

## **Proteome Discoverer 1.3**

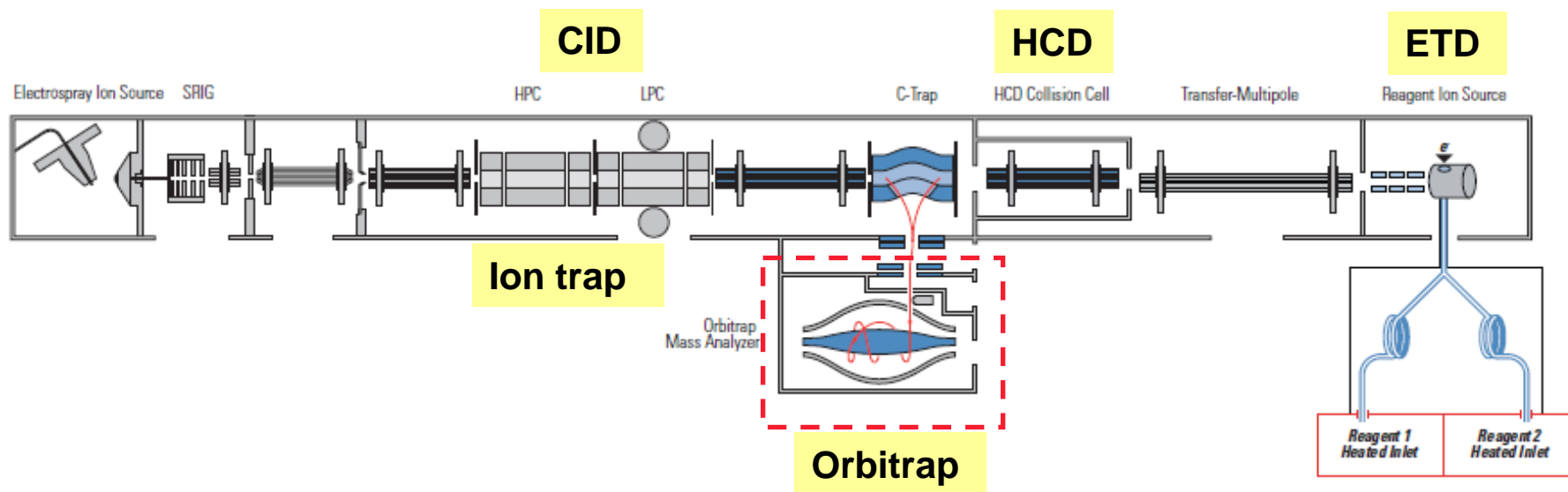
# **● Software: Enhanced Tools For Protein Identification**

David Horn

Proteomics Software Strategic Marketing Manager

September 22, 2011

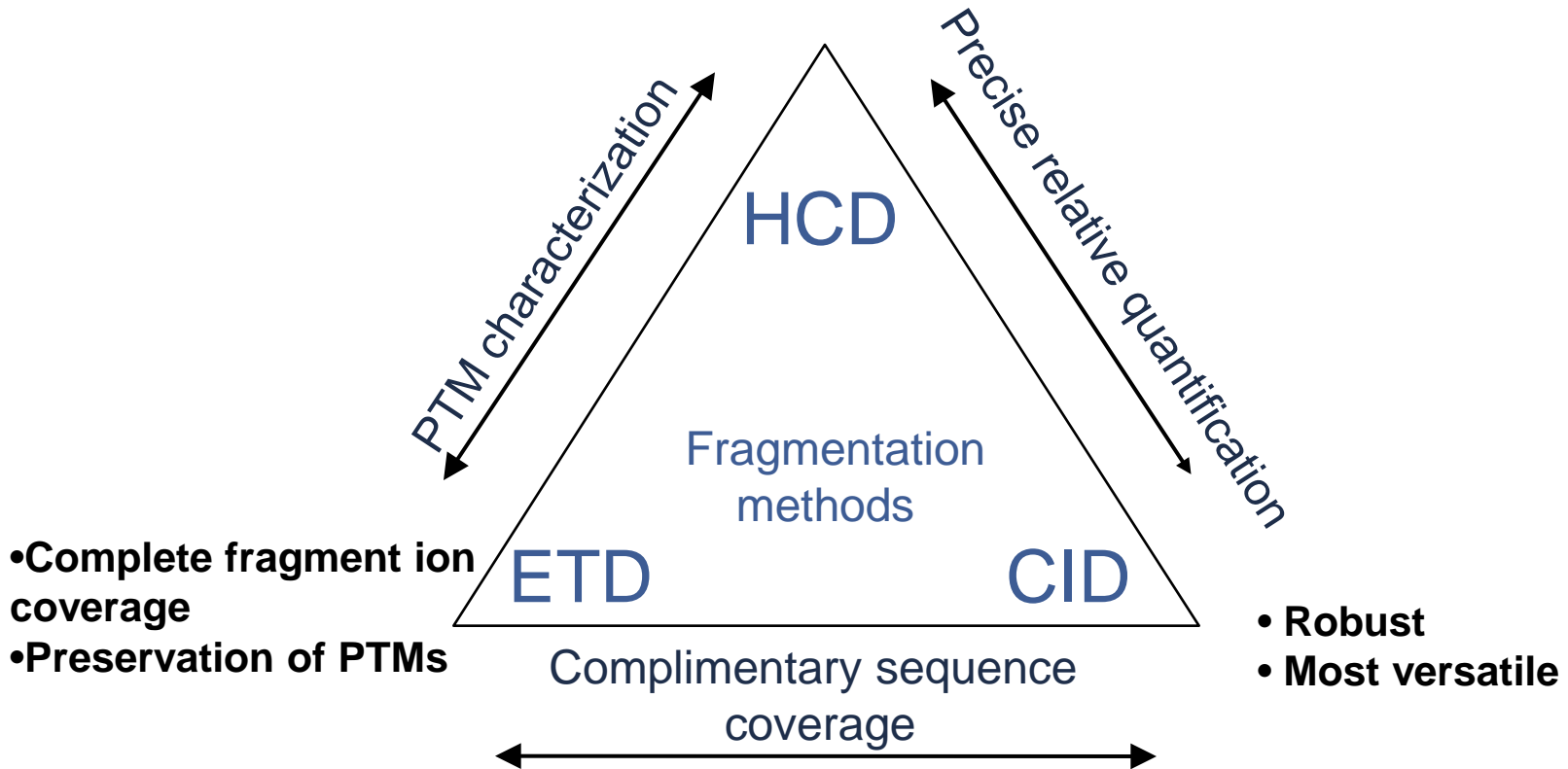
# Hybrid Orbitrap Mass Spectrometers



- Full flexibility of fragmentation techniques (CID, HCD, ETD)
- Up to 240,000 resolving power for MS, low-ppm mass tolerance
- Highest sensitivity (i.e. ion trap with SRIG: 5-10X vs. LTQ)
- Top down (Intact protein analysis), Bottom up (Protein ID), PTMs (especially those requiring MS<sup>n</sup> such as glycomics/glycoproteomics), Comprehensive Quan (SILAC, label free, targeted peptide quan with HRAM)

# Multiple Fragmentation Methods Provide Complementary Information

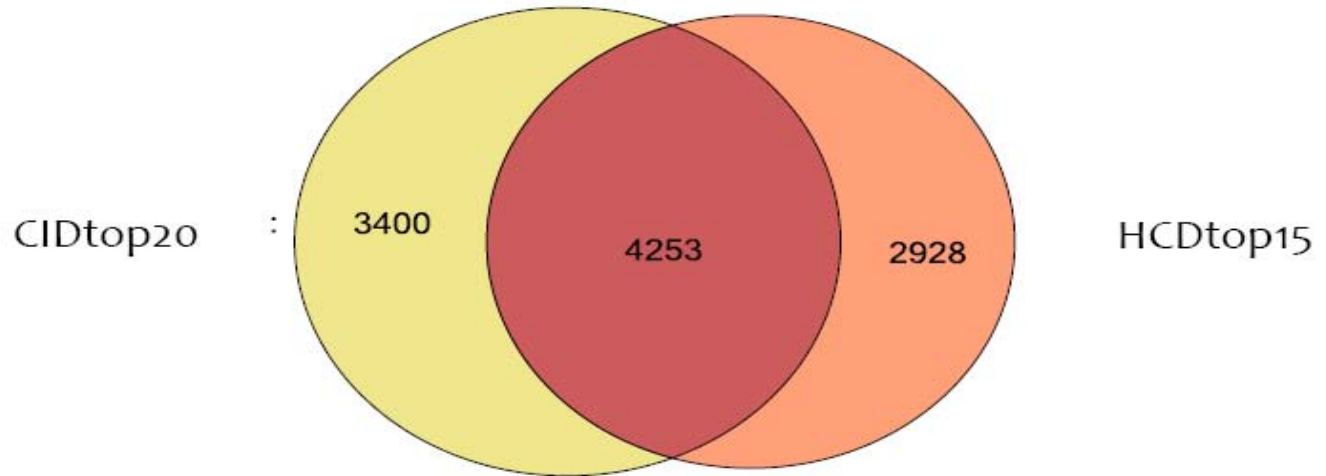
- High mass accuracy
- High energy fragmentation
- High intensity at low m/z
- Diagnostic ions (immonium & reporter ions)



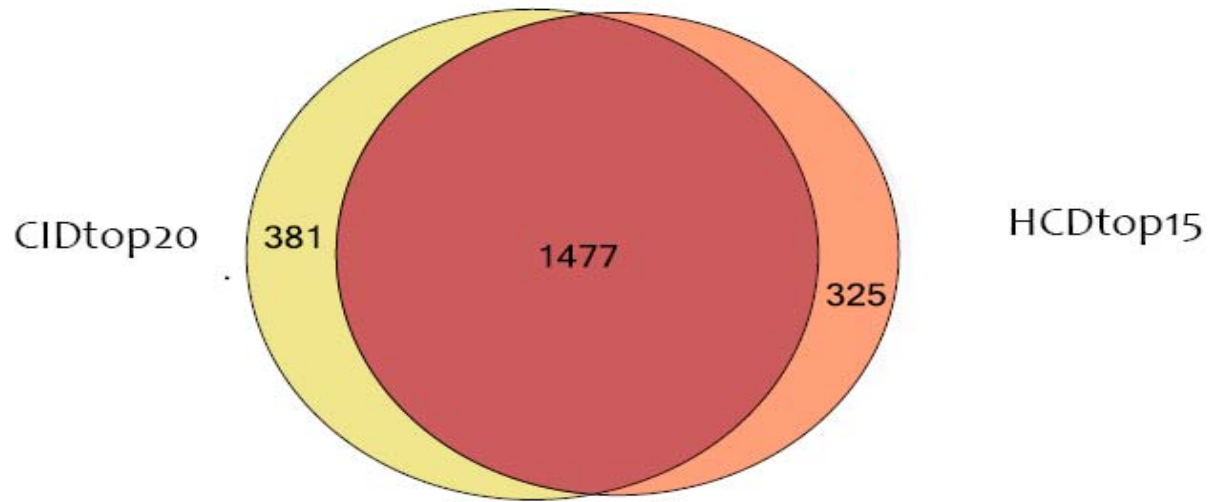
**Benefits of multiple fragmentation methods**

# Multiple modes of fragmentation = x 1.5 more unique peptides

## Peptides

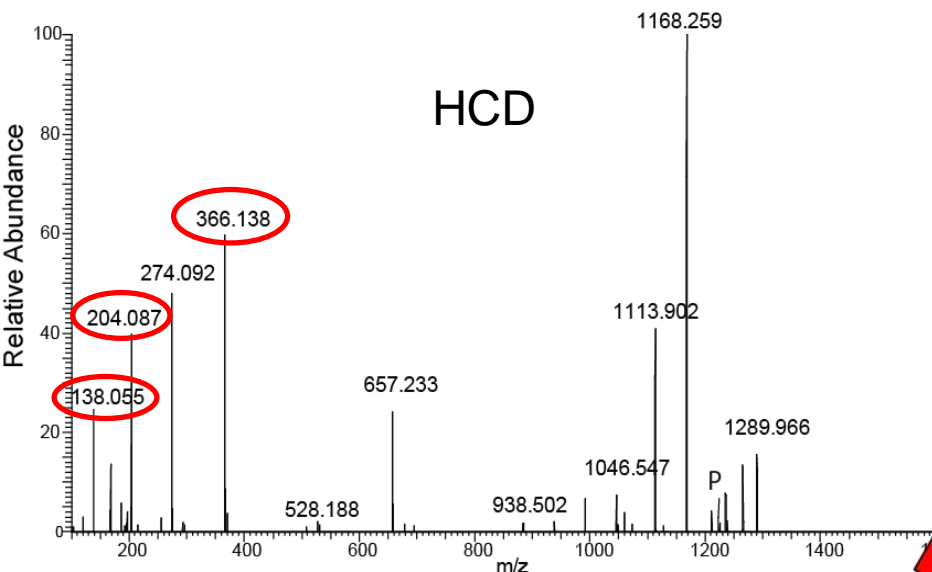


## Proteins



Sample: 200 ng HeLa lysate; tryptic digest; 90 min gradient.

# Example: HCD-triggered ETD for Identification of Glycopeptides



**Data Dependent Settings**

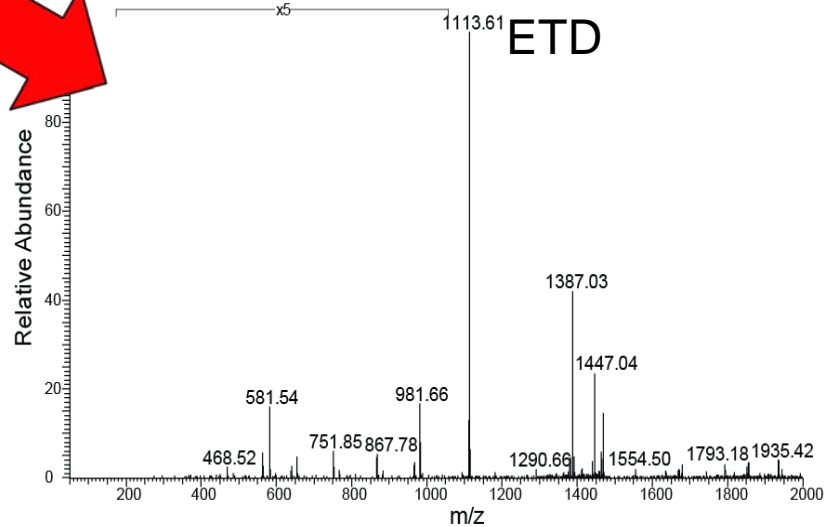
- Global
- Mass Widths
- Dynamic Exclusion
- Isotopic Data Depend
- Analog
- Product
- Segment
- Current Segment
- Chromatography
- Parent Mass List
- Reject Mass List
- Charge State
- Product Mass List
- Add Sub
- Scan Event
- Current Scan Event

Product masses:		
#	Mass	Name
1	138.0546	hexNac com
2	168.0653	hex nac frag
3	204.0867	hexNac
4		
5		
6		
7		
8		
9		
10		

Within top N: 20

10 ppm mass accuracy

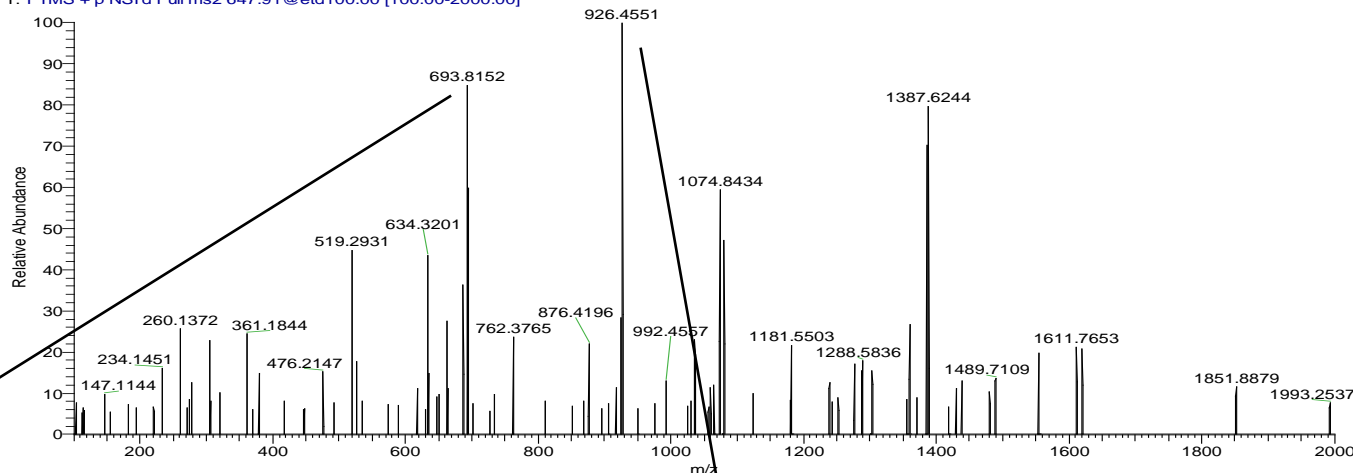
<b>Oxonium Ions=Product Ions</b>	
<b>Hex</b>	<b>m/z 163.06</b>
<b>HexNAc</b>	<b>m/z 204.087</b>
<b>HexNAc fragment ions</b>	<b>m/z 138.055</b>
	<b>m/z 186.076</b>
<b>Hex-HexNAc</b>	<b>m/z 366.138</b>



# ETD triggered Top 2 MS<sup>3</sup> CID of disulfide-linked peptides

ETD MS<sup>2</sup>

lman\_BSA\_ETD\_CIDms3\_Top2.raw #1139 RT: 19.25 AV: 1 NL: 1.91E5  
T: FTMS + p NSId Full ms2 647.91 @etd100.00 [100.00-2000.00]



Extracted from: C:\Users\lman\Desktop\Thermo Data\BSA\_Disulfide\lman\_BSA\_ETD\_CIDms3\_Top2.raw.raw #1141 RT: 19.28  
ITMS, CID, z=2, Mono m/z=693.81525 Da, MH+=1386.62322 Da, Match Tol=1.2 Da

CID MS<sup>3</sup>

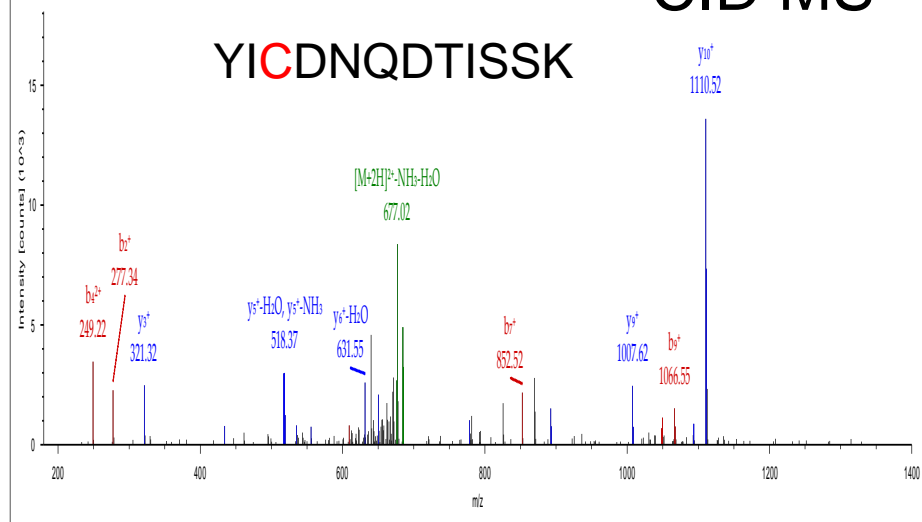
YICDNQDTISSK

y<sub>11</sub><sup>+</sup>

1110.52

[M+2H]<sup>2+</sup>-NH<sub>2</sub>-H<sub>2</sub>O

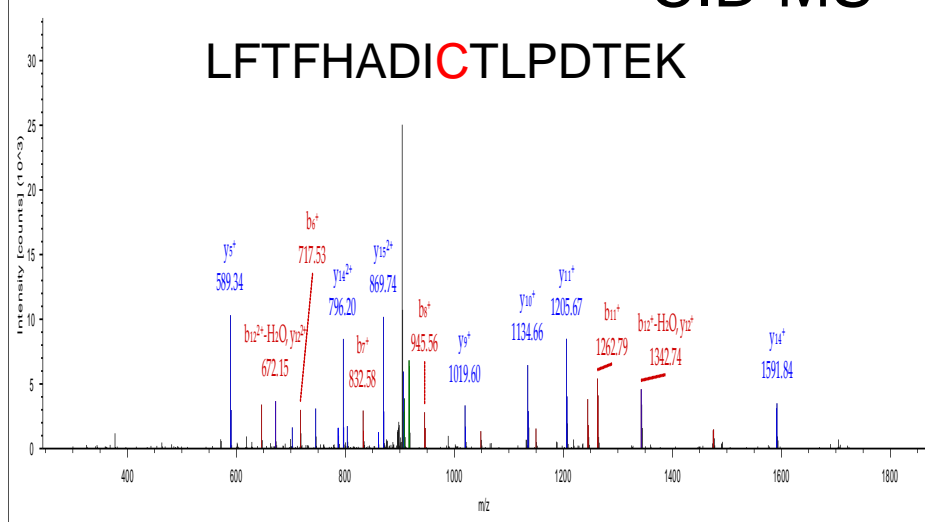
677.02



Extracted from: C:\Users\lman\Desktop\Thermo Data\BSA\_Disulfide\lman\_BSA\_ETD\_CIDms3\_Top2.raw.raw #1140 RT: 19.26  
ITMS, CID, z=2, Mono m/z=925.95477 Da, MH+=1850.90227 Da, Match Tol=1.2 Da

CID MS<sup>3</sup>

LFTFHADICTLPDTEK



# Analysis of Complex Proteomics Datasets Containing Mixed MS/MS modes

- Flexibility, throughput, and sensitivity of Orbitrap and ion trap systems produce large, highly complex datasets that most database search software packages are unable to utilize effectively
- Some examples:
  - Data-dependent decision tree (ETD on larger peptides, CID or HCD on smaller) to increase the number of ID's in a run
  - HCD-triggered ETD for labile PTM detection (e.g. glycans)
  - MS<sup>3</sup> of phosphopeptides
  - ETD MS/MS-triggered CID MS<sup>3</sup> for disulfide mapping
  - Many different quantification experiments, including SILAC, TMT, dimethyl labeling, spectral counting, precursor ion label-free quantification

Solution: **Proteome Discoverer 1.3**

- **Workflow-based system for proteomics “deep sequencing”**
- **Key Features**
  - Workflow editor enables highly flexible searches
  - Easily searches complex experiments produced by hybrid Orbitrap systems (e.g. data dependent decision tree, HCD-triggered ETD for glycan analysis, ETD->CID MS<sup>3</sup> disulfide mapping, etc.)
  - SEQUEST and Mascot
  - Labeled quantification (SILAC, TMT)
  - Automated calculation of false discovery rate (Percolator) – **New**
  - Biological annotation node and automated upload to ProteinCenter – **New**
  - Phosphorylation site localization – **New**



# Proteome Discoverer Workflow Editor

The screenshot displays the Proteome Discoverer Workflow Editor interface. The main window title is "Thermo Proteome Discoverer 1.3.0.339". The menu bar includes "File", "Search Report", "Quantification", "Processing", "Workflow Editor", "Administration", "Tools", "Window", and "Help". The toolbar contains various icons for file operations and search engines like SEQUEST, Mascot, and ZCore.

The "Workflow Editor (WF\_Q\_Exactive\_SEQUEST\_vs\_Mascot\_Search\_Percolator)" window is active. It shows a workflow diagram with the following nodes:

- Spectrum Files** (0)
- Spectrum Selector** (1)
- SEQUEST** (3)
- Percolator** (5)
- Annotation** (7)
- phosphoRS** (6)

The workflow is a linear sequence from Spectrum Files to Spectrum Selector to SEQUEST. From SEQUEST, the workflow branches into three parallel paths: Percolator, Annotation, and phosphoRS.

The "Parameters" panel on the right shows the following settings:

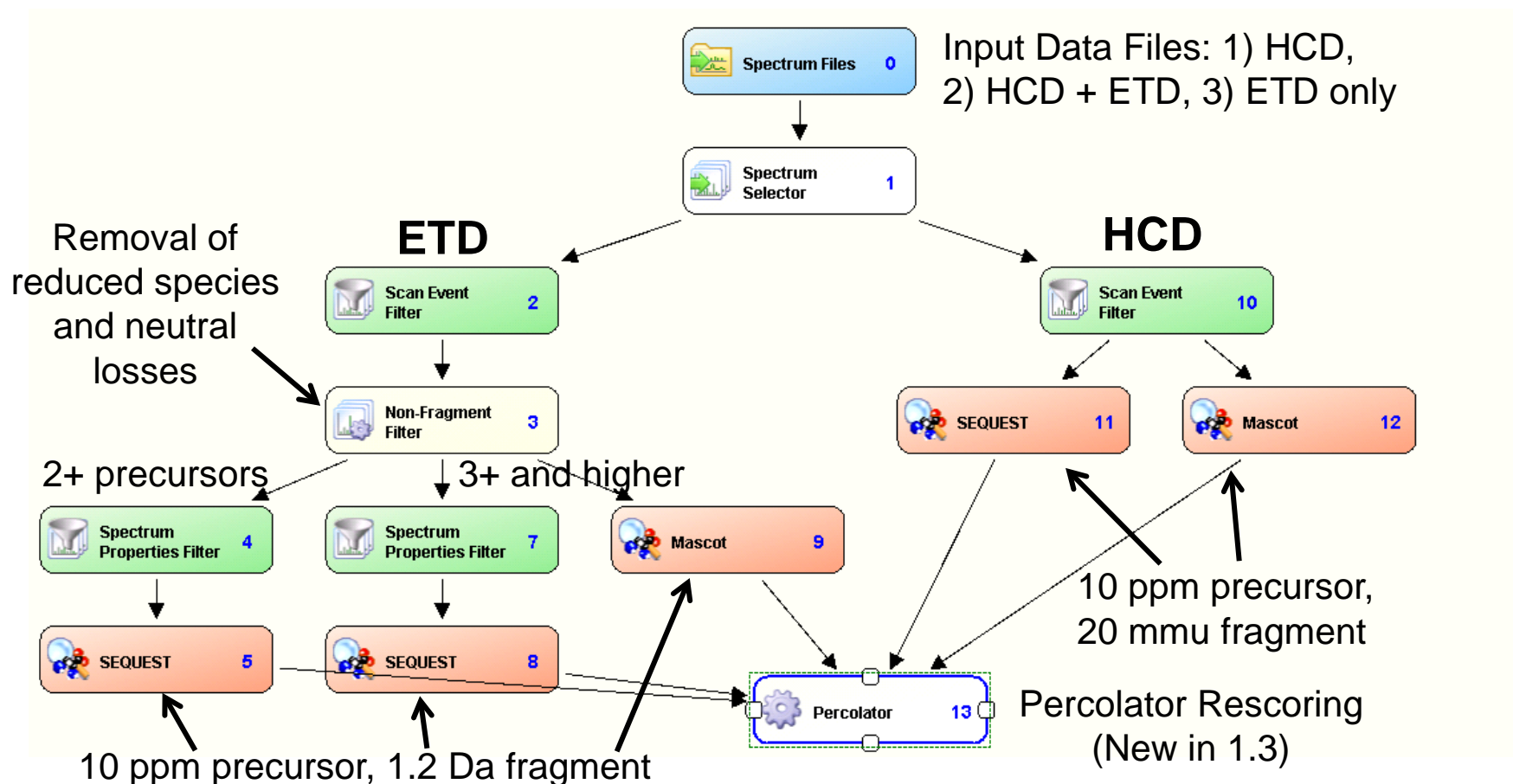
- 1. Input Data**
  - Protein Database: uniprot\_sprot2.fasta
  - Enzyme Name: Trypsin (Full)
  - Maximum Missed Cleavage: 2
- 2. Tolerances**
  - Precursor Mass Tolerance: 15 ppm
  - Fragment Mass Tolerance: 20 mmu
  - Use Average Precursor Ma: False
  - Use Average Fragment Ma: False
- 3. Ion Series**
  - Use Neutral Loss a Ions: True
  - Use Neutral Loss b Ions: True
  - Use Neutral Loss y Ions: True
  - Weight of a Ions: 0
  - Weight of b Ions: 1
  - Weight of c Ions: 0
  - Weight of x Ions: 0
  - Weight of y Ions: 1
  - Weight of z Ions: 0
- 4. Dynamic Modifications**
  - N-Terminal Modification: None
  - C-Terminal Modification: None
  - 1. Dynamic Modification: None
  - 2. Dynamic Modification: None

The "Protein Database" section indicates "The sequence database to be searched."

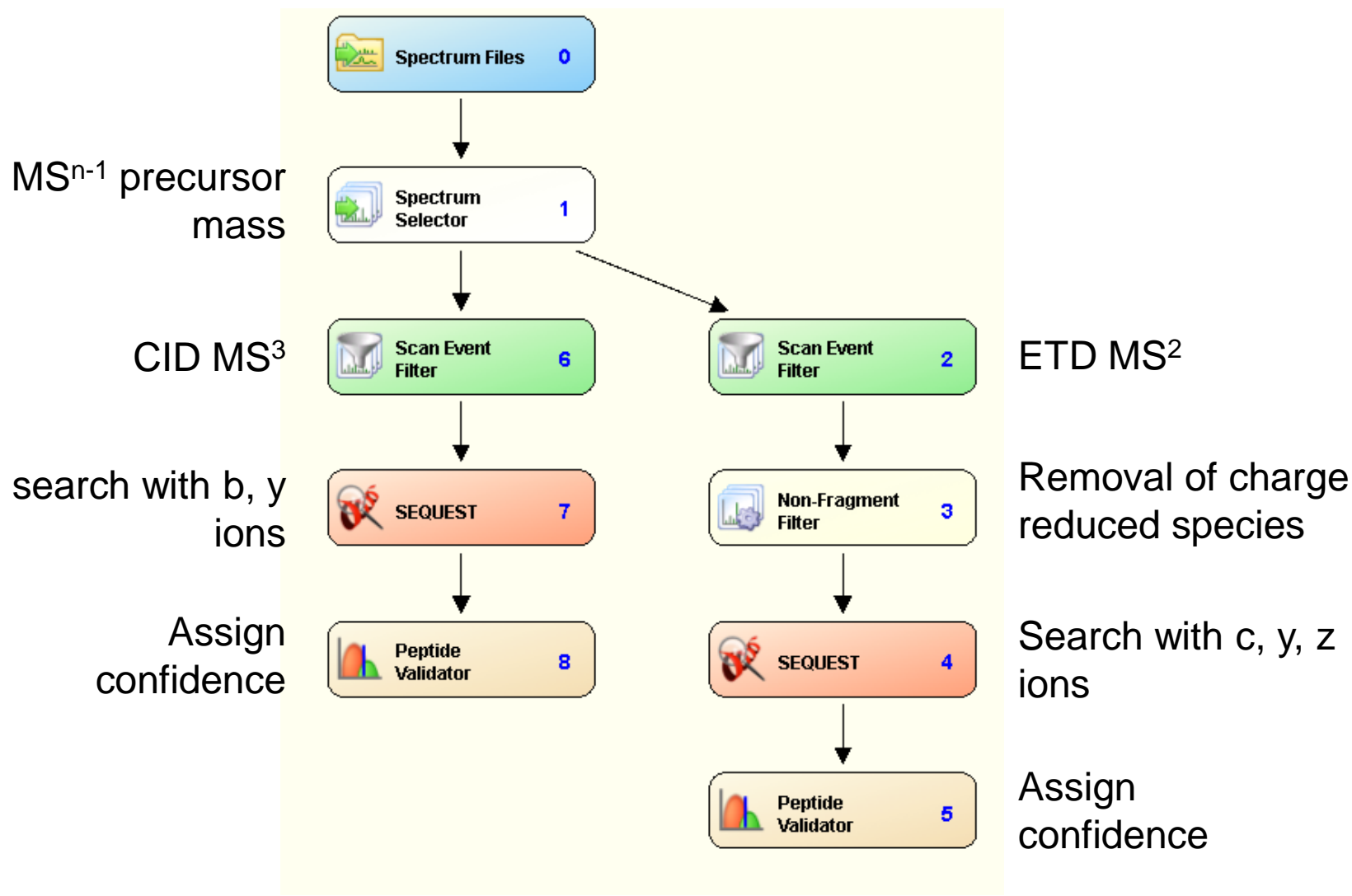
Red text annotations are present: "Nodes" with arrows pointing to the workflow diagram, "Workflow" with arrows pointing to the workflow diagram, and "Node Parameters" with arrows pointing to the Parameters panel.

# HCD-triggered ETD of O-GlcNAc Proteins

- Proteome Discoverer: 5 different database searches (3 SEQUEST, 2 Mascot), ETD-specific peak processing, Percolator post-processing

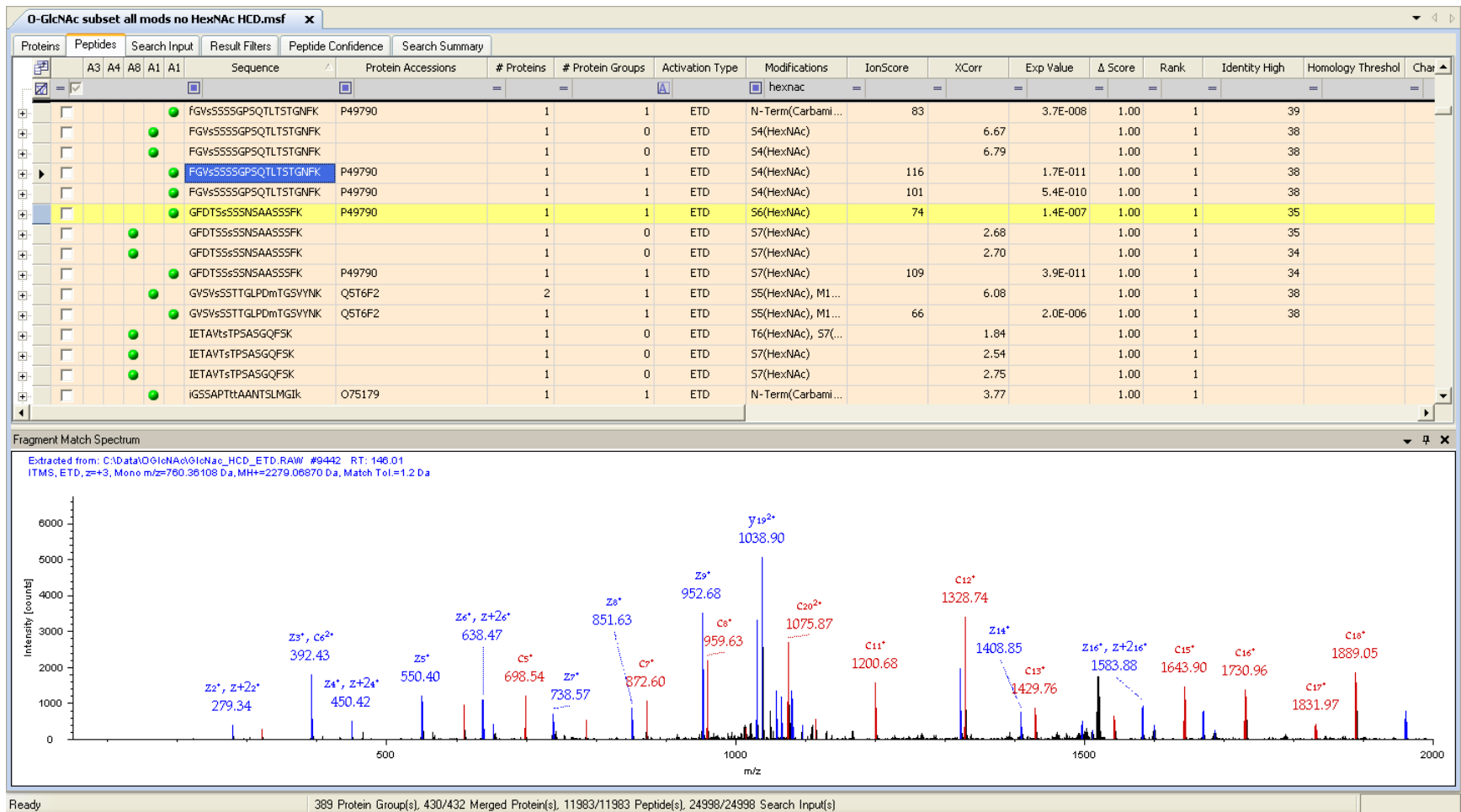


# Workflow for ETD-triggered CID MS<sup>3</sup>



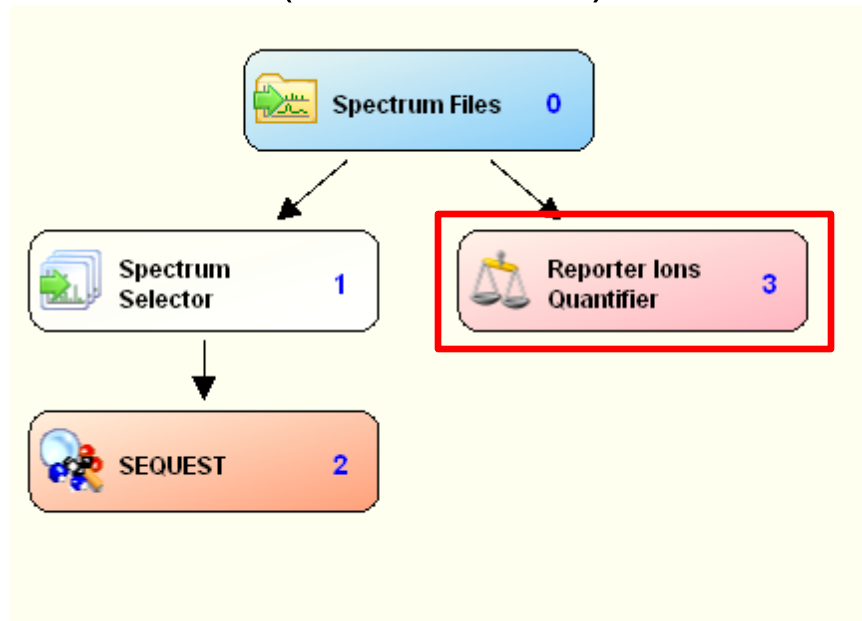
# HCD-triggered ETD of O-GlcNAc Proteins

- Proteome Discoverer Results:

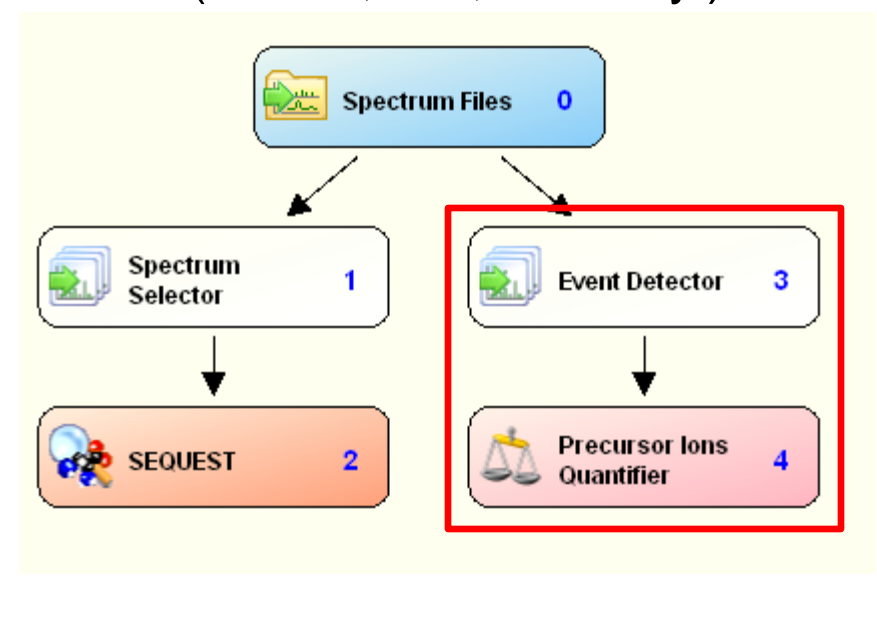


# Expression Profiling with Proteome Discoverer

## Reporter Ion Quantification (iTRAQ, TMT)



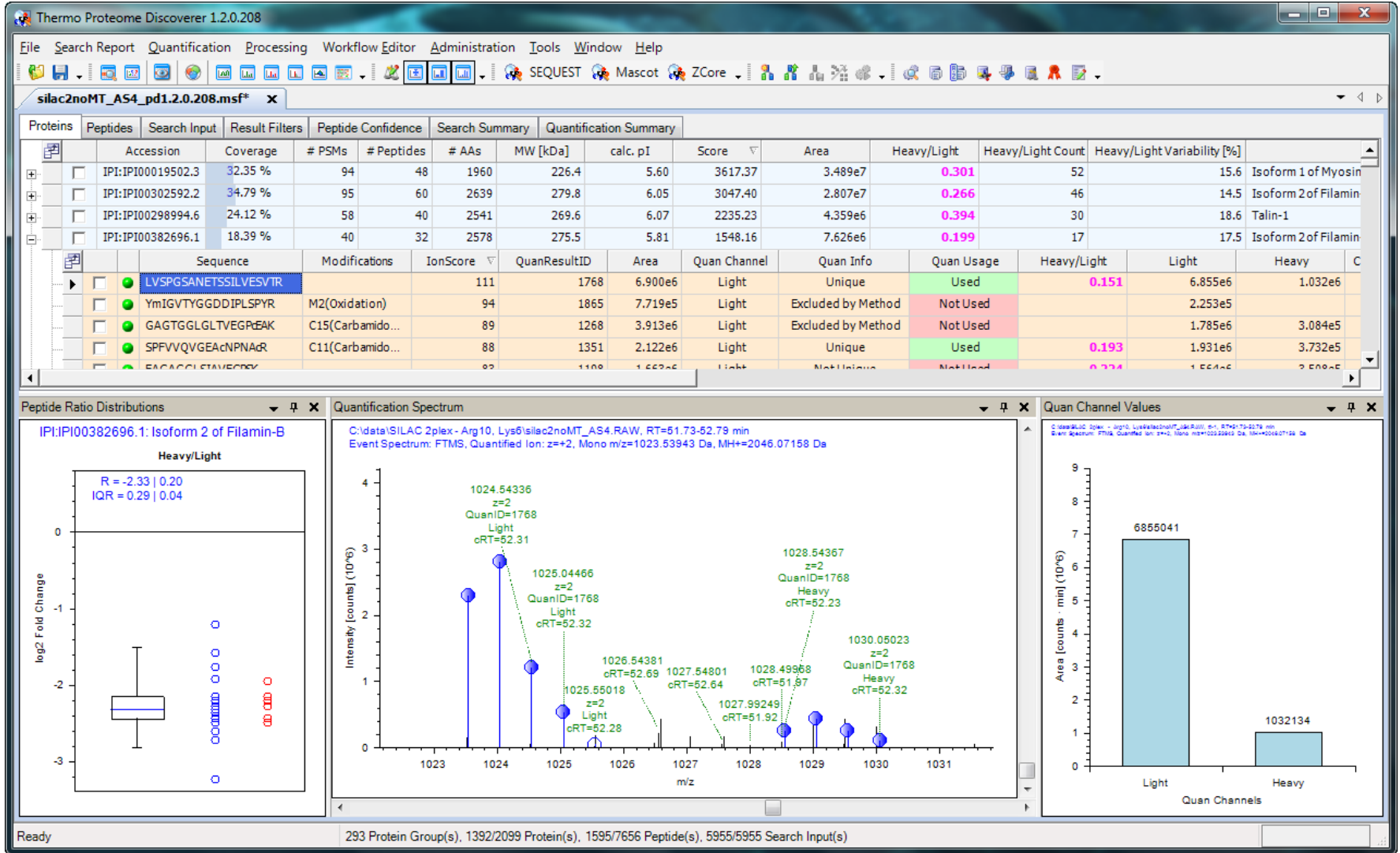
## Precursor Ion Quantification (SILAC, <sup>18</sup>O, Dimethyl)



## Expression ratios and variabilities for proteins and peptides:

Proteins	Peptides	Search Input		Result Filters		Peptide Confidence		Search Summary		Quantitation Summary		
		127/126	127/126 Counts	127/126 Variability [%]	128/126	128/126 Counts	128/126 Variability [%]	129/126	129/126 Counts	129/126 Variability [%]	130/126	130/126 Counts
0	0.993	56;44	3.5	1.020	55;43	2.9	0.993	56;44	3.5	0.983	55;43	
2	0.966	20;10	3.5	0.984	20;10	3.2	0.972	20;10	4.3	0.963	20;8	
5	1.041	19;22	2.8	1.021	19;22	6.7	1.045	19;22	6.4	1.057	19;22	
5	0.941	22;11	1.8	0.995	22;11	10.9	0.919	22;11	2.3	0.958	21;10	
5	0.979	55;42	1.2	1.001	55;42	1.2	0.986	55;42	2.8	1.009	54;42	

# Precursor Quan: Quan Result Display

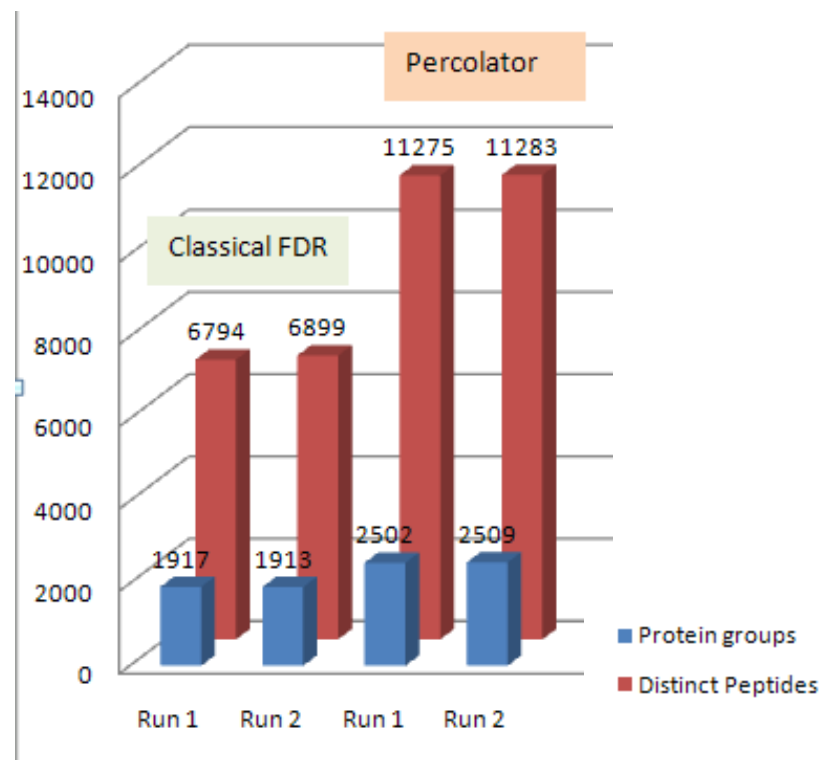


# Deep data mining using Percolator

- Percolator uses >30 features of a peptide spectral match (PSM) to distinguish true positives from random matches
- Result: more peptides and proteins identified with high confidence (1% FDR)
  - CID, HCD: >30% increase in PSM's
  - ETD: Up to 80% increase in PSM's
- Publication
  - Käll *et al*, Nature Methods 4:923-925 (2007)

## ASMS poster MP25

Breaking the 2000 proteins barrier in a standard LC run using a new benchtop Orbitrap instrument and multiple search engines.



# Percolator in Proteome Discoverer

The screenshot displays the Thermo Proteome Discoverer 1.3.0.339 interface. The main window shows a table of search results for peptide groups. A red box highlights the 'q-Value' and 'PEP' columns for peptide groups. A yellow callout box points to these columns with the text 'q-values and PEPs for peptide groups'. Below this, the 'Peptide Group Members' table is visible, with another red box highlighting its 'q-Value' and 'PEP' columns. A second yellow callout box points to these columns with the text 'q-values and PEPs for PSMs'. The status bar at the bottom indicates '(s), 5758/5758 Peptide(s), 42326/42326 PSM(s), 70574/70574 Search Input(s)'. The status bar also shows 'Ready'.

Protein	Sequence	# PSMs	# Proteins	# Protein Groups	Protein Group Accessions	Modifications	q-Value	PEP	IonScore	Exp Value	ΔCn
4911	RTsLSYLNK	2	1	1	YBL060W	S3(Phospho)	0.00598	0.105	22	7.1E-002	0.0000
4912	EcADEMktTPk	1	1	1	YPR018W	C2(Carbamidomethyl); K7...	0.00598	0.105	13	3.5E-001	0.0000
4913	ksSPATkVPSkPDr	4	1	1	YGL207W	K1(Lys8); S2(Phospho); K7...	0.00605	0.106	17	2.3E-001	0.0000
4914	ALDAsNAIDR	3	1	1	YJR068W	S5(Phospho)	0.00607	0.106	20	7.8E-002	0.0000
4915	kkDAsQEESLI	2	1	1	YBR127C	K1(Lys8); K2(Lys8); S5(Ph...	0.00609	0.107	13	5.5E-001	0.0000
4916	tcrHFISVILNSRGLETGK	2	1	1	YBR214W	T1(Phospho); C2(Carb...	0.00609	0.107	13	1.4E+000	0.0000
4917	MsPVLtPkr	2	1	1	YIL106W	S2(Phospho); T7(Phospho)	0.00616	0.107	20	1.3E-001	0.0000
4918	FLDkLGLsR					K4(Lys8); S8(Phospho)	0.00629	0.11	15	3.1E-001	0.0000
4919	KKDAsQEESLI					S5(Phospho); S9(Phospho)	0.00636	0.111	25	2.6E-002	0.0000
4920	NSTPSDASSTKntDHIV					T13(Phospho)	0.00641	0.112	12	1.0E+000	0.0000
4921	HLNTIILTK					T6(Phospho); K9(Lys8)	0.00643	0.112	18	1.8E-001	0.0000
4922	TkPAEEKsAEPEVk					K2(Lys8); K7(Lys8); S8(Ph...	0.00643	0.112	11	1.1E+000	0.0000
4923	FSNGGASsR					S8(Phospho)	0.00647	0.113	14	8.9E-002	0.0000
4924	LLDYFk	1	1	1	YLR355C	K6(Lys8)	0.0065	0.114	16	1.5E-001	0.0000
4925	RLsGIMR	1	1	1	YBR156C	S3(Phospho)	0.00661	0.114	28	1.1E-002	0.0000

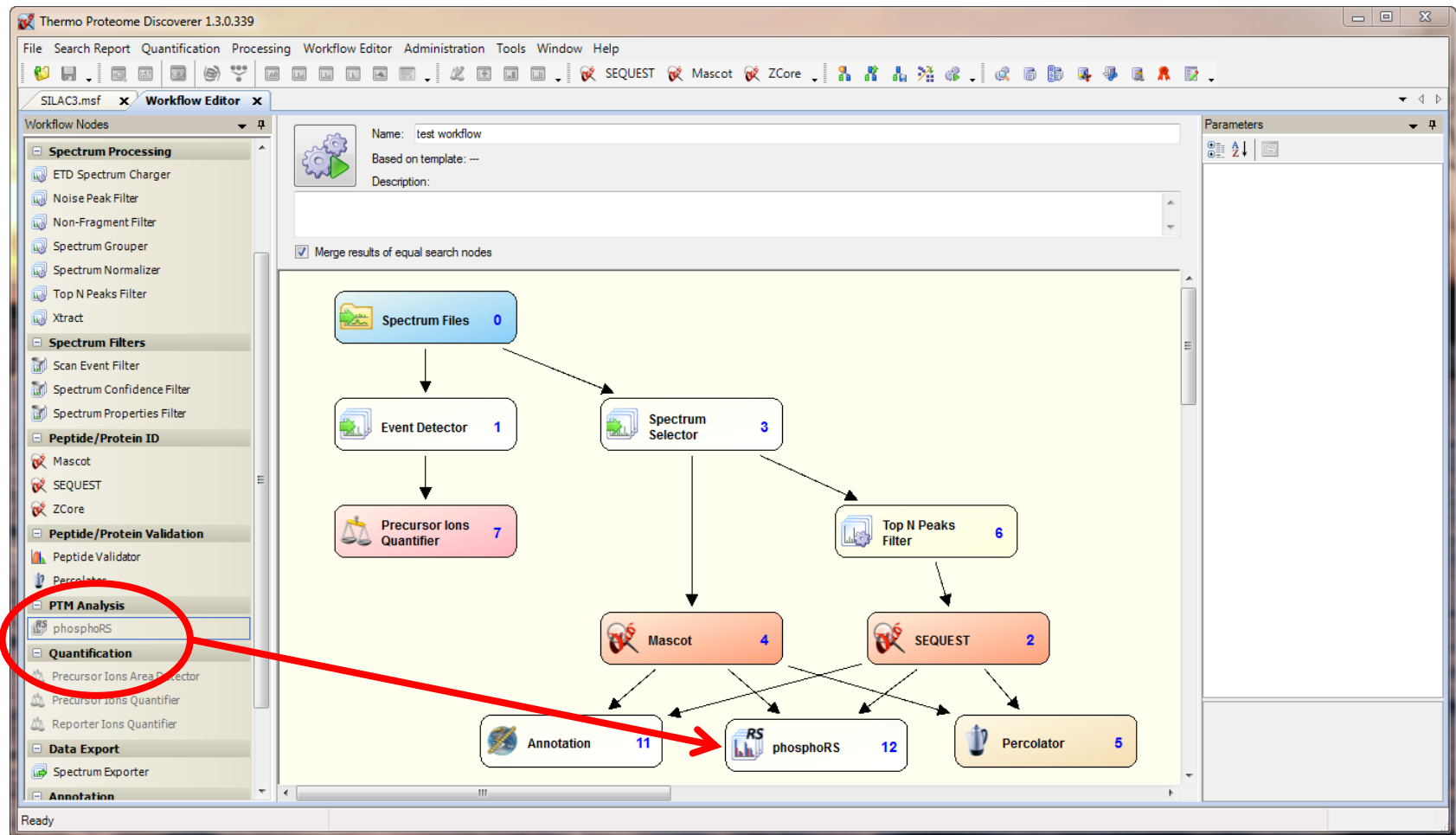
  

Peptide Group Members	Sequence	PSM Ambiguity	Protein Group Accessions	Modifications	Rank	q-Value	PEP	IonScore	ΔScore	ΔCn	Exp Value	Search Eng
1	ksSPATkVPSkPDr	Unambiguous	YGL207W	K1(Lys8); S2(Phospho); K7...	1	0.00605	0.106	17	0.2353	0.0000	2.3E-001	
2	kSsPATkVPSkPDr	Unambiguous	YGL207W	K1(Lys8); S3(Phospho); K7...	2	0.0357	0.572	15	1.0000	0.0000	3.8E-001	
3	ksSPATkVPSkPDr	Unconsidered	YGL207W	K1(Lys8); S2(Phospho); K7...	3	0.0641	0.843	15	1.0000	0.0000	3.8E-001	
4	kSsPATkVPSkPDr	Unconsidered	YGL207W	K1(Lys8); S3(Phospho); K7...	4			13		0.2353	5.7E-001	



# phosphoRS node in Proteome Discoverer

- Developed in conjunction with Karl Mechtler's group at University of Vienna for determination of phosphorylation site confidence



# phosphoRS in Proteome Discoverer

The screenshot displays the Thermo Proteome Discoverer 1.3.0.339 interface. The main window shows a table of peptide data with columns for Peptides, Search Input, Result Filters, Peptide Confidence, Search Summary, and Quantification Summary. A red box highlights the pRS Score, pRS Probability, and pRS Site Probabilities columns for the peptide groups. A yellow callout box with a red border points to these columns, containing the text "pRS scores for peptide groups".

Peptides	Sequence	# PSMs	# Proteins	# Protein Groups	Protein Group Accessions	Modifications	pRS Score	pRS Probability	pRS Site Probabilities
4911	RTsLSYLNK	2	1	1	YBL060W	S3(Phospho)	8	4.0 %	T(2): 4.0; S(3): 4.0; S(5): 46.0; Y(6): 46.0
4912	EcADEMktTPk	1	1	1	YPR018W	C2(Carbamidomethyl); K7(...)	20	84.2 %	T(8): 84.2; T(9): 15.8
4913	ksSPATkVPSkPDr	4	1	1	YGL207W	K1(Lys8); S2(Phospho); K7(...)	24	32.3 %	S(2): 32.3; S(3): 32.3; T(6): 32.3; S(10): ...
4914	ALDAsNAIDR	3	1	1	YJR068W	S5(Phospho)	39	100.0 %	S(5): 100.0
4915	kkDAsQEESLI	2	1	1	YBR127C	K1(Lys8); K2(Lys8); S5(Ph...)	40	98.1 %	S(5): 98.1; S(9): 1.9
4916	trHFISVILNSRGLETGK	2	1	1	YBR214W	T1(Phospho); C2(Carbamidomethyl); K7(...)	54	94.2 %	T(1): 94.2; S(7): 5.8; S(12): 0.0; T(17): 0.0
4917	MsPVLtPkr	2	1	1	YIL106W	S2(Phospho); T7(Phospho); K7(...)	73	50.0 %	S(2): 100.0; T(6): 50.0; T(7): 50.0
4918	FLDKLGLSR	1	1	1	YGL207W	K4(Lys8); S8(Phospho)	46	100.0 %	S(8): 100.0
4919	KKDAsQEESLI	1	1	1	YBR127C	S5(Phospho); S9(Phospho)	80	100.0 %	S(5): 100.0; S(9): 100.0
4920	NSTPSDASSTKNNDHIV	2	1	1	YGL207W	T3(Phospho)	23	8.7 %	S(2): 0.1; T(3): 0.3; S(5): 1.3; S(8): 8.7; ...
4921	HLNITLTK	3	1	1	YGL207W	T6(Phospho); K9(Lys8)	34	10.3 %	T(4): 89.5; T(6): 10.3; T(8): 0.3
4922	TkPAEEKsAEPEVk	1	1	1	YGL207W	K2(Lys8); K7(Lys8); S8(Ph...)	26	88.4 %	T(1): 11.6; S(8): 88.4
4923	FSNGGASr	1	1	1	YGL207W	S8(Phospho)	36	89.3 %	S(2): 0.1; S(7): 10.6; S(8): 89.3
4924	LLDYFk	1	1	1	YLR355C	K6(Lys8)	60	100.0 %	S(3): 100.0
4925	RLsGIMR	1	1	1	YBR156C	S3(Phospho)	77	100.0 %	T(6): 100.0
4926	AINGLMLK	1	0	0	YGL207W	T6(Phospho)	77	100.0 %	T(6): 100.0
4927	EASksPISSFVNDYr	3	1	1	YDR096W	K4(Lys8); S5(Phospho); R1(...)	21	16.2 %	S(3): 16.2; S(5): 16.2; S(8): 62.4; S(9): 4...

The bottom window shows the Peptide Group Members table, which details individual PSMs for the peptide ksSPATkVPSkPDr. A red box highlights the pRS Score, pRS Probability, and pRS Site Probabilities columns for these PSMs. A yellow callout box with a red border points to these columns, containing the text "pRS scores for PSMs".

Peptide Group Members	Sequence	PSM Ambiguity	Protein Group Accessions	Modifications	Rank	pRS Score	pRS Probability	pRS Site Probabilities	q-Value
1	ksSPATkVPSkPDr	Unambiguous	YGL207W	K1(Lys8); S2(Phospho); K7(...)	1	24	32.3 %	S(2): 32.3; S(3): 32.3; T(6): 32.3; S(10): 3.1	0.00605
2	ksSPATkVPSkPDr	Unambiguous	YGL207W	K1(Lys8); S3(Phospho); K7(...)	2	14	31.8 %	S(2): 31.8; S(3): 31.8; T(6): 31.8; S(10): 4.6	0.0357
3	ksSPATkVPSkPDr	Unconsidered	YGL207W	K1(Lys8); S2(Phospho); K7(...)	3	14	31.8 %	S(2): 31.8; S(3): 31.8; T(6): 31.8; S(10): 4.6	0.0641
4	ksSPATkVPSkPDr	Unconsidered	YGL207W	K1(Lys8); S3(Phospho); K7(...)	4	24	32.3 %	S(2): 32.3; S(3): 32.3; T(6): 32.3; S(10): 3.1	

# Visualization of Found and Known PTMs

- The Protein ID Details view now presents a comprehensive overview of all modifications found in the MS/MS data of each protein

**Protein Identification Details**

CDC19 SGDDID:S000000036, Chr I from 71787-73289, Verified ORF. "Pyruvate kinase, functions as a homotetramer in glycolysis to convert phosphoenolpyruvate to pyruvate, the input for aerobic (TCA cycle) or anaerobic (glucose fermentation) respiration"

Annotate PTMs reported in Uniprot  
 Show only PTMs  
 Include PSMs that are filtered Out

Coverage: 48.20%

**Found Modifications:**

- A Arg6 (R)
- C Carbamidomethyl (C)
- L Lys8 (K)
- O Oxidation (M)
- P Phospho (T.S)

**Uniprot PTMs:**

- P Phospho

**PTM Site Probabilities:**

- 25 - 45% (Pink)
- 45% - 75% (Yellow)
- 75% - 99% (Light Green)
- 99% - 100% (Dark Green)

**Sequence**

Modification List	1	11	21	31	41	51	61	71	81	91
Modifications		P	A	P	L	A	L	A		
Uniprot	MSRLRLTSL	NVAGSDLRR	TSIIIGTIGPK	TNNPETLVAL	RKAGLNIVRM	NFSHGSYEYH	KSVIDNARKS	EELYPCRPLA	IALDTKGPEI	RIGTITNDVD
Modifications			O	L						L
Uniprot	YPIPPNHEMI	FTTDDKYAKA	CDDKIMVVDY	KNITKVISAG	RIIYVDDGVL	SFQVLEVVD	KILKVKALNA	GKICSHKGVN	LPGQVDLPA	LSEKDKEDLR
Modifications				A	L	L	L	O	A	L
Uniprot	FGVKNQVMV	FASFIRFAND	VLTIREVLCG	QGKQVKLIUV	IENQQGVNPF	DEILKVTDCV	MVARGDLGIE	IPAPEVLAVQ	KKLIAKSNLA	GKPVICATQM
Modifications								C	L	L
Uniprot	LESMTYNPRP	TRAEVSDVGN	AILDGADCVM	LSGETAKGNY	PINAVTTMAE	TAVIAEQAIA	YLPNYDDMRN	CTPKPTSTTE	TVAASAVAAV	FEQKAKAIVV
Modifications				A			A	L	L	L
Uniprot	LSTSGTTPRL	VSKYRPNCPV	ILVTRCPRAA	RFSHLYRGVF	PFVFEKEPVS	DWTDDVEARI	NFGIEKAKEF	GILKKGDTYV	SIQGFKAGAG	HSNTLQVSTV
Modifications										
Uniprot	PP	P				P		PP		P

Toggle between PTMs only and all mods

double click a protein of interest

Site probabilities are color coded

Found and documented PTM are annotated on the protein sequence

# Visualization of Identified PTMs

- Find detailed information about the found PTMs in the Modification List

Protein Identification Details

Coverage **ProteinCard**

CDC19 SGDID:S000000036, Chr I from 71787-73289, Verified ORF. "Pyruvate kinase, functions as a homotetramer in glycolysis to convert phosphoenolpyruvate to pyruvate, the input for aerobic (TCA cycle) or anaerobic (glucose fermentation) respiration"

Annotate PTMs reported in Uniprot  
 Show only PTMs  
 Include PSMs that are filtered Out

Coverage: **48.20%**

**Found Modifications:**

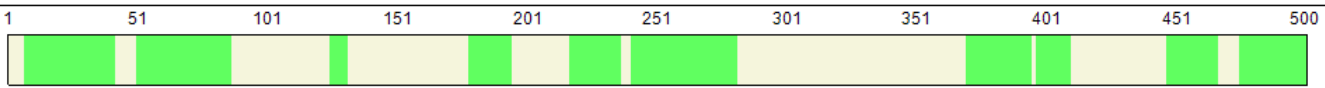
P Phospho (T,S)

**Uniprot PTMs:**

P Phospho

**PTM Site Probabilities:**

- 25 - 45%
- 45% - 75%
- 75% - 99%
- 99% - 100%



Sequence Modification List

Position	Target	Modification	Classification	Highest PTM Score	Highest Peptide Confidence	Sequence Motif
8	T	Phospho	Post-translational	0.5	High	SRLERLrSLNVVA
9	S	Phospho	Post-translational	99.9	High	RLERLrLNvVAG
16	S	Phospho	Post-translational	0	High	LNvVAGsDLRRIS
21	T	Phospho	Post-translational	2.2	High	GSDLRRrSIIGTI
22	S	Phospho	Post-translational	97.4	High	SDLRRrSIIGTIG
26	T	Phospho	Post-translational	0.4	High	RISIIGrIGPKIN

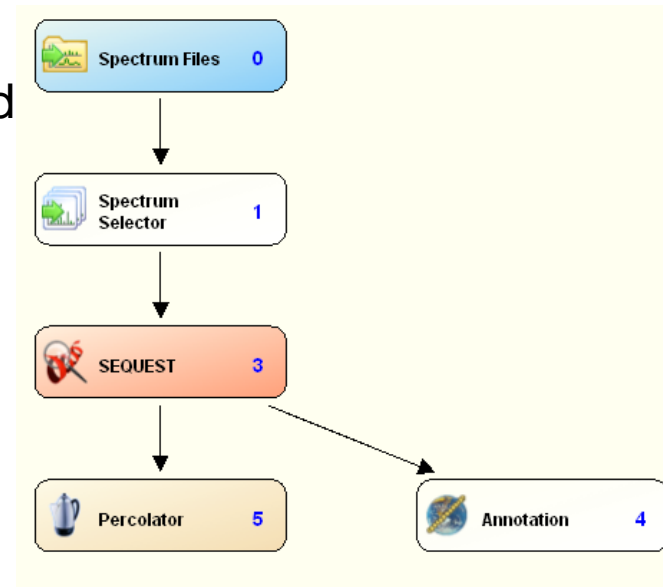
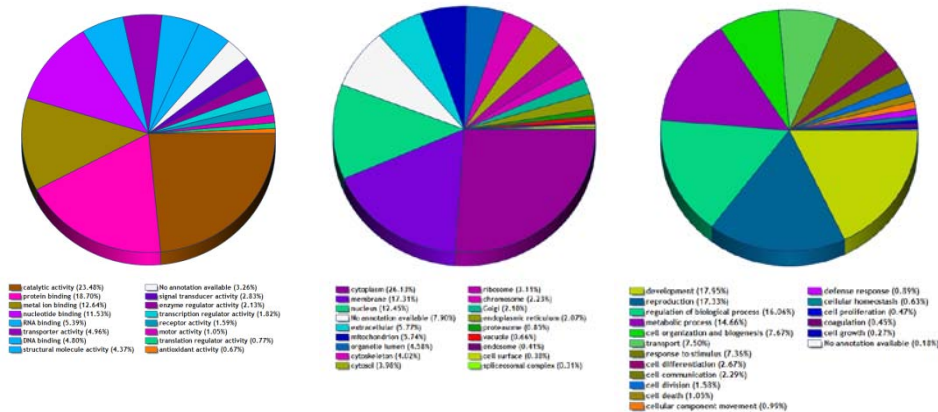
6 Number of amino acids displayed before and after the modification

OK Help

# Providing biological context using the Annotation node

- 3 year support subscription includes annotation node:
  - Automatically queries hosted ProteinCenter server in Denmark for GO, protein family (Pfam), PTM, and ProteinCard annotations
    - All can be filtered
    - Displayed in the PD Viewer
  - Previous results can be annotated or re-annotated

Molecular Function      Cellular Component      Biological Process



ASMS Poster ThP22 401

“Integration of a central protein repository into a standard data processing application for mining proteomics data”

# ProteinCard for Identified Proteins

- ProteinCard can be displayed for every identified protein whose accession is tracked in ProteinCenter

Thermo Proteome Discoverer 1.3.0.339

File Search Report Quantification Processing Workflow Editor

SILAC3.msf x

Proteins	Peptides	Search Input	Result Filters	Peptide Confidence
Accession	Description			
1	<input type="checkbox"/>	YLR249W	YEF3 SGDID:S000004239, Chr XII from	
2	<input type="checkbox"/>	YBR189W	RPS9B SGDID:S000000393, Chr II from	
3	<input type="checkbox"/>	YBR127C	VMA2 SGDID:S000000331, Chr II from	
4	<input type="checkbox"/>	YHR064C	SSZ1 SGDID:S000001106, Chr VIII from	
5	<input type="checkbox"/>	YER091C	MET6 SGDID:S000000893, Chr V from	
6	<input type="checkbox"/>	YMR186W	HSO2 SGDID:S000004798, Chr XIII from	
7	<input type="checkbox"/>	YPL240C	HSP82 SGDID:S000006161, Chr XVI from	
8	<input type="checkbox"/>	YGL253W	HXK2 SGDID:S000003222, Chr VII from	
9	<input type="checkbox"/>	YGR192C	TDH3 SGDID:S000003424, Chr VII from	
10	<input checked="" type="checkbox"/>	YAL038W	CDJ19 SGDID:S000000036, Chr I from	
11	<input type="checkbox"/>	YJR009C	TDH2 SGDID:S000003769, Chr X from	
12	<input type="checkbox"/>	YGR135W	PRE9 SGDID:S000003367, Chr VII from	
13	<input type="checkbox"/>	YNR016C	ACC1 SGDID:S000005299, Chr XIV from	
14	<input type="checkbox"/>			
15	<input type="checkbox"/>			
16	<input type="checkbox"/>			
17	<input type="checkbox"/>			
18	<input type="checkbox"/>			
19	<input type="checkbox"/>			
20	<input type="checkbox"/>			
21	<input type="checkbox"/>			
22	<input type="checkbox"/>			
23	<input type="checkbox"/>	YDR365W-B	YDR365W-B SGDID:S000007401, Chr	
24	<input type="checkbox"/>	YBL027W	RPL19B SGDID:S000000123, Chr II from	
25	<input type="checkbox"/>	YER026C	CHO1 SGDID:S000000828, Chr V from	
26	<input type="checkbox"/>	YHR174W	ENO2 SGDID:S000001217, Chr VIII from	

Protein Identification Details

Coverage ProteinCard

General Keys Features Molecular Functions Cellular Components Biological Processes Diseases External Links

### External Links

- [SGD](#)
- [Entrez Gene](#)
- [BLINK](#)
- [UniRef100](#)
- [UniRef90](#)
- [UniRef50](#)
- [PDB](#)
- [PubMed](#)
- [SNPs](#)
- [Nt](#)
- [UCSC](#)
- [NCBI map](#)
- [Homologene](#)
- [GEO profiles](#)
- [UniGene](#)
- [IntActAll](#)

OK Help

Ready 1122/1129 Protein Group(s), 1321/1321 Protein(s), 5758/5758 Peptide(s), 42326/42326 PSM(s), 70574/70574 Search Input(s)

double click a protein of interest



# Filtering identified proteins using GO and Pfam Annotations

The screenshot displays the Thermo Proteome Discoverer 1.3.0.339 interface. The main window shows a table of identified proteins with columns for Accession, Description, Score, Coverage, Molecular Function, Cellular Component, Biological Process, Pfam IDs, and Area. A dialog box is open over the table, allowing the user to select filter terms. The 'protein binding' term is selected in the 'Or' section of the dialog. A red arrow points from a text box to the 'protein binding' selection. The text box contains the instruction: 'select GO slim terms as filter from the Row Filter dropdown'.

Accession	Description	Score	Coverage	Molecular Function	Cellular Component	Biological Process	Pfam IDs	Area
1	YLR249W YEF3 SGDID:S000004239, Chr XII from 636782-639916, V...	32459.85	14.27 %				Pf00005	1.767
2	YBR189W RPS9B SGDID:S000000393, Chr II from 604503-604509,60...	15804.15	56.92 %				Pf00163; Pf01479	4.035
3	YBR127C VMA2 SGDID:S000000331, Chr II from 492816-491263, re...	9426.84	14.70 %				Pf00006; Pf00306;...	1.211
4	YHR064C S5Z1 SGDID:S000001106, Chr VIII from 227143-225527, r...	7497.94	13.20 %				Pf00012; Pf06723	8.607
5	YER091C MET6 SGDID:S000000893, Chr V from 342163-339860, re...	5526.08	28.68 %				Pf01717	5.329
6	YMR186W HSC82 SGDID:S000004798, Chr XIII from 632354-634471,...	5353.10	35.18 %				Pf00183; Pf02518	1.018
7	YPL240C HSP82 SGDID:S000006161, Chr XVI from 98625-96496, re...	5091.64	31.88 %				Pf00183; Pf02518	1.018
8	YGL253W HXK2 SGDID:S000003222, Chr VII from 23935-25395, Veri...	4912.87	4.12 %				Pf00349; Pf03727	7.568
9	YGR192C TDH3 SGDID:S000003424, Chr VII from 883815-882817, r...	4656.40	49.70 %				Pf00044; Pf02800	1.259
10	YAL038W CDC19 SGDID:S000000036, Chr I from 71787-73289, Verif...	4094.51	48.20 %				Pf00224; Pf02887;...	1.047
11	YJR009C TDH2 SGDID:S000003769, Chr X from 454673-453675, re...	3941.63	47.9 %				Pf00044; Pf02800	3.851
12	YGR135W PRE9 SGDID:S000003367, Chr VII from 761397-762173, V...	3788.8	4.65 %				Pf00227	3.462
13	YNR016C ACC1 SGDID:S000005299, Chr XIV from 661377-654676, r...	3509.93	6.90 %				Pf00289; Pf00364;...	1.086
14	YPL081W RPS9A SGDID:S000006002, Chr XVI from 404947-404953...	3427.11	59.90 %				Pf00163; Pf01479	1.940
15	YEL046C GLY1 SGDID:S000000772, Chr V from 68792-67620, Fever...	3415.90	3.62 %				Pf00155	9.163
16	YDL229W SSB1 SGDID:S000000001, Chr III from 1206990-1206990, V...	3355.5	3.95 %				Pf00012; Pf06723	3.200
17	YKL152C GPM1 SGDID:S000000001, Chr III from 1206990-1206990, V...	3277.7	3.7 %				Pf00300	6.871
18	YDL182W LYS20 SGDID:S000000001, Chr III from 1206990-1206990, V...	3222.2	2.2 %				Pf00682	9.880
19	YLR441C RPS1A SGDID:S000000001, Chr III from 1206990-1206990, V...	3100.0	0.0 %				Pf01015	6.913
20	YGL008C PMA1 SGDID:S000000001, Chr III from 1206990-1206990, V...	3003.3	0.3 %				Pf00122; Pf00690;...	7.184
21	YML063W RPS1B SGDID:S000000001, Chr III from 1206990-1206990, V...	2909.9	0.9 %				Pf01015	6.913
22	YCR012W PGK1 SGDID:S000000001, Chr III from 1206990-1206990, V...	2811.1	1.1 %				Pf00162	4.252
23	YDR365W-B YDR365W-B SGDID:S0000007401, Chr IV from 1206990-12...	2558.46	10.77 %				Pf00665; Pf01021	9.241
24	YBL027W RPL19B SGDID:S000000123, Chr II from 168426-168427,1...	2507.36	35.45 %				Pf01280	3.246
25	YER026C CHO1 SGDID:S000000828, Chr V from 208473-207643, re...	2269.25	14.49 %				Pf01066	4.068

select GO slim terms as filter from the Row Filter dropdown

# Annotated PTM's from UniProt

Coverage ProteinCard

TDH2 SGDID:S000003769, Chr X from 454673-453675, reverse complement, Verified ORF, "Glyceraldehyde-3-phosphate dehydrogenase, isozyme 2, involved in glycolysis and gluconeogenesis; tetramer that catalyzes the reaction of glyceraldehyde-3-phosphate to 1,3 bis-phosphoglycerate; detected in the cytoplasm and cell wall"

Annotate PTMs reported in Uniprot  
 Show only PTMs  
 Include PSMs That Are Filtered Out

Coverage: 47.59%

Found Modifications:

P Phospho (Y,T,S)

Uniprot PTMs:

P Phospho

PTM Site Probabilities:

- 25 - 45%
- 45% - 75%
- 75% - 99%
- 99% - 100%

Sequence Modification List

	1	11	21	31	41	51	61	71	81	91
Modifications	1									
YJR009C	MVRVRIINGFG	RIGRLVMRIA	LQRKNVEVVA	LNDPFISNDY	SAYMFKYDST	HGRYAGEVSH	DDKHIIVDGH	KIATFQERDP	ANLPWASLNI	DIADISTGVF
Uniprot						PP	P	P		
Modifications					P					P
YJR009C	KELDTAQKHI	DACAKKVVIT	APSSTAPMEV	MGVNEEKYTS	DLKIVSNASC	TTNCLAPLAK	VINDAFGIEE	GLMTTVHSMT	ATQKTVDGPS	HKDWRGGRTA
Uniprot						P				P
Modifications	201	PPP		P						
YJR009C	SGNIIPSSTG	AAKAVGVKLP	ELQGLTGMA	ERVPTVDVSV	VDLTVKLNKE	TTYDEIKKVV	KAAAECKLAC	VLCYTEDAVV	SSDFLGDNSN	SIFDAAAGIQ
Uniprot	P									
Modifications										
YJR009C	LSPKFKLVLS	WYDNEYGYST	RVVDLVEHVA	KA						
Uniprot		P		PP						

Known phosphorylation sites from UniProt annotated in protein sequence



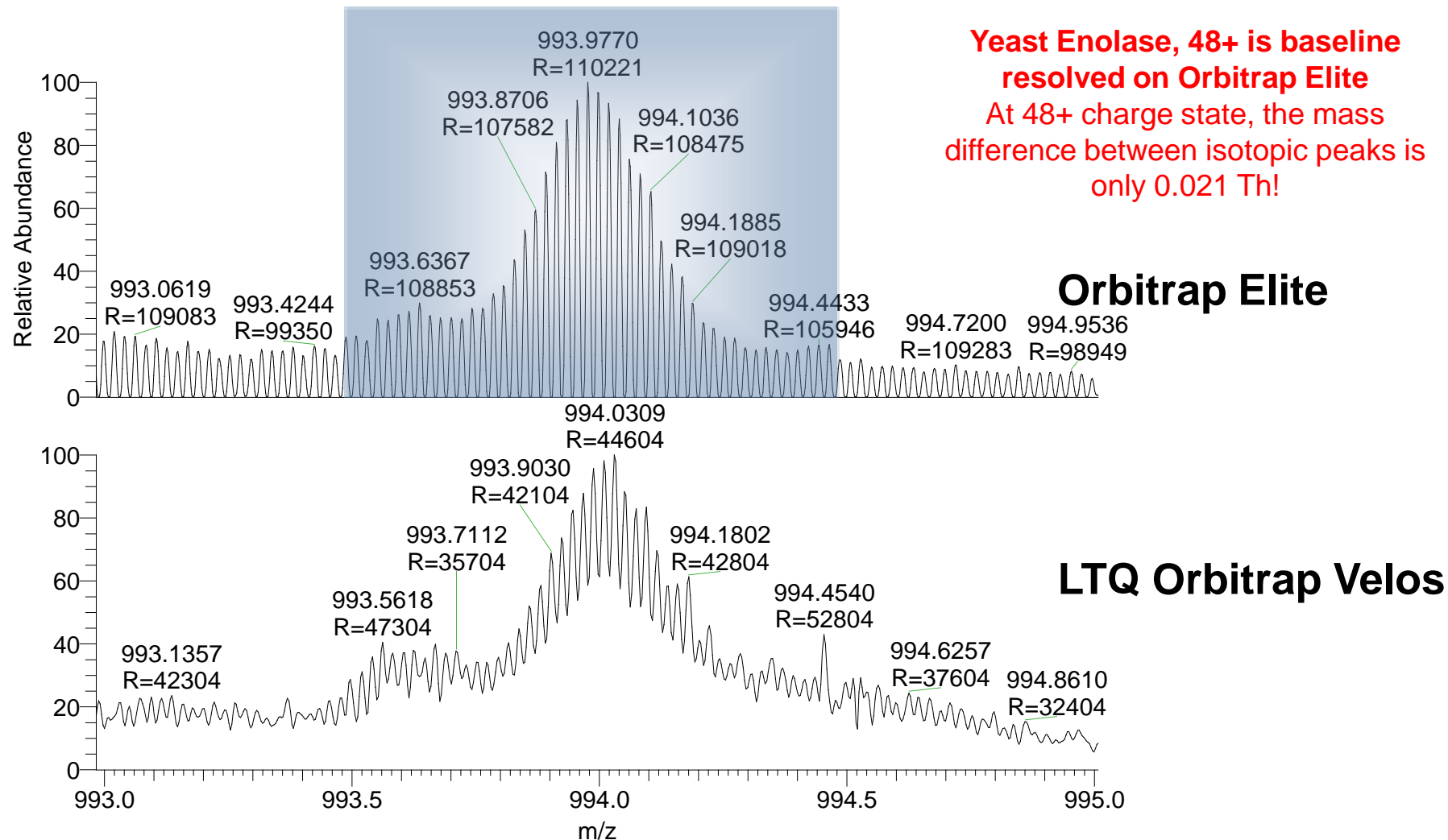
# Conclusions

---

- Proteome Discoverer 1.3 has the most comprehensive set of tools for analysis of proteomics data from Thermo mass spectrometers
- The new nodes developed in collaboration with 3<sup>rd</sup> parties demonstrate the flexibility of the Proteome Discoverer node-based workflow engine and hint at the expansion in the number of tools available for our customers
- Integration with ProteinCenter annotation adds biological inference to SEQUEST and Mascot search results

● **Supplementary Slides**

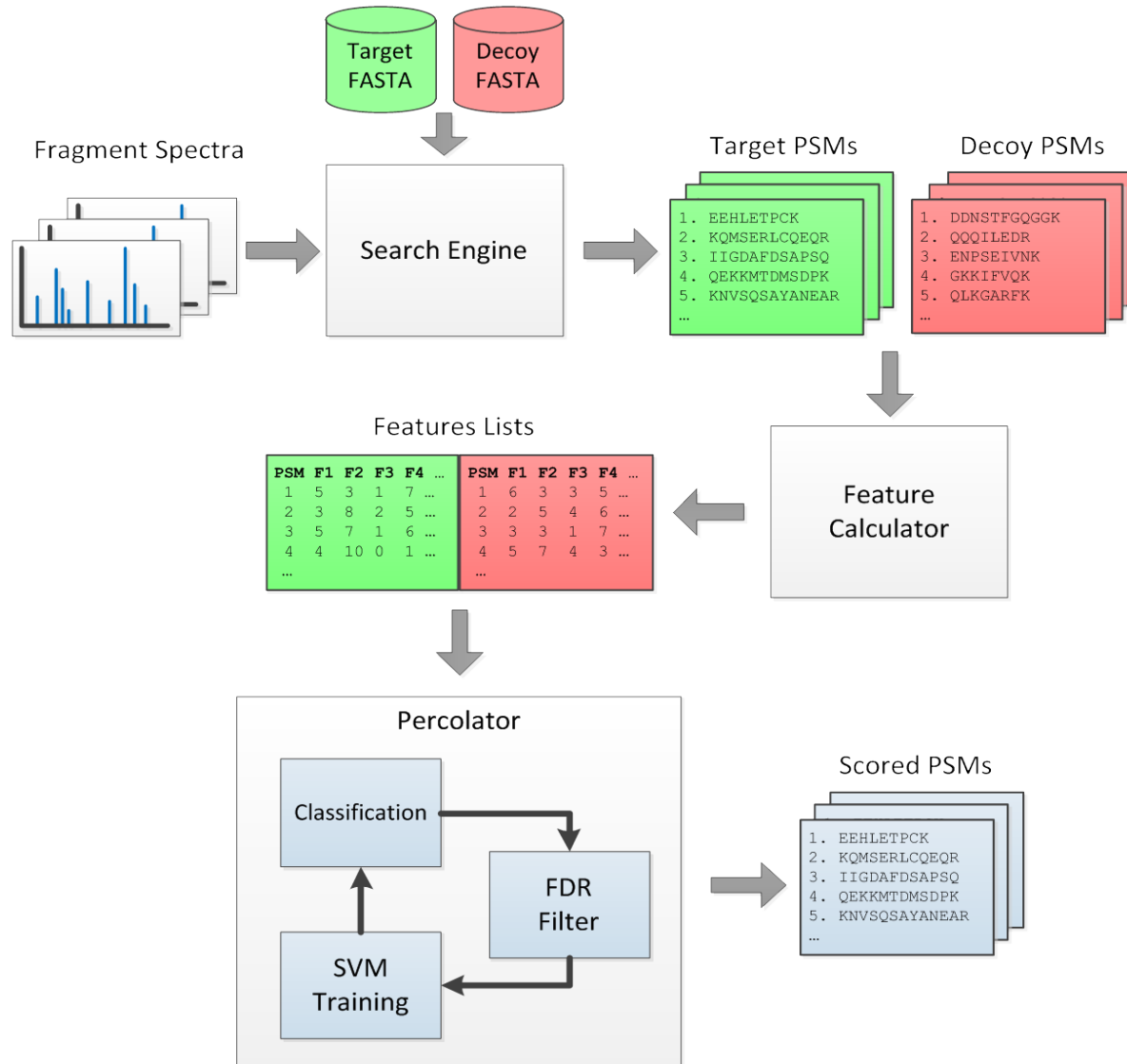
# Intact protein analysis example: Yeast Enolase (48 kDa)



# Percolator

- The Percolator algorithm (<http://per-colator.com>) uses a set of features related to the quality of the peptide-spectrum matches (PSMs) with a semi-supervised method to train a machine learning algorithm called a support vector machine (SVM) to discriminate between correct and incorrect matches
- Does not require any expert-driven or subjective decisions, thereby eliminating any artificial biases
- The learnt classifier is specifically adapted and unique for each data set, thus, adapting to variations in data quality, protocols and instrumentation
- Percolator improves the sensitivity of existing database search algorithms at a constant false discovery rate
- Furthermore, Percolator assigns a statistically meaningful q-value to each PSM, as well as the probability of the individual PSM being incorrect

# Percolator



# q-values and Posterior Error Probabilities

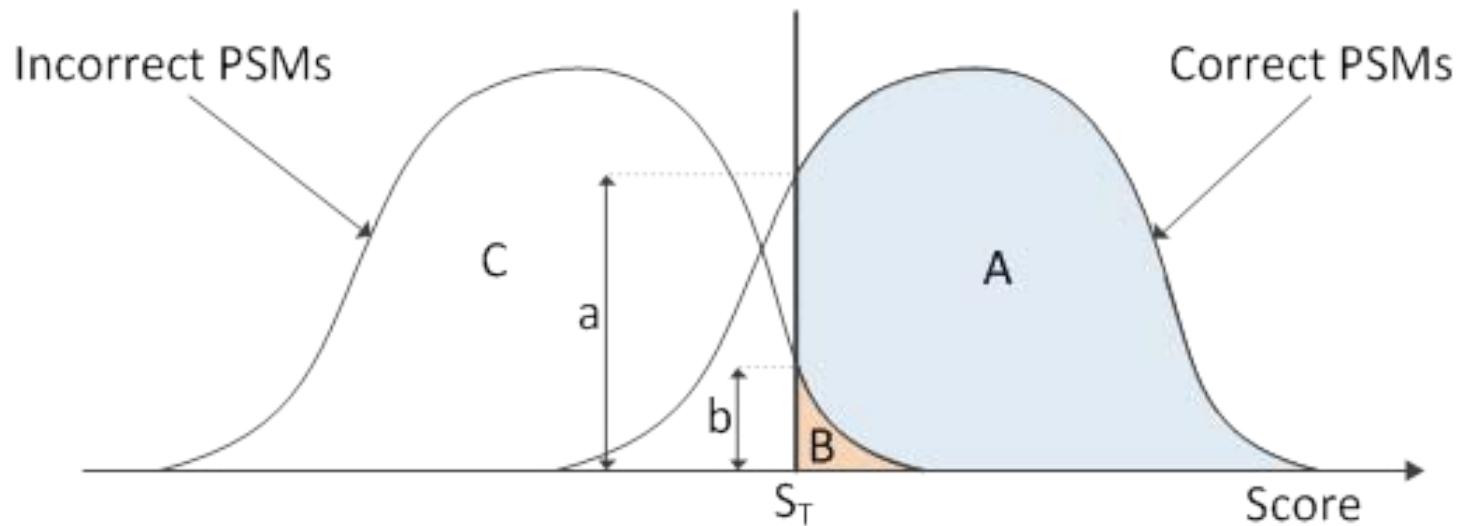
## q-value

- the minimal FDR at which the identification is deemed correct
- Although associated with a single PSM, the q-value depends upon the data set in which the PSM occurs
- A q-value of 0.01 for peptide EAMRQPK matching spectrum  $s$  means that, if we try all possible FDR thresholds, then 1% is the minimal FDR at which this PSM will appear in the output list

## Posterior Error Probability (PEP)

- quite simply, the probability that the observed PSM is incorrect.
- The PEP can be thought of as a local version of the FDR ( “local FDR”). Whereas the FDR measures the error rate associated with a collection of PSMs, the PEP measures the probability of error for a single PSM.
- if the PEP of PSM (EAMRQPK, $s$ ) is 5%, there is a 95% chance that peptide EAMRQPK was in the spectrometer when the spectrum  $s$  was generated

# q-values and Posterior Error Probabilities



$$FDR = \frac{A}{A + B}$$

$$PEP = \frac{a}{a+b}$$

# q-value or PEP: Which one is better?

PEPs and q-values are complementary, and are useful in different situations:

- If you are interested in determining which proteins are expressed in a certain cell type under a certain set of conditions, or if your follow-up analysis will involve looking at groups of PSMs, for example, considering all proteins in a known pathway, evaluating enrichment with respect to GO categories, or performing experimental validation on a group of proteins, then the q-value is an appropriate measure.
- If the goal of your experiment instead is to determine the presence of a specific peptide or protein, then the PEP is more relevant. For example, imagine that you are interested in determining whether a certain protein is expressed in a certain cell type under a certain set of conditions. In this scenario you should examine the PEPs of your detected PSMs. Likewise, imagine that you have identified a large set of PSMs using a q-value threshold, and among them, you identify a single PSM that is intriguing. Before deciding to dedicate significant resources to investigating a single result, you should examine the PEP associated with that PSM. This is because, although the q-value associated with that PSM may be 0.01, the PEP is always greater than or equal to 0.01. In practice, the PEP values for PSMs near the  $q = 0.01$  threshold are likely to be much larger than 1%.



# Percolator in Proteome Discoverer

- Available as a new node under “Peptide Validation”

The screenshot displays the Thermo Proteome Discoverer 1.3 Workflow Editor. The left sidebar shows a tree of workflow nodes, with "Peptide/Protein Validation" highlighted and circled in red. A red arrow points from this category to the "Percolator" node (node 3) in the main workflow diagram. The workflow diagram shows a sequence of nodes: "Spectrum Files" (0) leads to "Spectrum Selector" (1), which then branches into "Mascot" (2) and "SEQUEST" (4). Both "Mascot" and "SEQUEST" lead to the "Percolator" (3) node. The "Parameters" panel on the right shows the following settings:

Parameters	
Show Advanced Parameters	
1. Input Data	
Maximum Delta Cn	0.05
2. Decoy Database Search	
Target FDR (Strict)	0.01
Target FDR (Relaxed)	0.05
Validation based on	q-Value

Below the parameters, a section titled "Maximum Delta Cn" explains: "Peptides having a delta Cn better or equal than this are used for Percolator. Minimum value = 0, Maximum value = 1".

**Note:** Percolator needs a sufficient number of PSMs from the target and the decoy search. If less than 200 target or decoy PSMs were identified, the percolation is rejected. The same is true if less than 20% decoy PSMs are available compared to the number of target matches.

# Percolator in Proteome Discoverer

- Proteome Discoverer uses a set of more than 30 features describing the quality of a PSM

Search engine/scoring specific features			
Mascot Ions Score		Only used for Mascot searches	
SEQUEST Xcorr		Only used for SEQUEST searches	
SEQUEST	<b>Peptide/Precursor related features</b>		
X! Tandem	% Isolation Interference	Fraction of ion current in the isolation window not attributed to the identified precursor	
Delta Cn			
MH+ [Da]	<b>Fragment series related features</b>		
Delta Mass	Fragment Coverage Series A, B, C [%]	Coverage of the N-terminal fragment ion series. The coverage is separately calculated for each series used and the maximum coverage is used.	
Delta Mass			
Absolute De	Fragment Coverage Series X, Y, Z [%]	Coverage of the C-terminal fragment ion series.	
Absolute De	<b>Fragment related features</b>		
Peptide Len	Log Matched	IQR Fragment Delta Mass [Da]	
Is z=1		Inter-quartile range of the distribution of mass errors in Dalton of all fragments considered.	
Is z=2	Log Matched	IQR Fragment Delta Mass [ppm]	
Is z=3		Inter-quartile range of the distribution of mass errors in ppm of all fragments considered.	
Is z=4	Longest Sequ	Mean Fragment Delta Mass [Da]	
Is z=5		Arithmetic mean range of the distribution of	
Is z>5	Longest Sequ	<b>Digestion related features</b>	
		# Missed Cleavages	Number of missed cleavages
		<b>FASTA related features</b>	
		Log Peptides Matched	Logarithm of the number of candidates in the precursor mass window
		<b>Spectrum related features</b>	
		Log Total Intensity	Logarithm of the total ion current of the fragment spectrum
		Fraction Matched Intensity [%]	Fraction of the total ion current of the fragment spectrum that is matched by fragments of the PSM

# Percolator in Proteome Discoverer

The screenshot shows the Thermo Proteome Discoverer 1.3.0.339 interface. The 'Peptide Confidence' tab is active, displaying the 'Percolator' section with a dropdown menu and buttons for 'Set Filter Type', 'Apply Filters', and 'Apply FDRs'. Below this is a table with columns for 'Processing Node Name', 'Workflow Name', and 'Short Display Name', showing 'Mascot', 'SILAC3', and 'A2'. To the right, there are two filter settings panels: 'Modest Confidence Filter Settings' and 'High Confidence Filter Settings', both showing 'FDR Settings' with a 'Threshold' of 0.05 and 'Validation based on' set to 'q-Value'. Red callouts provide instructions: one points to the 'Apply Filters' button, another points to the 'q-Value' field in the 'High Confidence Filter Settings' panel, and a third points to the 'Apply Filters' button again.

Click "Apply Filters" after changes

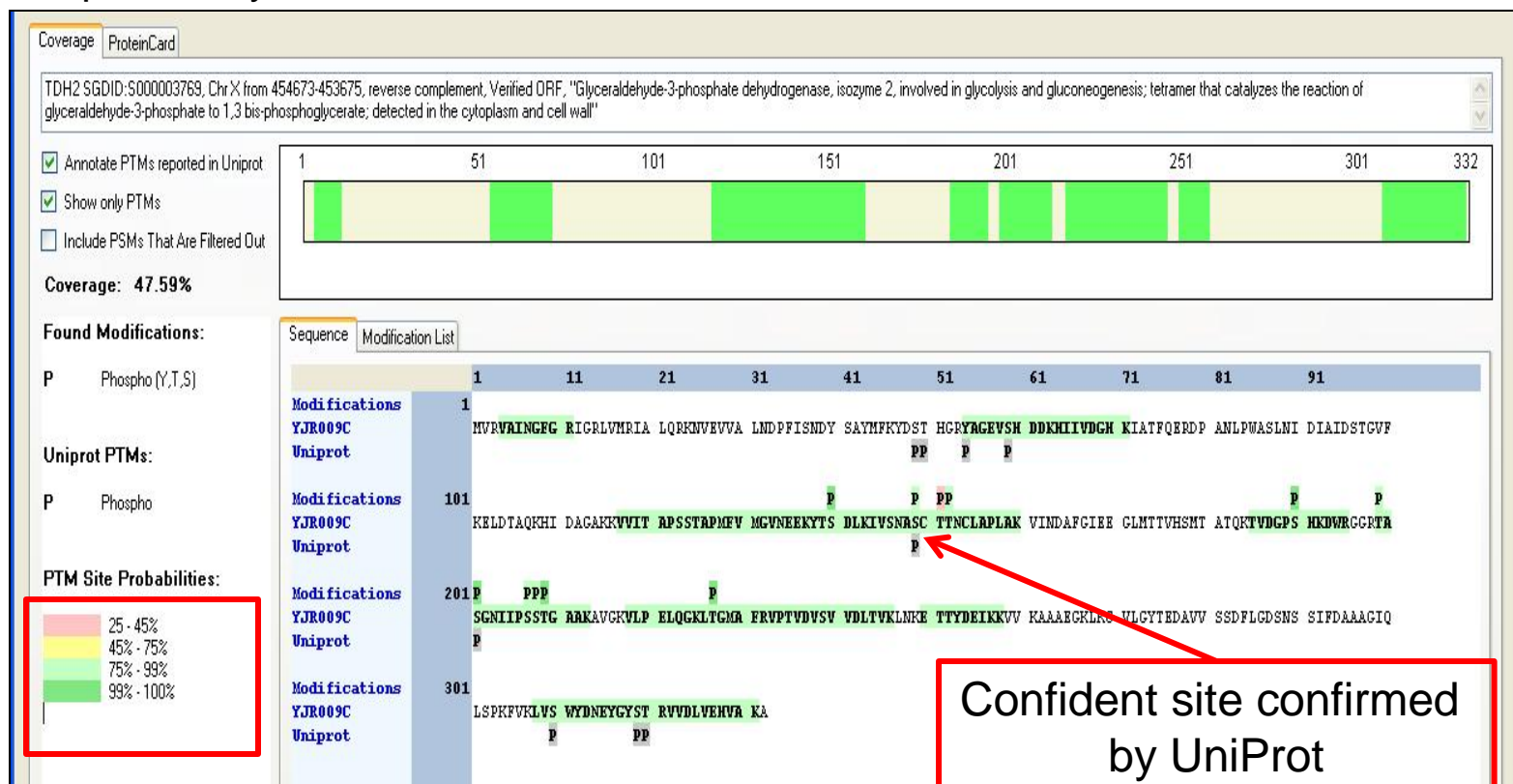
Define the q-value (or PEP) thresholds to divide PSMs in high, medium, and low confident hits on the Peptide Confidence tab

Define whether the confidence estimation is based on Percolator scores (default when Percolator was used) or classical FDR on score thresholds

Ready 1122/1129 Protein Group(s), 1321/1321 Protein(s), 5758/5758 Peptide(s), 42326/42326 PSM(s), 70574/70574 Search Input(s)

# Confident PTM analysis Using PhosphoRS

- Developed in collaboration with Karl Mechtler's lab at University of Vienna for phosphorylation site confidence determination
  - phosphoRS Score: is it phosphorylated?
  - site probability Score: confidence in site localization



# phosphoRS

---

- automated calculation of individual site probabilities for each putative phospho-site
- works for all common fragmentation techniques (CID, ETD, and HCD) and all available database search engines
- validated and optimized by analysis of LC-MS/MS data of more than 150 synthetic phospho-peptides with precisely known phospho-sites

# phosphoRS



Probability  $P_{S,i}$  to match  $k_S$  or more peaks purely by chance:

$$P_{S,i}(X \geq k_S) = \sum_{k=k_S}^n \binom{n}{k} p^k (1-p)^{n-k}$$

$n$ : number of theoretical fragment ions  
 $k_S$ : number of matched fragment peaks  
 $p$ : probability to match a fragment by chance

*pRS*-Scores:

$$pRS_{S,i} = -10 \cdot \log_{10}(P_{S,i}(X \geq k_S))$$

$$pRS \text{ Sequence Probability} = \frac{P_{S,i_{optimal}}^{-1}}{\sum_S P_{S,i_{optimal}}^{-1}}$$

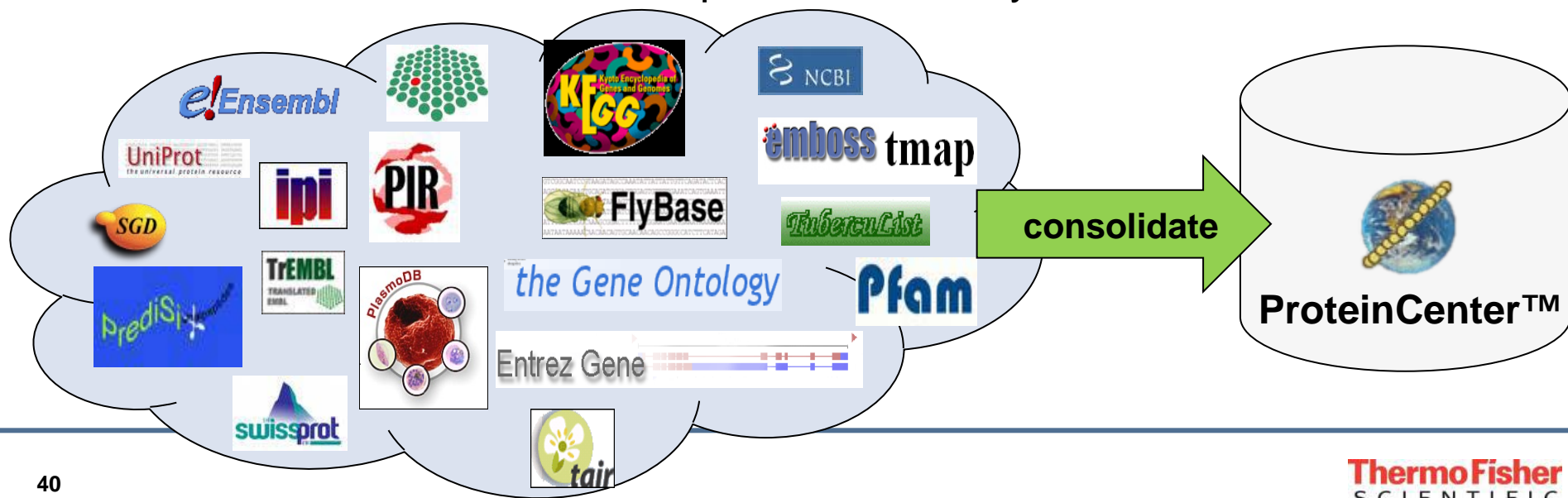
$$pRS \text{ Site Probability} = \sum_{\text{all isoforms where this site is phosphorylated}} pRS \text{ Seq. Prob.}_i$$

# phosphoRS

- pRS Score
  - This peptide score is based on the cumulative binomial probability that the observed match is a random event. The value of the pRS Score strongly depends on the data scored, but usually scores above 50 give good evidence for a good PSM.
- pRS Sequence Probability
  - This value estimates the probability (0-100%) that the respective isoform is correct.
- pRS Site Probabilities
  - For each phosphorylation site this is an estimation of the probability (0-100%) for the respective site being truly phosphorylated. pRS Site Probabilities above 75% are good evidence that the respective site is truly phosphorylated.

# ProteinCenter™

- Protein-centric data warehouse specifically designed for interpretation of proteomics data
- Enables the comparison of data sets searched against different databases and different versions of databases
- > 16 million protein sequences from the major public protein databases distilled from 130 million accession codes from past and present versions
- The consolidated database is updated bi-weekly



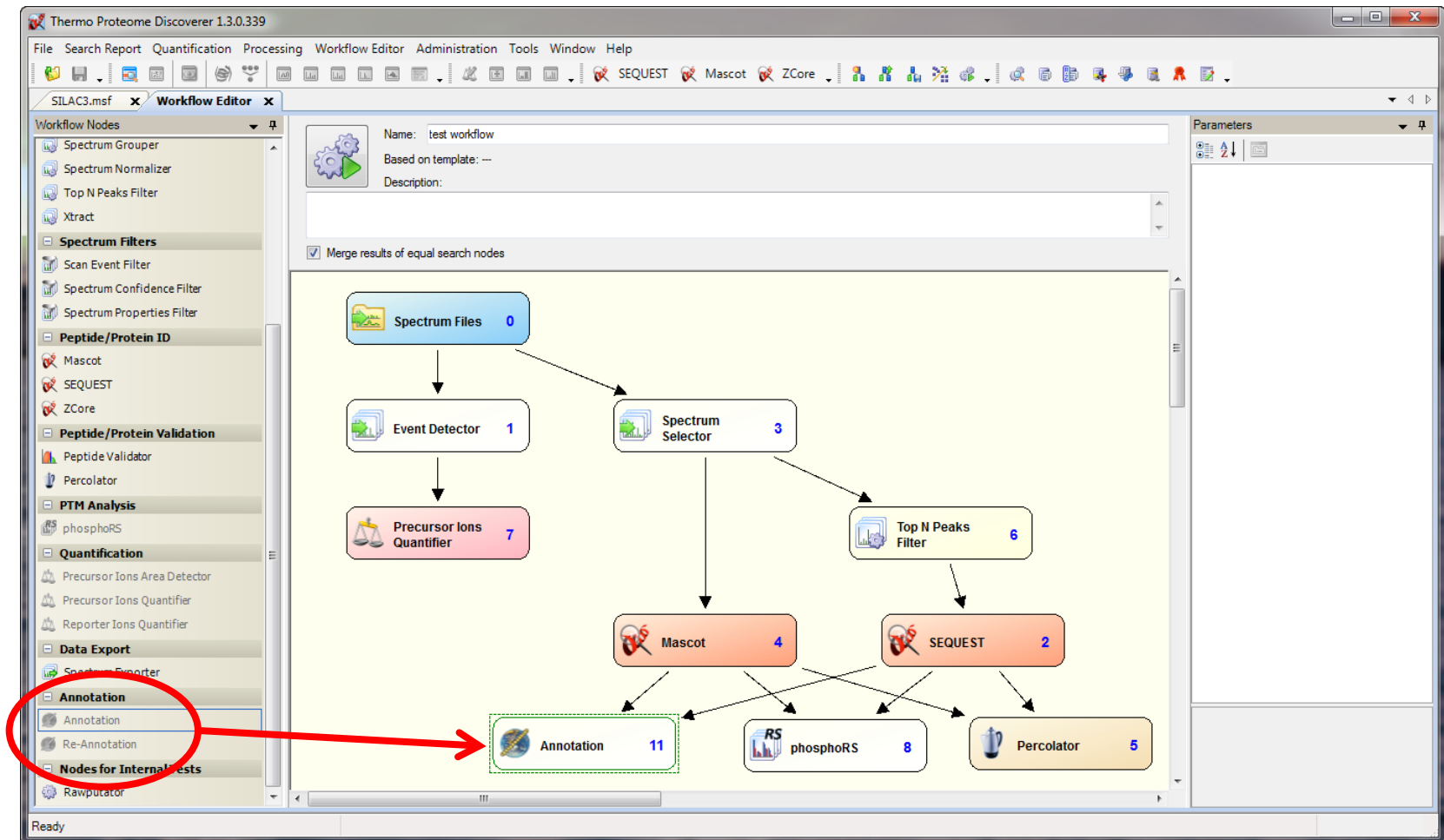


# Annotate Identified Proteins Using the ProteinCenter™ Repository

- Provides concise, precise and focused annotation for a given protein
- Use the Annotation node to directly retrieve protein annotations from the web service provided by ProteinCenter
- Currently the following annotations are retrieved:
  - Gene Ontology (GO, <http://www.geneontology.org>) and GO slim annotations
    - provides a widely used controlled vocabulary to describe the function, localization and process a protein is assigned to
    - approx. 35000 terms is organized in dependency hierarchies.
    - Subsets taken on a high level of this hierarchy (termed GO slims) are widely used to give an overview of the biological impact of a molecule
  - Protein family (Pfam, Wellcome Trust Sanger Institute) annotations
  - Modifications documented in the Uniprot database

# Use the Annotation Node in a Workflow

- The Annotation node automatically connects with every search node



# (Re) Annotate Existing .msf Files

- Use the Re-Annotation node to annotate existing result files that do not yet contain annotations or update existing annotations
- Can be automated with the Discoverer Daemon

The image shows two screenshots from Thermo Proteome Discoverer 1.3. The top screenshot is the Workflow Editor, showing a workflow named 'SILAC\_HeLa' with a 'Re-Annotation' node highlighted. The bottom screenshot is the Discoverer Daemon 1.3 interface, showing a list of .msf files in the 'Load Files' section and a 'Start' button. Red arrows point from the workflow editor to the daemon interface, and yellow callout boxes provide instructions for each step.

2. provide a folder name on the server

1. use the provided Re-Annotation workflow template

3. select the .msf files to annotate

4. Start processing

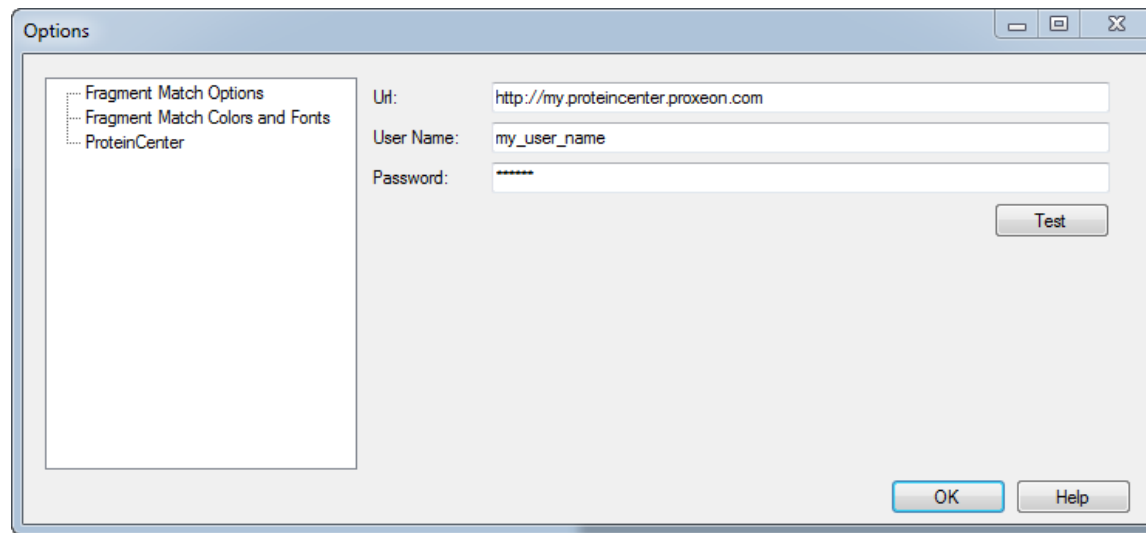
Workflow Editor: Name: SILAC\_HeLa, Based on Template: --, Description: --, Merge Results of Equal Search Nodes:

Discoverer Daemon 1.3: Spectrum Files, Batch processing selected, Load Files: Add..., Remove, Start, List of files: D:\ExampleData\Reannot\SILAC\_HeLa\_01.msf, D:\ExampleData\Reannot\SILAC\_HeLa\_02.msf, D:\ExampleData\Reannot\SILAC\_HeLa\_03.msf, D:\ExampleData\Reannot\SILAC\_HeLa\_04.msf, D:\ExampleData\Reannot\SILAC\_HeLa\_05.msf

Discoverer Daemon 1.3: Workflow: Re-Annotation, Server Output Directory: Subdirectory on: localhost Path: z:\MagellanRoot\PublicFiles\DiscovererDaemon\SpectrumFiles, ReAnnot: [text box], Output Filename: [text box], Log output: Processing completed workflow 'SILAC\_HeLa\_05', Processing completed workflow 'SILAC\_HeLa\_04', Processing completed workflow 'SILAC\_HeLa\_03', Processing completed workflow 'SILAC\_HeLa\_02', Processing completed workflow 'SILAC\_HeLa\_01', Processing workflow 'SILAC\_HeLa\_05', Registering 'SILAC\_HeLa\_05.msf', Processing workflow 'SILAC\_HeLa\_04', Registering 'SILAC\_HeLa\_04.msf', Processing workflow 'SILAC\_HeLa\_03', Registering 'SILAC\_HeLa\_03.msf', Processing workflow 'SILAC\_HeLa\_02', Registering 'SILAC\_HeLa\_02.msf', Processing workflow 'SILAC\_HeLa\_01', Registering 'SILAC\_HeLa\_01.msf', Construct spectrum file collection 'ReAnnot\_20110222'

# Automatic Transfer of Results from Proteome Discoverer to ProteinCenter™

- You need a user account on a ProteinCenter™ server
- Configure your account settings under Tools > Options > ProteinCenter



- Export all or selected proteins with Tools > Export to ProteinCenter