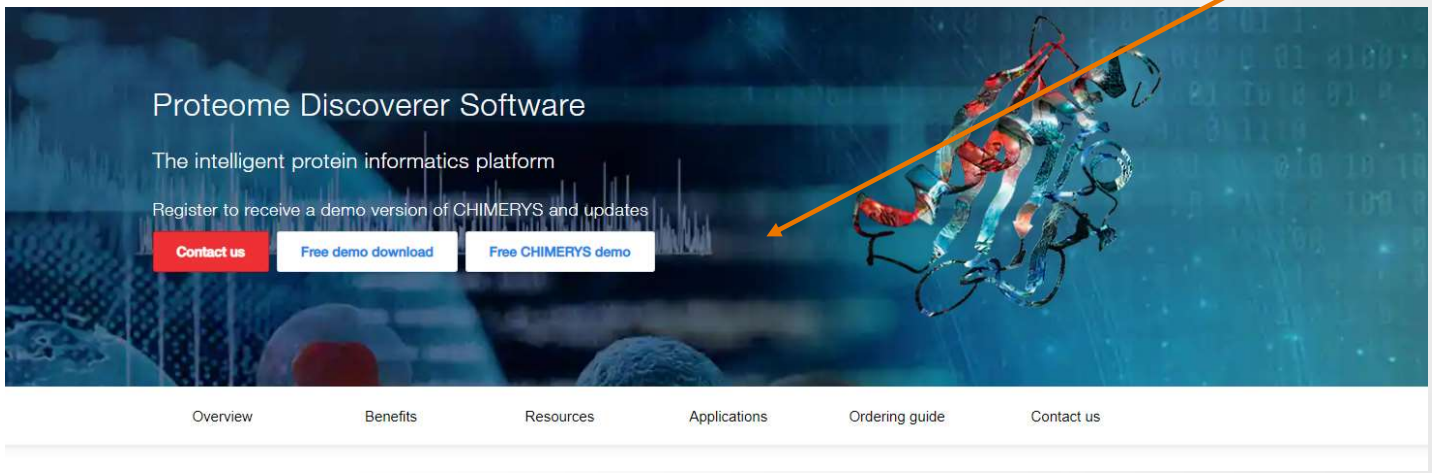
- CHIMERYS produces substantially more protein IDs, but fewer are quantified due to CHIMERYS coefficient <0.8.
- **For RTS-based approaches, Sequest HT + INFERYS Rescoring + Comet produces the best results.**

How can I try CHIMERYS?

Visit Proteome Discoverer product page

- https://www.thermofisher.com/us/en/home/industrial/mass-spectrometry/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-software/multi-omics-data-analysis/proteome-discoverer-software.html
- Click “Free CHIMERYS demo” and fill out form
- We will send you a temporary account to search 5-10 data files

Link to demo request sends e-mail chimerys.demo@thermofisher.com



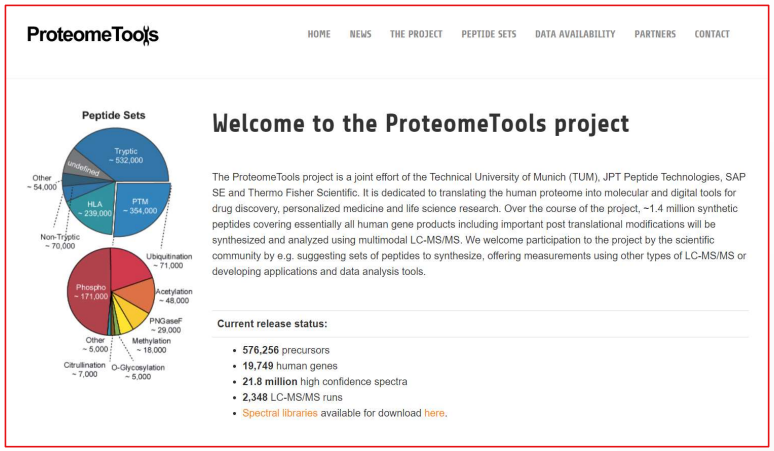
Questions

thermofisher.com/proteomediscoverer

INFERYS Prediction and Rescoring

INFERYS is developed using Proteome Tools and other curated data

- Deep learning-based MS/MS spectral prediction
- Developers of INFERYS at MSAID™ were part of the ProteomeTools and Prosit projects with the Kuster lab
- Many years of experience with deep learning and algorithm development



What is INFERYS?

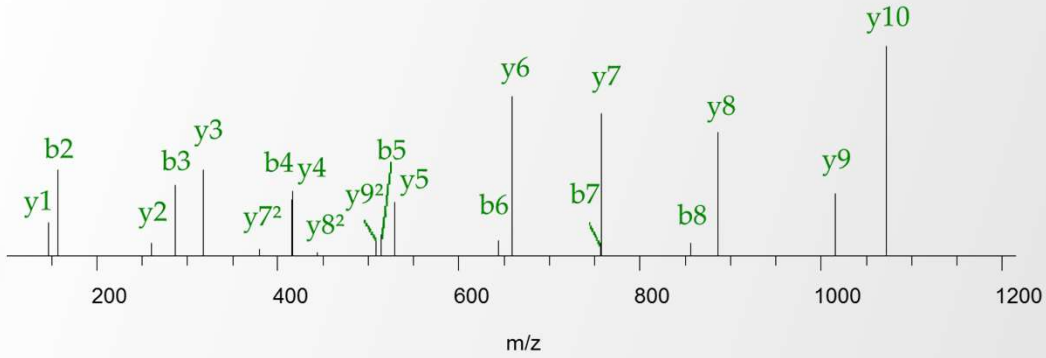
Deep learning-based model for peptide MS/MS spectral prediction

VGEEVEIVGIK

z = 2+

NCE = 26

INFERYS



Predicted MS/MS spectrum

INFERYS® Rescoring Node

Additional features calculated as input into Percolator

- Search against FASTA file
- Produce a list of PSMs from the target and decoy database searches
- Produce up to rank 10 hits per spectrum

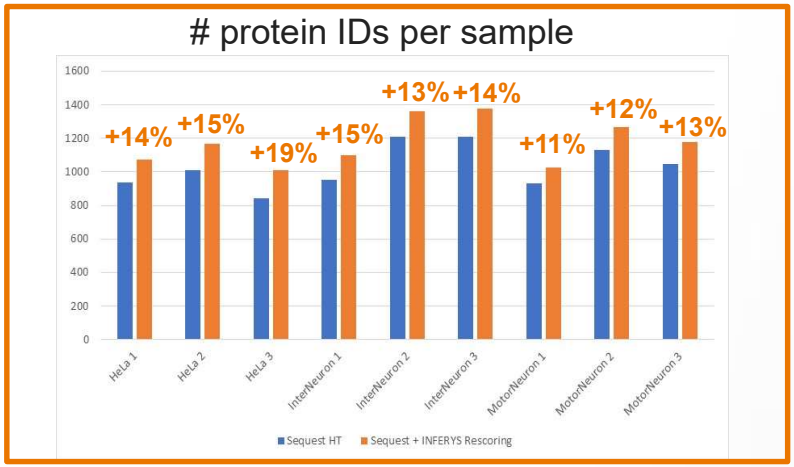
- Determine optimal collision energy for input into INFERYS algorithm
- Calculate all predicted spectra for all target and decoy hits
- Calculate spectral angle and other figures-of-merit for correlation of predicted and measured spectra

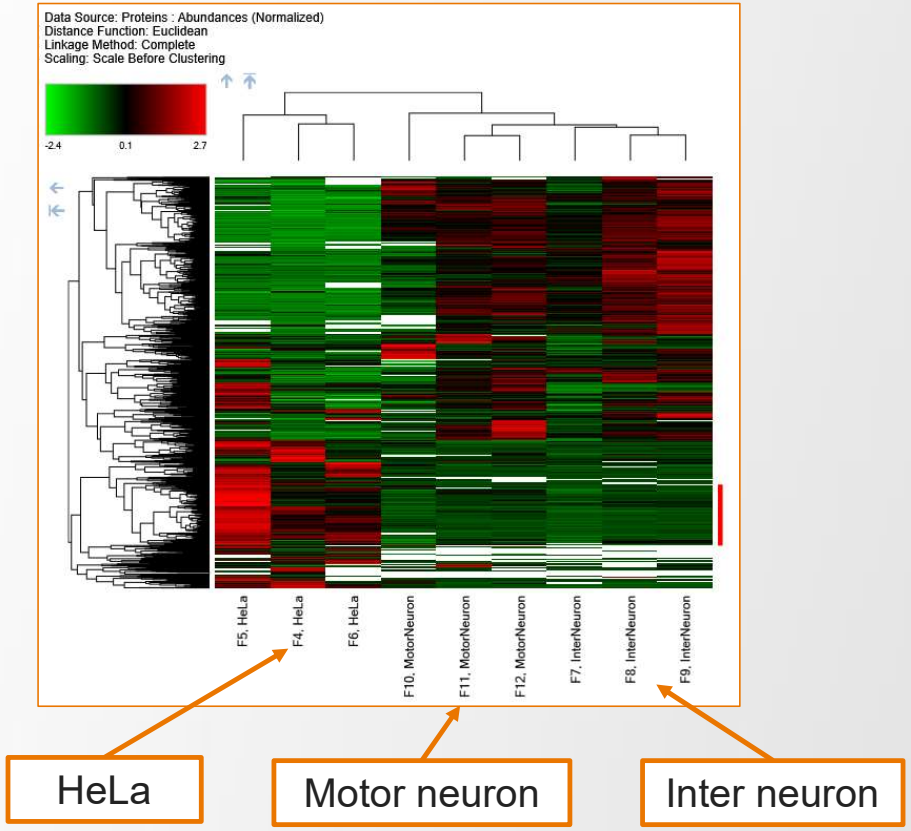
- Use the standard 35 features
- If INFERYS Rescoring node is used, add 15 more features
- Calculate FDR thresholds using SVM score calculated from all input features
- For information about Percolator, read Matthew The *et* al, JASMS 2016

Single cell analysis using INFERYS Rescoring

Yongzheng Cong *et al*, Chem Sci, 2021 Advance article, FAIMS 2 CV (-55, -70)

- Thermo Scientific™ FAIMS Pro™ interface with Thermo Scientific™ Orbitrap Eclipse™ Tribrid™ Mass Spectrometer, IT HCD MS/MS
- With INFERYS Rescoring
  - **13-19%** increase in identified and quantified proteins
  - **20-24%** increase in identified and quantified peptides

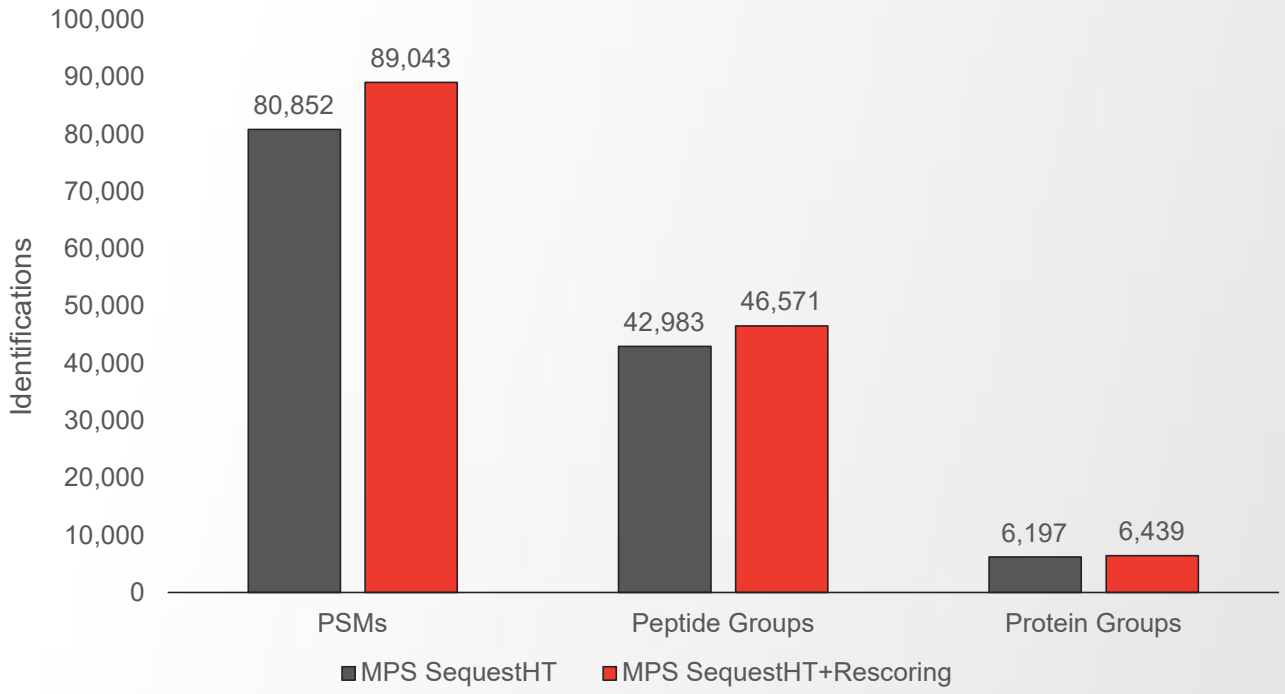




Results – Standard bottom-up proteomics data set

200 ng HeLa Thermo Scientific™ Obitrap Exploris™ 480 MS with FAIMS CV50 CV70

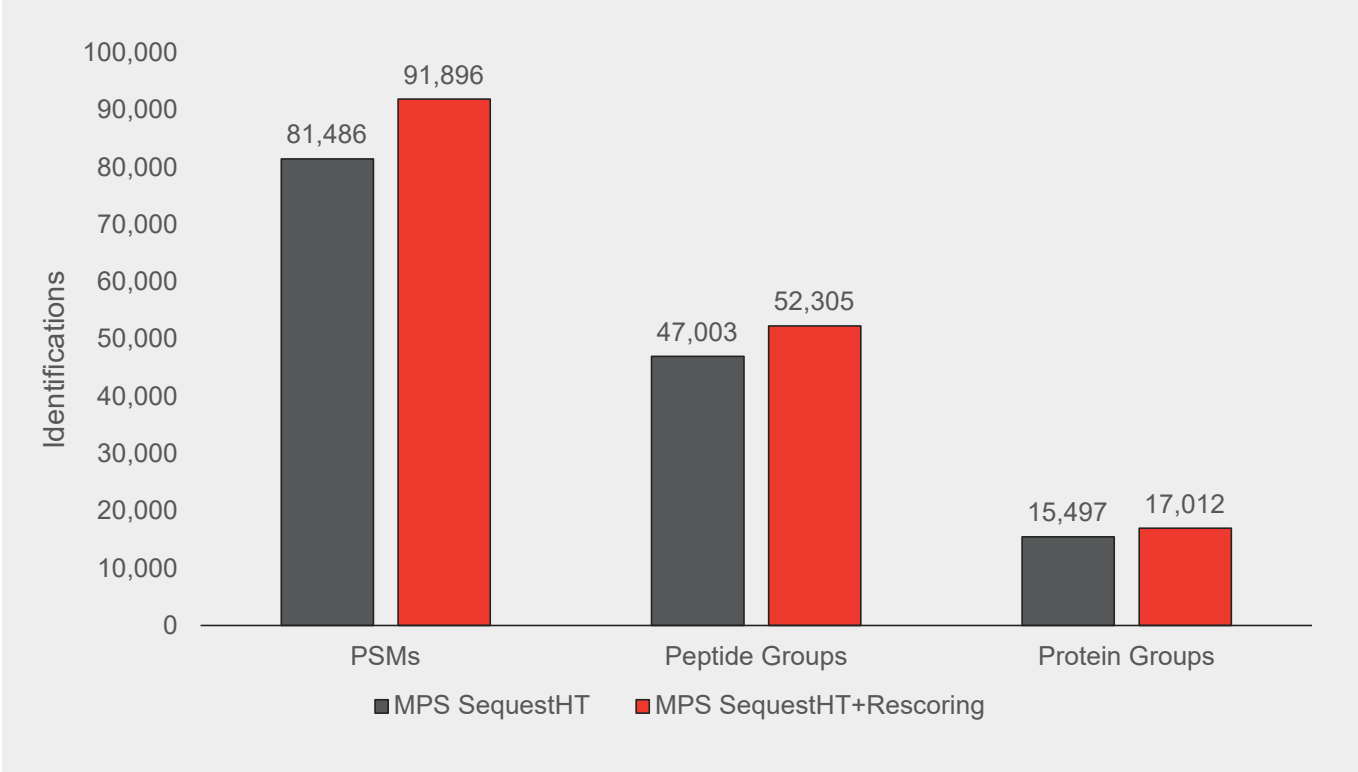
- Standard tryptic HeLa digest dataset
- 10% increase in identifications at PSM, 8% at unique peptide and 4% at protein level
- Rescues short peptides and peptides where classical scores fail to separate targets from decoys



Results – Standard bottom-up proteomics data set

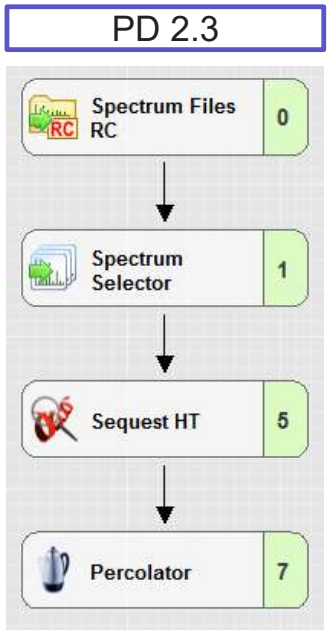
Metaproteomics – Human stool samples (*Rechenberger et al., Proteomes. 2019)*

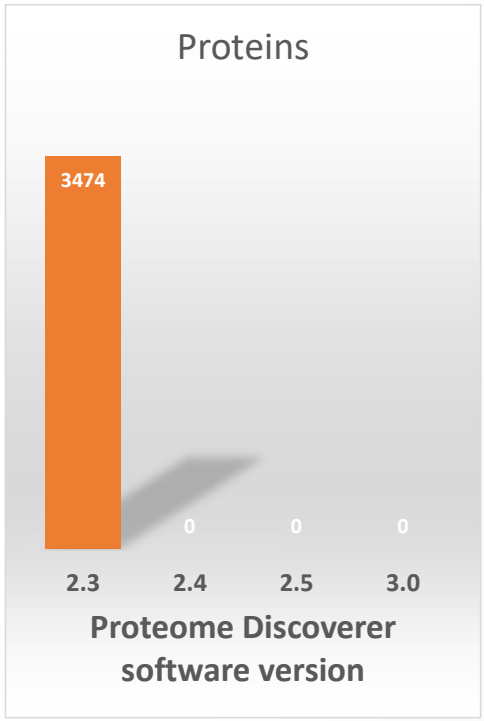
- Trypsin digest
- Complex sample with huge search space
- 13% increase in identifications at PSM, 11% at peptide and 10% at protein level
- Rescues short peptides and peptides where classical scores fail to separate targets from decoys

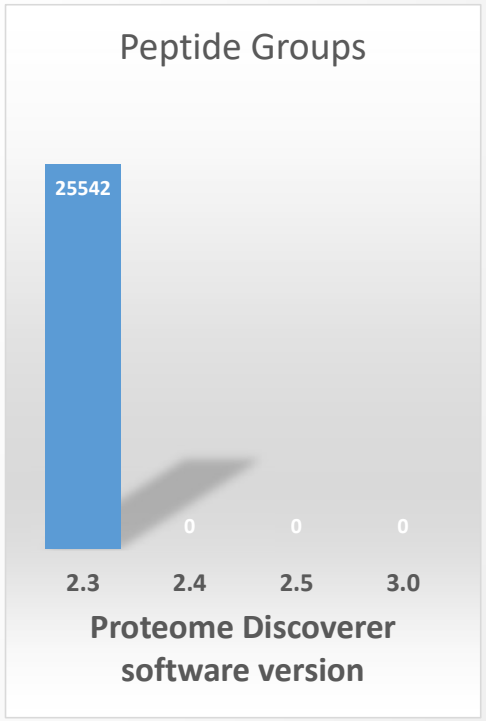


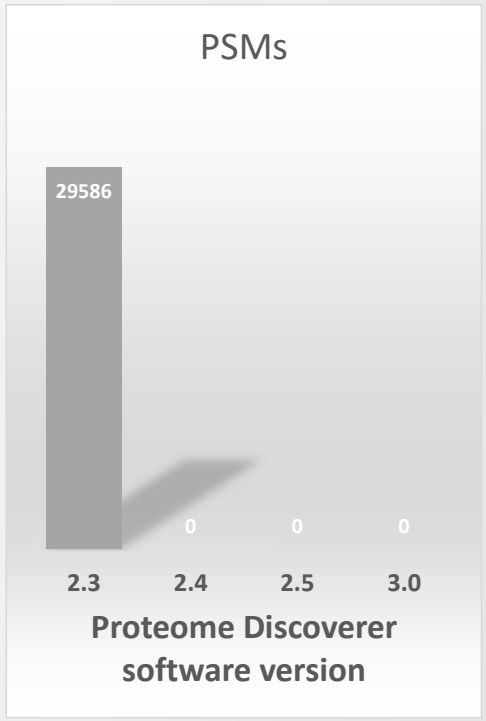
Improvement in peptide ID by release

HeLa dataset, 90-minute run, ID counts by Proteome Discoverer software release version



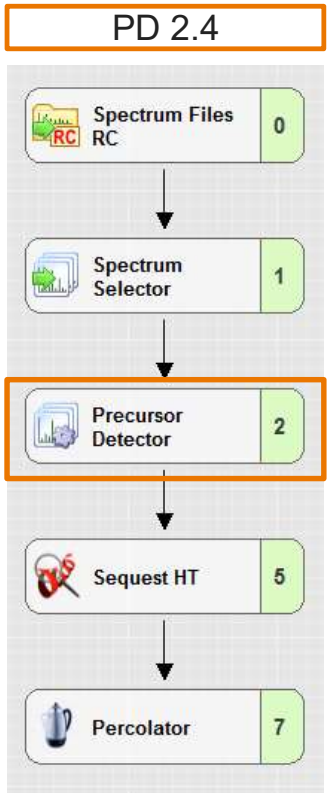


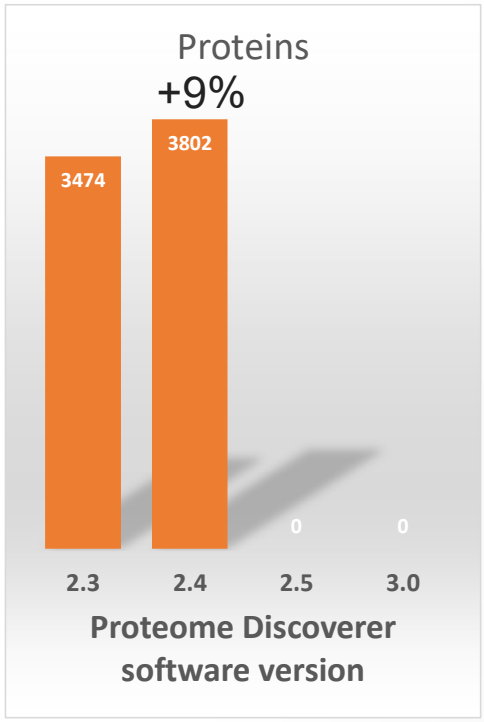




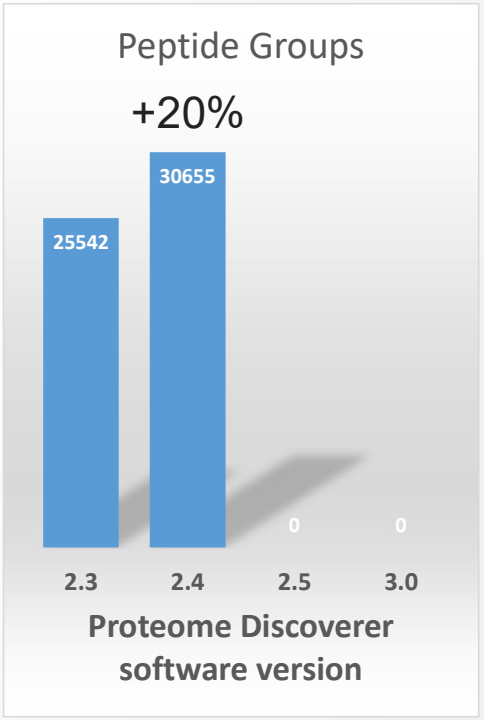
Improvement in peptide ID by release

HeLa dataset, 90-minute run, ID counts by Proteome Discoverer software release version

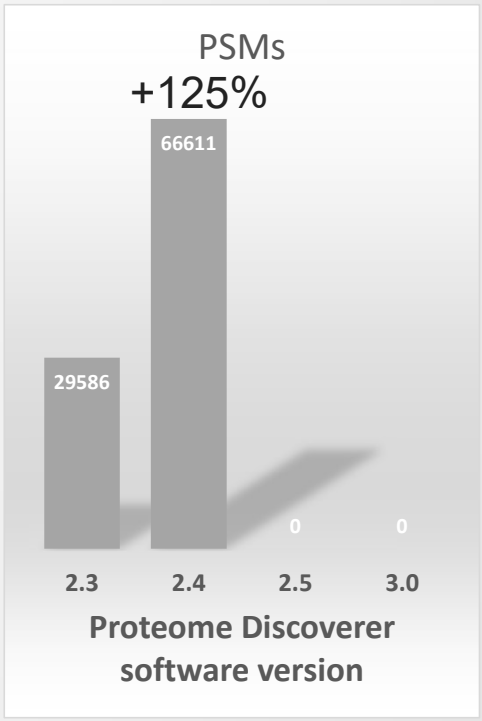




PD 2.4 produces 9% more proteins than PD 2.3



PD 2.4 produces 20% more proteins than PD 2.3

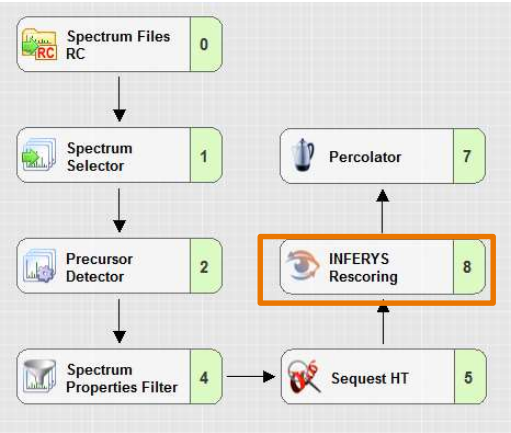


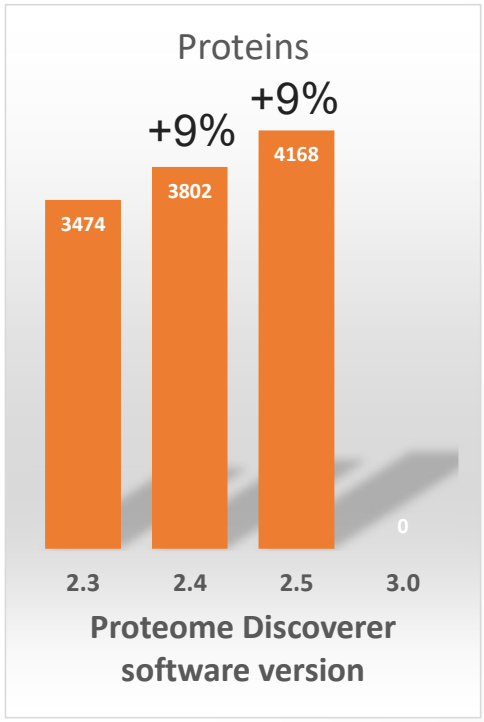
PD 2.4 produces 44% more PSMs than PD 2.3

Improvement in peptide ID by release

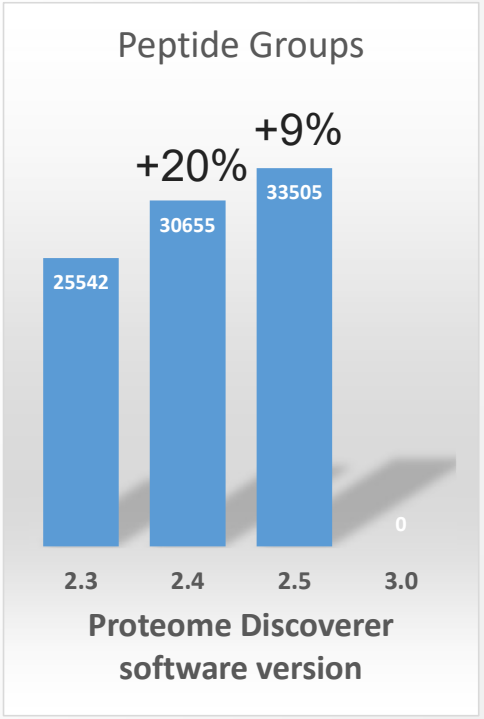
HeLa dataset, 90-minute run, ID counts by Proteome Discoverer software release version

PD 2.5

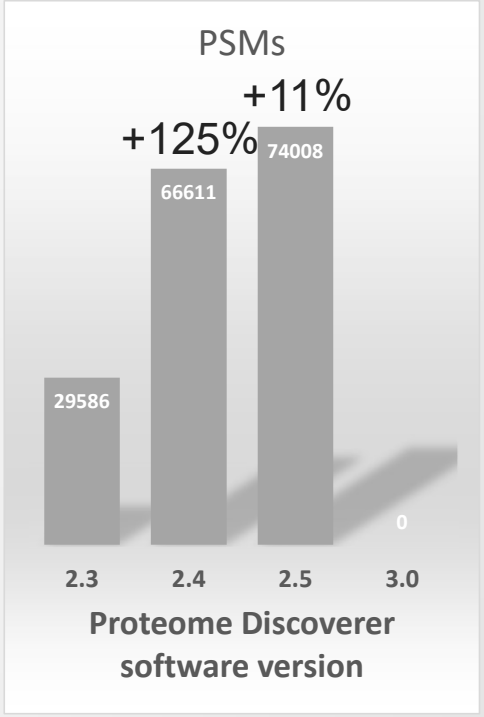




PD 2.5 produces 20% more proteins than PD 2.3



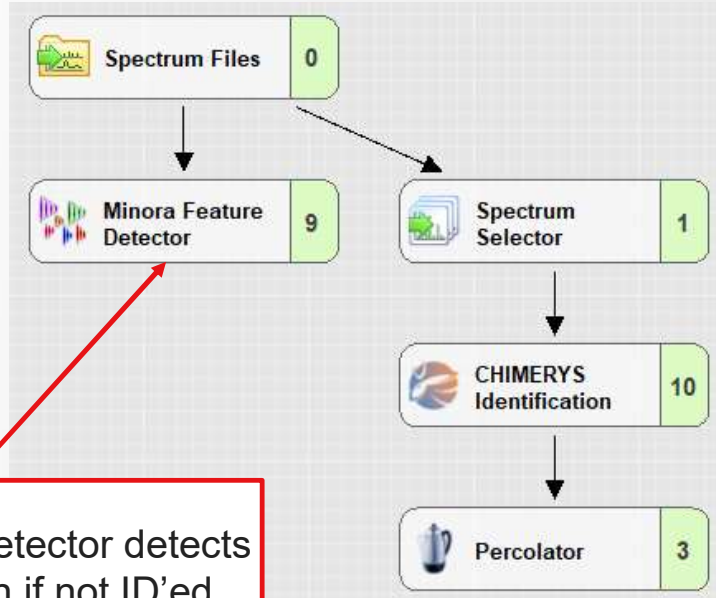
PD 2.5 produces 31% more unique peptides than PD 2.3



PD 2.5 produces 2.5x as many PSMs as PD 2.3

Wide Window Acquisition for LFQ

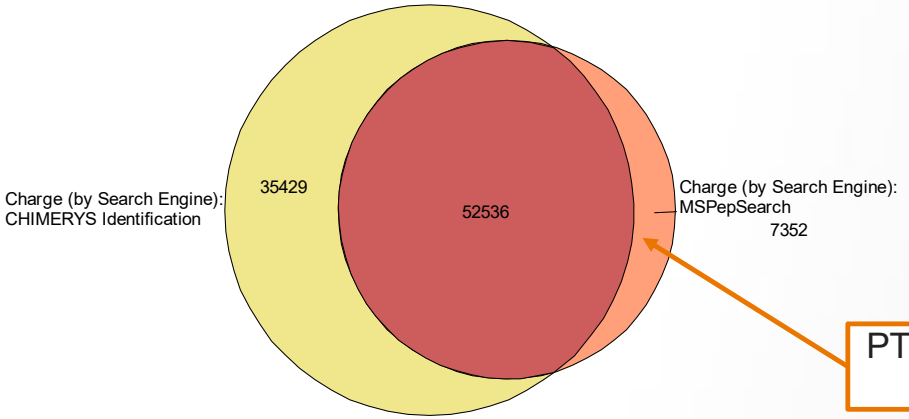
- Minora algorithm already detects >100,000 features in the MS1 data
- CHIMERYS helps explain more of the MS1 scan information, leading to significant increases in number of quantified peptides and proteins
- Wide window acquisition: 4 *m/z* for OT/OT and 1.5 *m/z* for OT/IT improve identification and quantification results



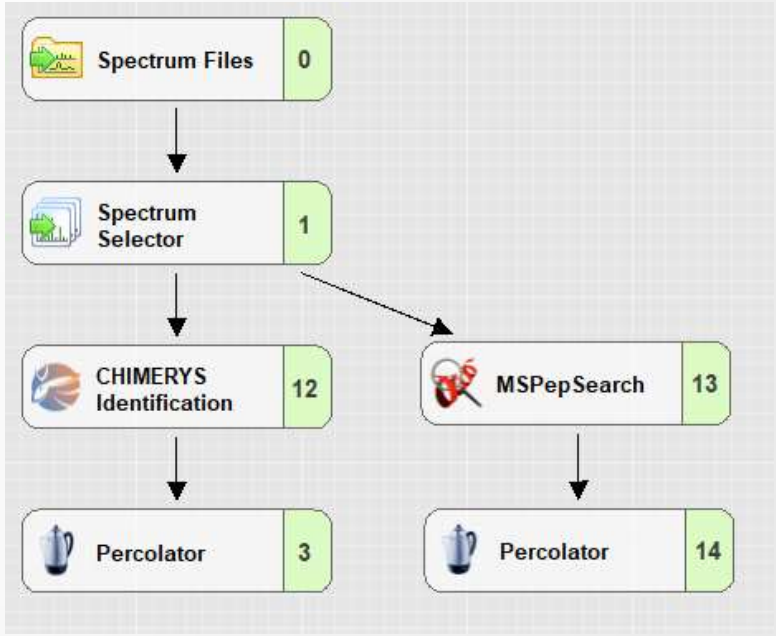
Combining CHIMERYS with other search engines

Increase peptides even further by using library search or other engines specializing in PTMs

- Use other search engines such as Sequest HT and MSPepSearch to find peptides not supported by INFERYS prediction models
- Can also combine with multi-stage searches to help maximize IDs



PTMs, semitryptic peptides



RTS TMT18plex results

Journal of Proteome Research, *2021*, 20, 2964-2972

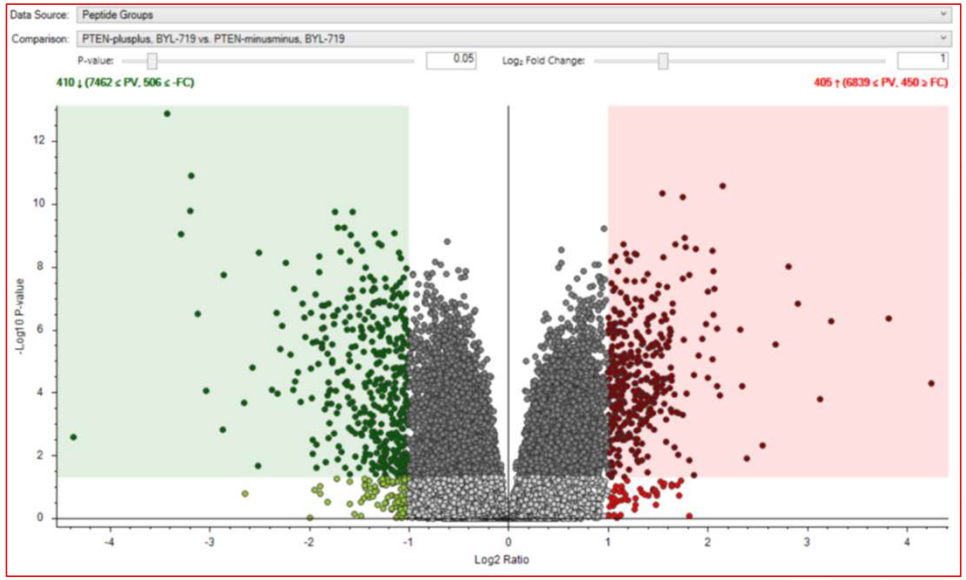
Data available on PRIDE (PXD024275)

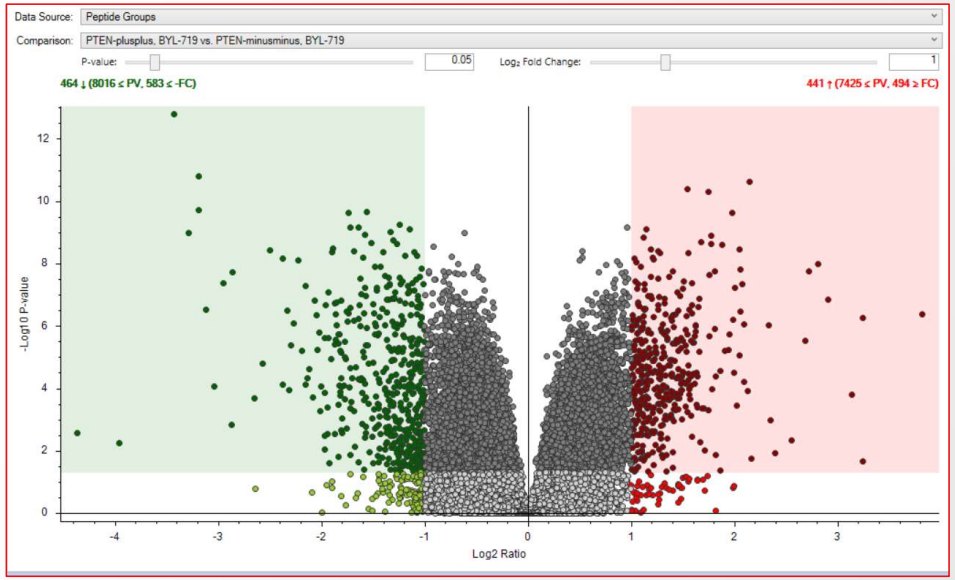
Search strategy	Protein groups	Peptide groups	PSMs
Sequest HT	9159	92275	170820
Sequest HT + INFERYS Rescoring + Comet	**9455 (+3.1%)**	**101594 (+8.4%)**	**347331**

- Sequest HT + INFERYS Rescoring + Comet workflow for real-time search analysis leads to increases in Peptide and Protein Group IDs
- The increase in identified peptides and proteins also produce an increase in the number of differentially expressed peptides and proteins

RTS TMT 18plex quan results

Peptide group ratios for PTEN++/PTEN– for BYL-716 stimulation



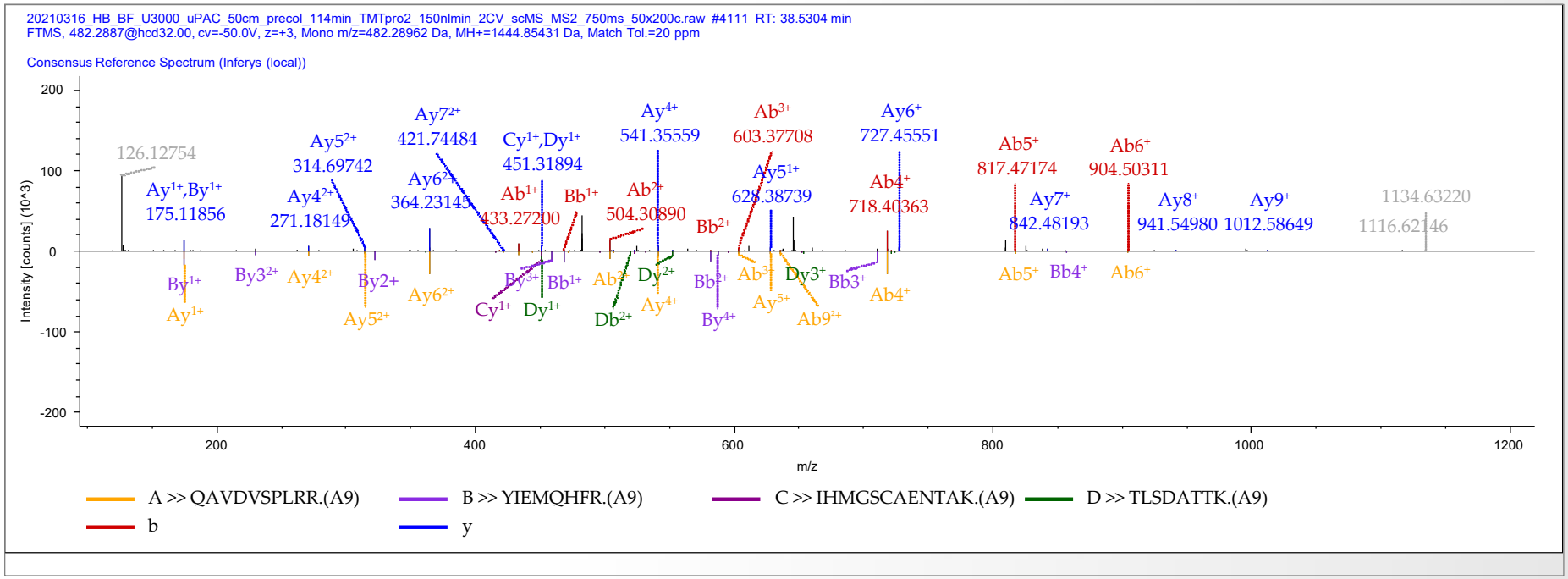


Sequest HT:
410 down-regulated,
405 up-regulated

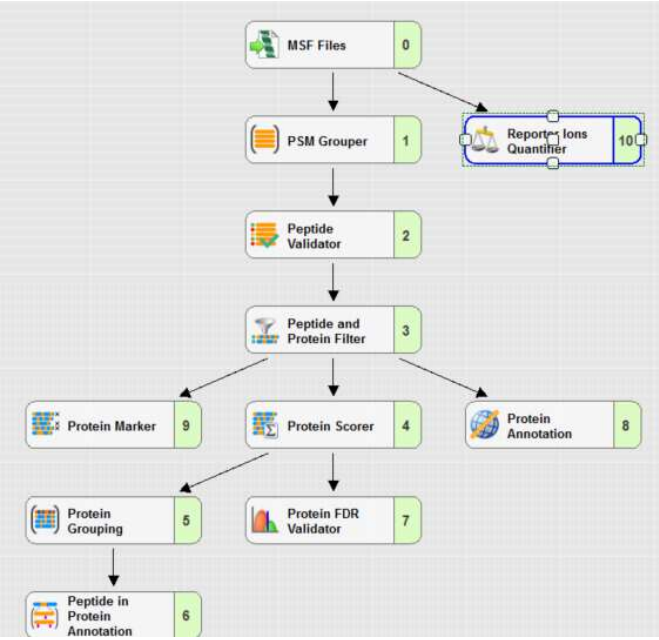
INFERYS Rescoring + Comet:
464 down-regulated,
441 up-regulated

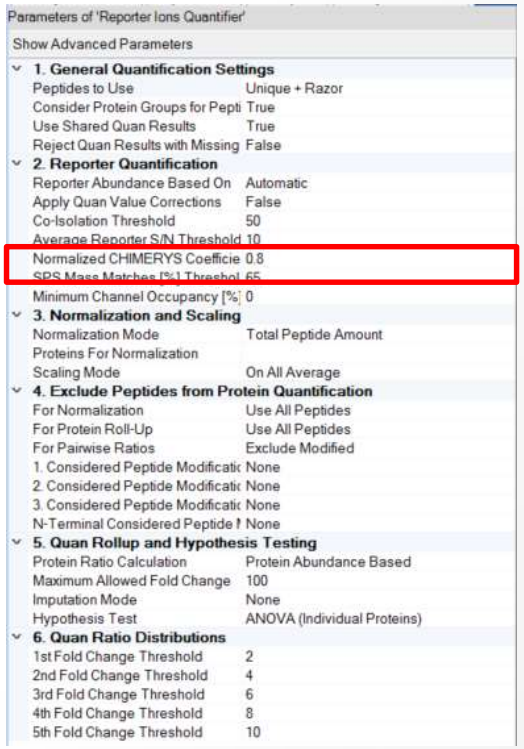
CHIMERYS result for TMT quantification dataset

Chimeric spectrum with 4 PSMs – are any peptides usable for quan?



CHIMERYS Coefficient parameter in Consensus workflow



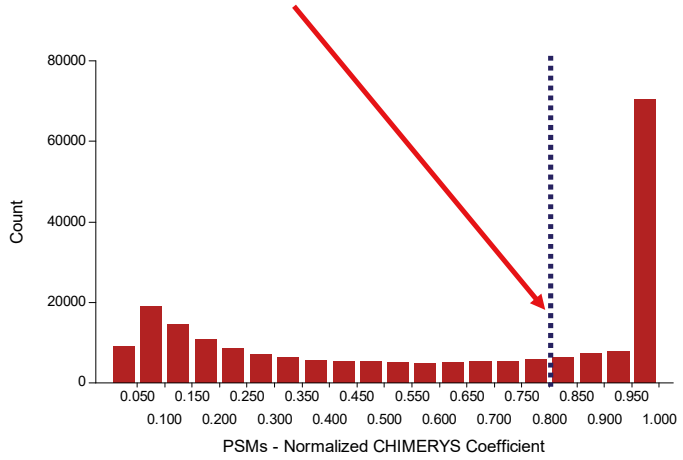


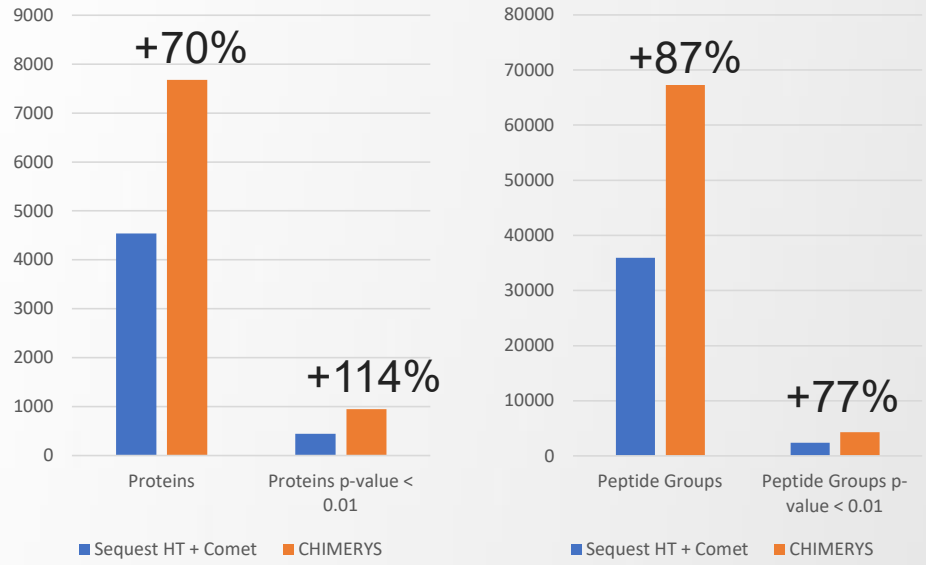
- New workflow “CWF_Comprehensive_Enhanced Annotation_Reporter_Quan_MS2.pdConsensusWF” includes a default of 0.8 for Normalized CHIMERYS Coefficient”
- “_MS3” version of Consensus workflow sets this value to 0. SPS MS3 threshold is sufficient even for CHIMERYS results

CHIMERYS for TMT MS2

Dramatic increase in the number of identified and quantified peptides

- CHIMERYS coefficient
  - Percentage of fragment abundance for PSM compared to all in same MS/MS spectrum
  - Coefficient > 0.8 considered acceptable





TMTpro data from human samples provided by Liwen Zhang at Ohio State
Data acquired and processed by Rosa Viner

Yeast TMT 10plex dataset – PXD020689

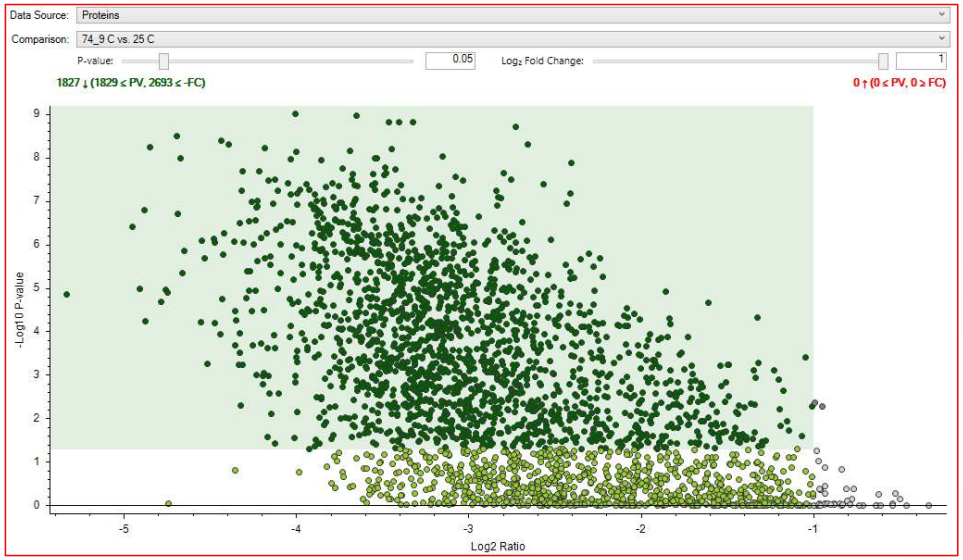
Example where CHIMERYS produces improved results

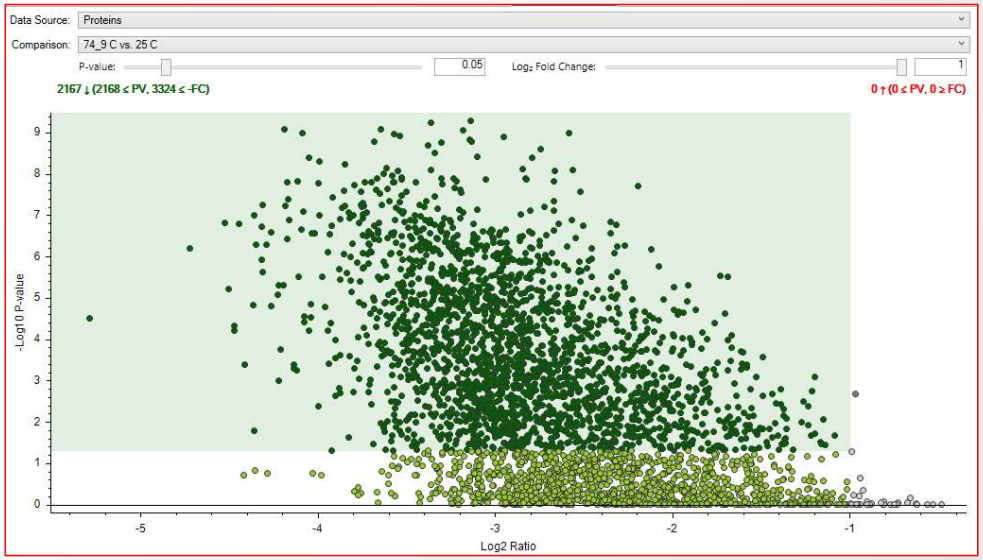
- “Boosting Detection of Low-Abundance Proteins in Thermal Proteome Profiling Experiments by Addition of an Isobaric Trigger Channel to TMT Multiplexes”, Justice *et al,* Analytical Chemistry, 2021.
- HCD MS/MS on Q Exactive HF-X mass spectrometer, yeast samples treated at different temperatures

Search strategy	Protein groups	Peptide groups	PSMs	Quantified peptides	Quantified proteins
Sequest HT	3448	41852	435424	36963	3253
Sequest HT + INFERYS Rescoring	3605	40867	424047	35400	3364
CHIMERYS	**4477**	**70211**	**1071875**	**45847**	**3998**

Yeast TMT 10plex dataset – PXD020689

CHIMERYS produces more differentially expressed proteins for this dataset





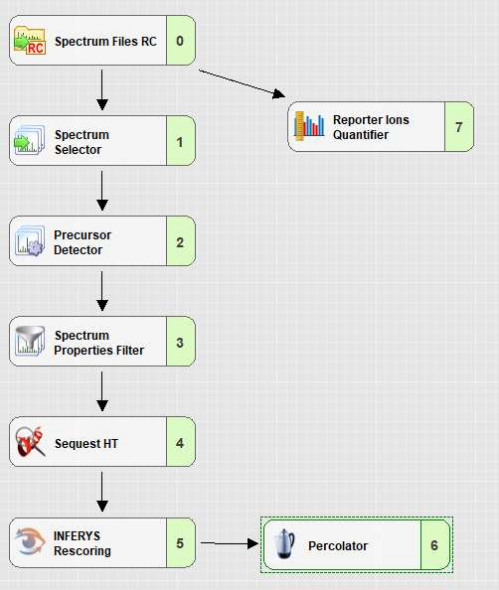
Sequest HT:
1827 down-regulated

CHIMERYS
2167 down-regulated

New INFERYS Rescoring Workflows

Default workflows for both TMT and TMTpro, MS2 and MS3 quan

- HCD MS2, SPS MS3



- Real-time search