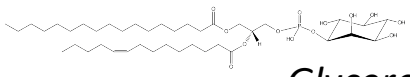
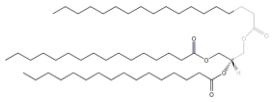
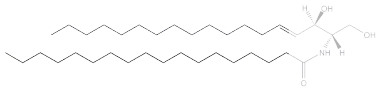


Glycerophospholipids



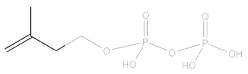
Glycerolipids

Fatty acyls

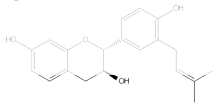


Sphingolipids

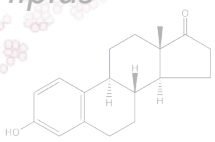
Prenol lipids



Polyketides



Sterol lipids



Computational Tools for Tandem Mass Spectrometry

Scott Walmsley, Ph.D.

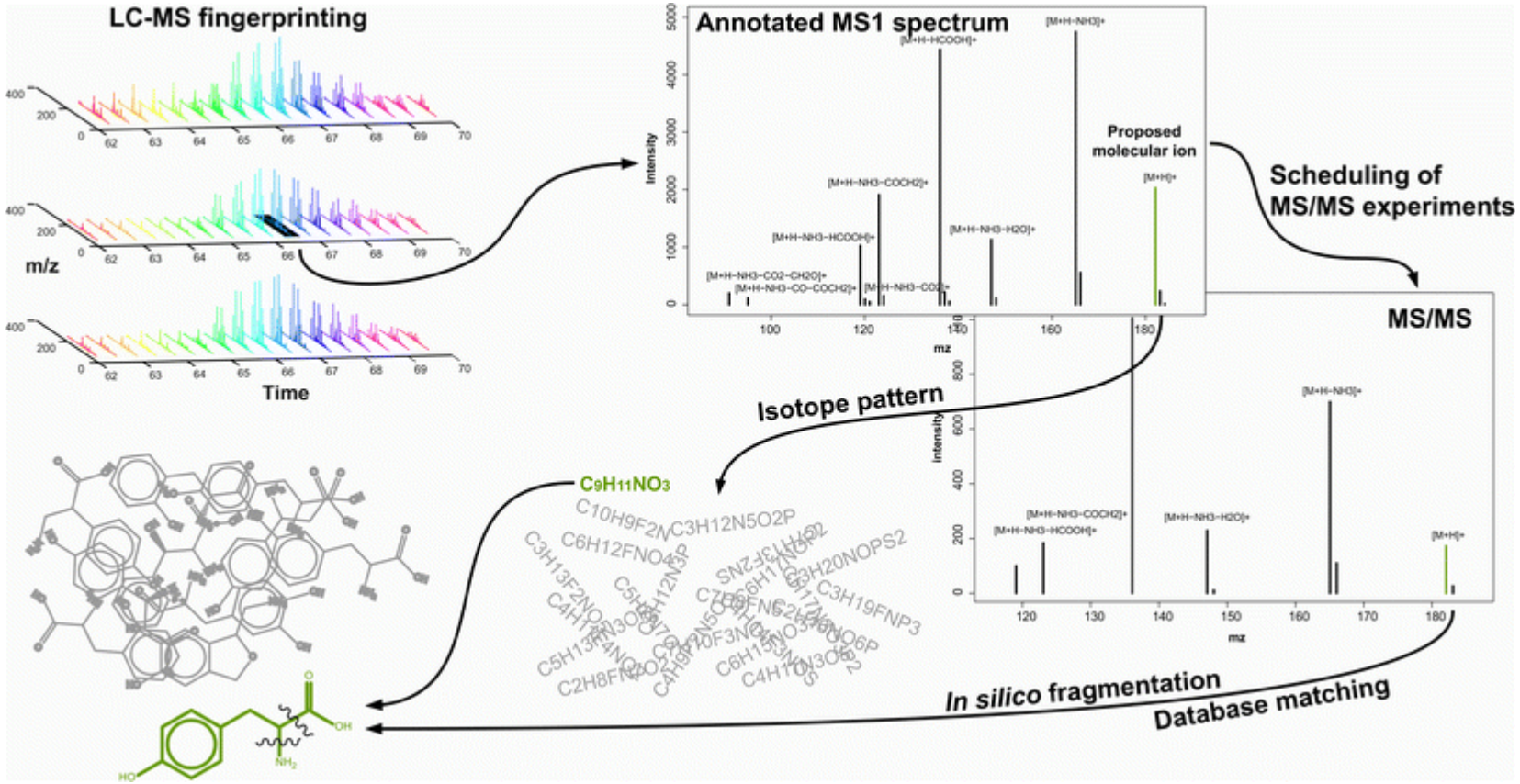
Mass Spectrometry Facility

DOPS-Skaggs SOP

UC Anschutz

Tandem Mass Spectrometry

- Peptide MS/MS
  - Proteomics
- Small Molecule MS/MS
  - Metabolomics, environmental, toxicology, etc…



1: Stanstrup J, Gerlich M, Dragsted LO, Neumann S. Metabolite profiling and beyond: approaches for the rapid processing and a nnotation of human blood serum mass spectrometry data. Anal Bioanal Chem. 2013 Jun;405(15):5037-48. doi: 10.1007/s00216-013-6954-6. Epub 2013 Apr 25. PubMed PMID: 23615935.

“Shotgun” proteomics

1: Meissner F, Mann M. Quantitative shotgun proteomics: considerations for a high-quality workflow in immunology. Nat Immunol. 2014 Feb;15(2):112-7. doi: 10.1038/ni.2781. PubMed PMID: 24448568.

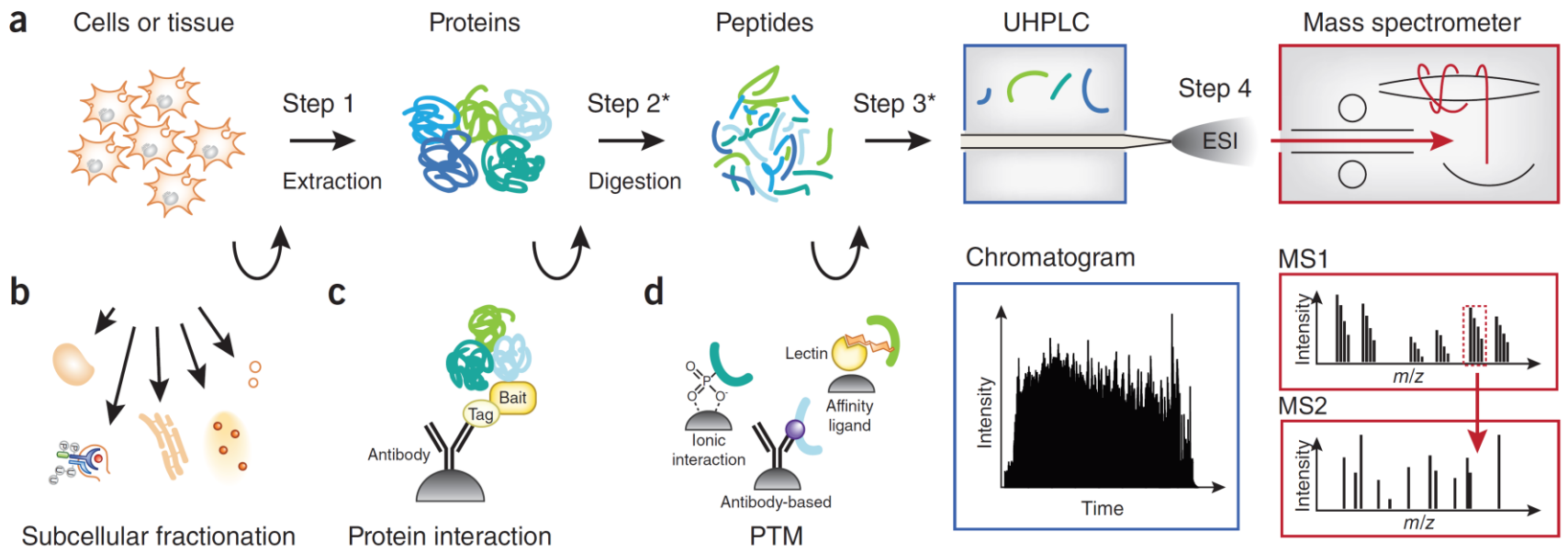
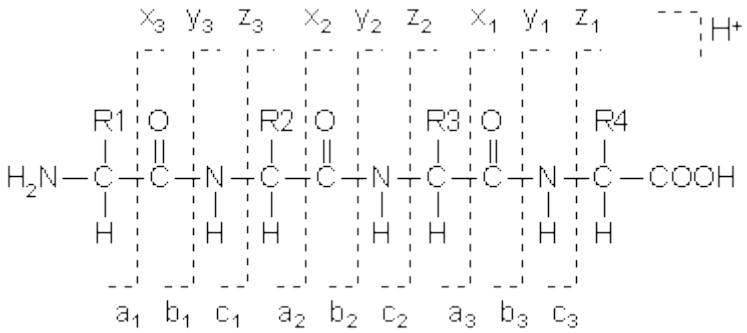
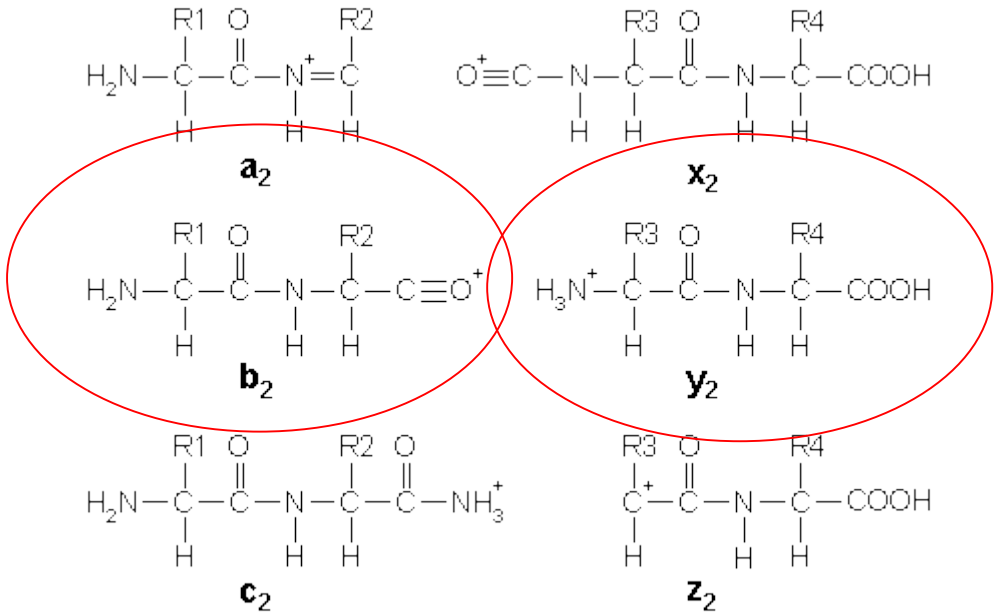


Figure 2 Shotgun proteomics workflow. (**a**) The generic workflow of modern LC-MS–based proteomics consists of four steps. In step 1, proteins are extracted from tissues, body fluids, cells or subcellular compartments. In step 2, proteins are proteolytically digested. In step 3, peptides are separated by UHPLC. In step 4, peptides are ionized by electrospray (electrospray ionization, ESI), and their masses and fragment masses are acquired in a mass spectrometer. In the workflow described here, LC-MS is performed with an UHPLC system coupled online to a Q Exactive (Thermo Fisher Scientific). (**b**–**d**) Variations of the workflow that include enrichment steps for proteins in subcellular compartments (**b**), for interacting proteins (**c**) and for peptides with PTMs (**d**). At steps 2 and 3 (*), additional fractionation of proteins or peptides is possible.

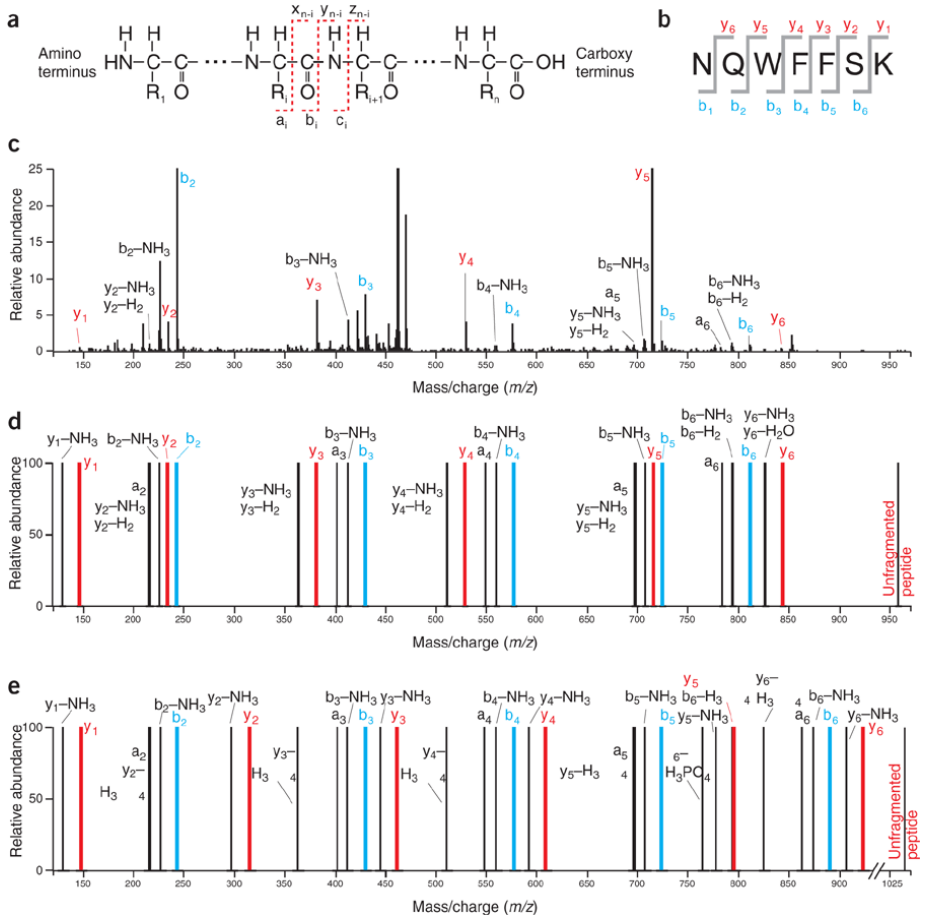
“Tryptic” peptides can be predictably fragmented in a Mass Spectrometer





http://www.matrixscience.com/

Fragmentation by MSMS



Expected fragmentation patterns are matched to an experimental spectra.
WE call this the peptide spectrum match (PSM)
This is performed using a “search engine”

Search engines rank order the “best hit” followed by the next best matches

Marcotte EM. How do shotgun proteomics algorithms identify proteins? Nat Biotechnol. 2007 Jul;25(7):755-7. PubMed PMID: 17621303.

Tandem Mass Spectra

- 1000’s to millions produced each file
- Peptide-spectra match (PSM) accomplished via “search engines”
- Examples of search engines that perform PSMs:
  - Theoretical spectral matching:
    - SEQUEST
    - MASCOT
    - X!Tandem
  - Experimental Library Matching:
    - NIST MS
    - SpectraST

Search engines
→ Match MSMS spectra to protein sequence

- X!TANDEM

  https://www.ncbi.nlm.nih.gov/pubmed/14976030

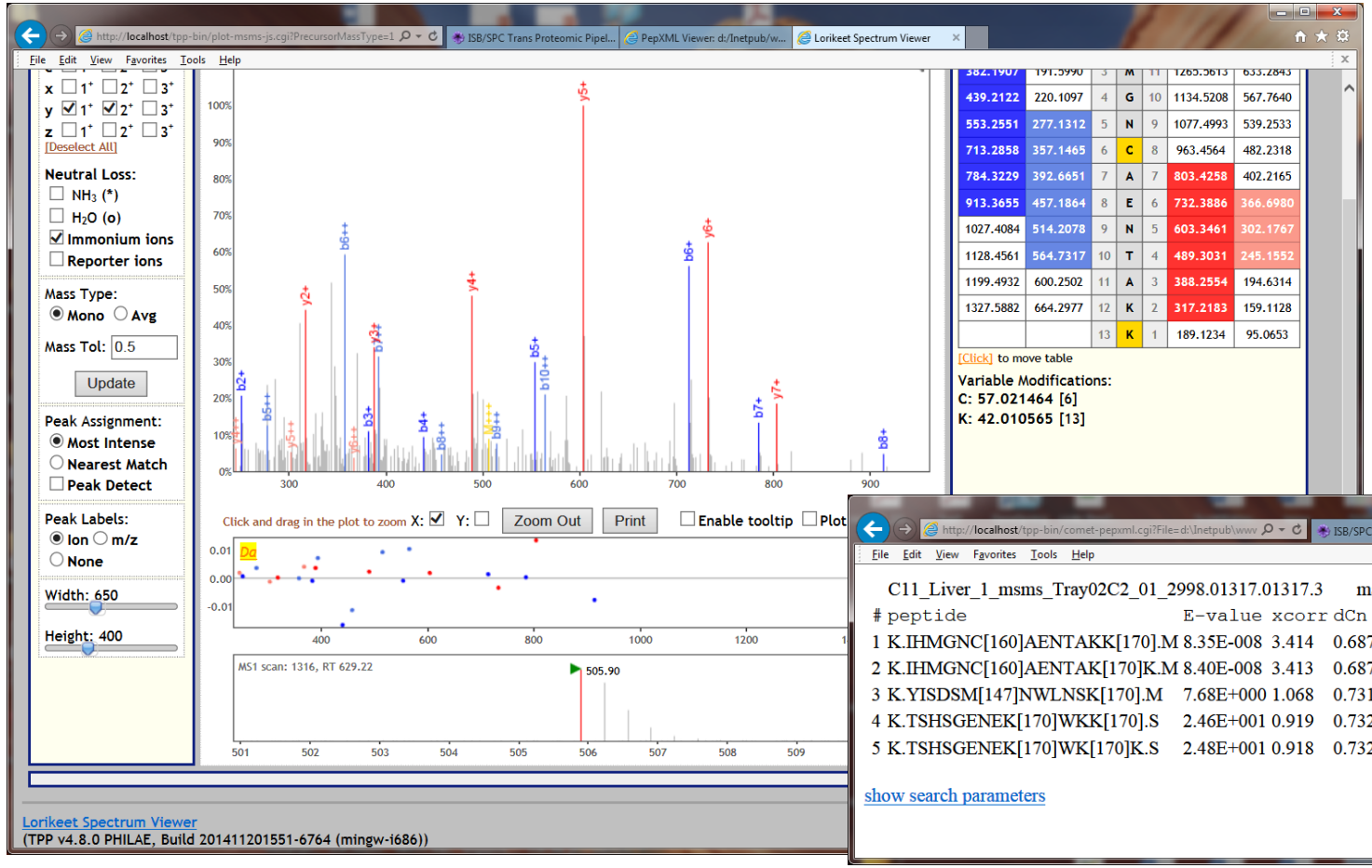
- COMET (SEQUEST)

  https://link.springer.com/article/10.1007%2Fs13361-015-1179-x

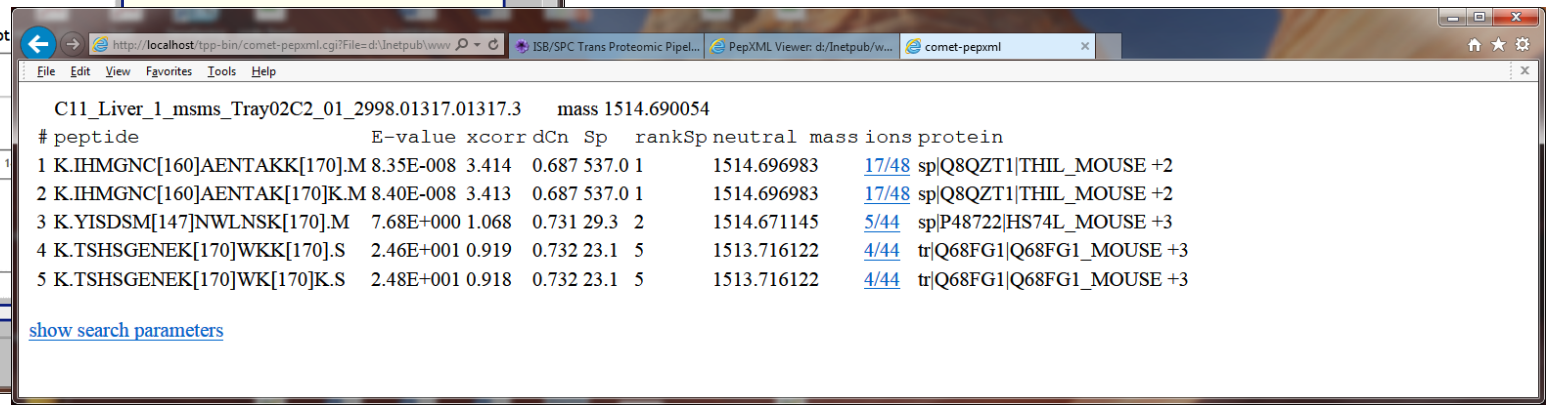
- MASCOT (Mowse)

  https://www.ncbi.nlm.nih.gov/pubmed/15335725?dopt=AbstractPlus

MSMS Search Result (Comet)

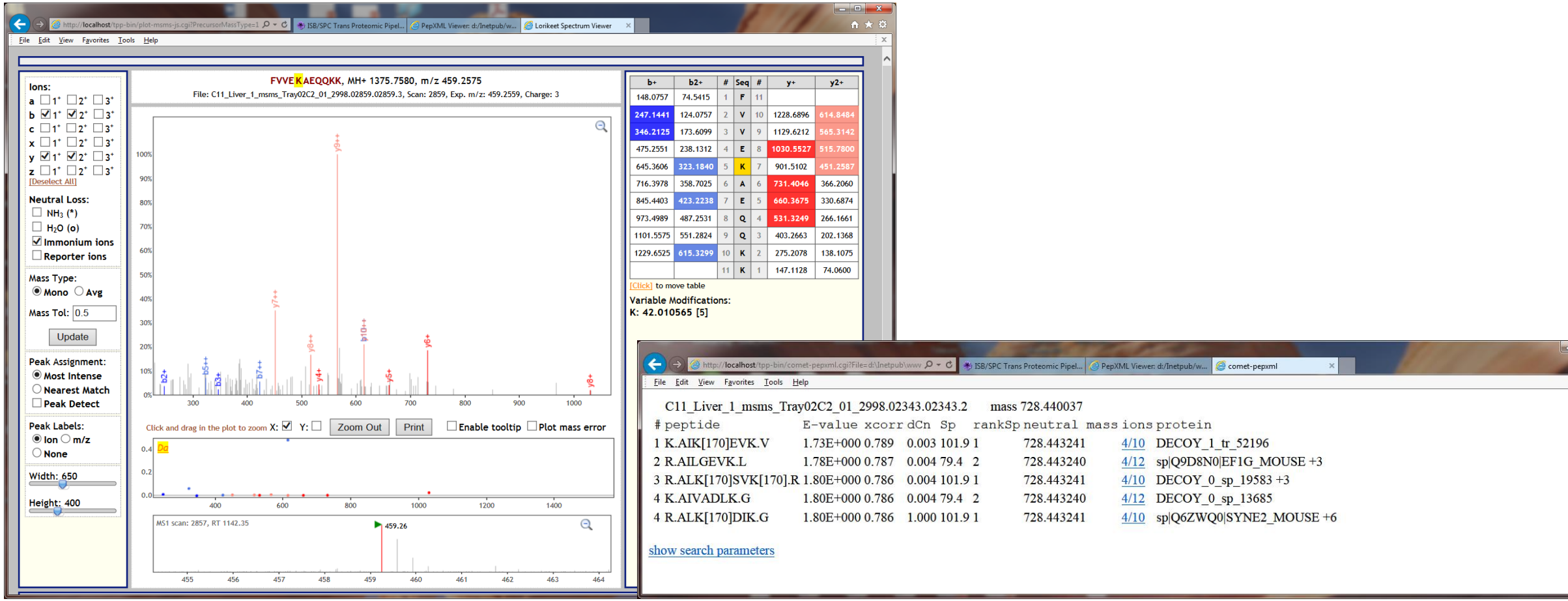


PSM for a good match



MSMS Search Result #2 (COMET)

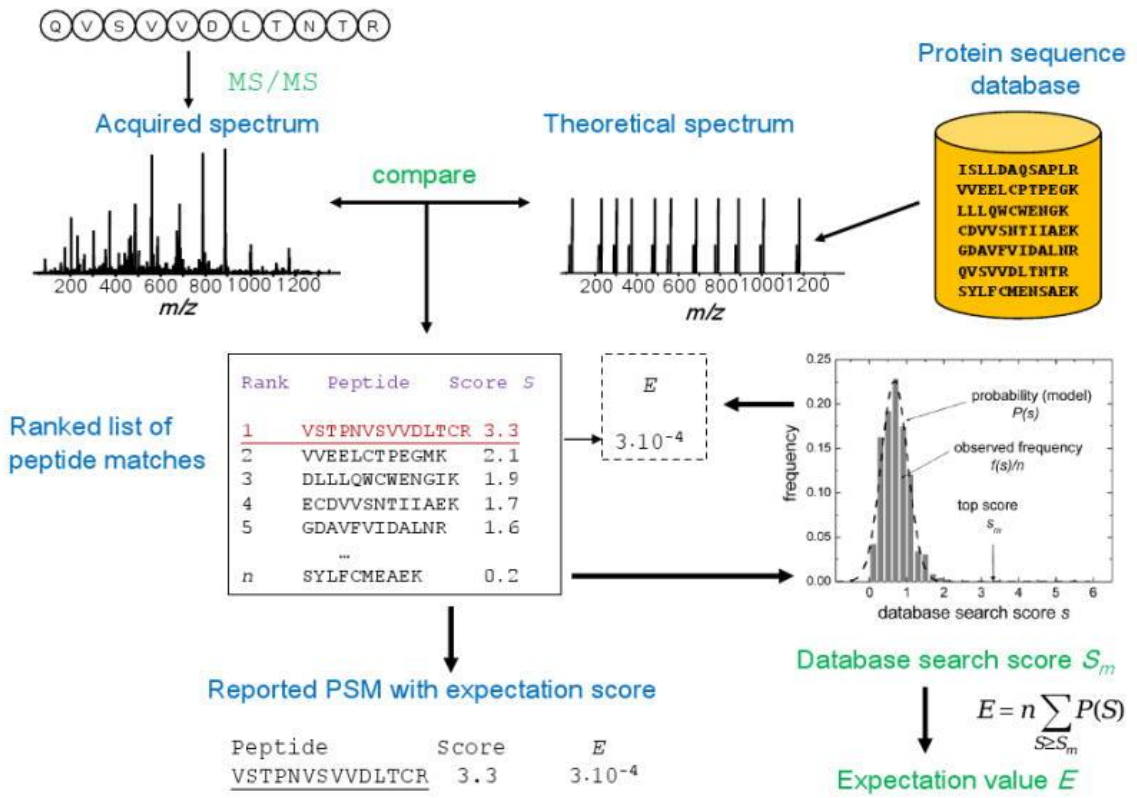
PSM for a bad match



b+	b2+	#	Seq	#	y+	y2+
148.0757	74.5415	1	F	11		
247.1441	124.0757	2	V	10	1228.6896	614.8484
346.2125	173.6099	3	V	9	1129.6212	565.3142
475.2551	238.1312	4	E	8	1030.5527	515.7800
645.3606	323.1840	5	K	7	901.5102	451.2587
716.3978	358.7025	6	A	6	731.4046	366.2060
845.4403	423.2238	7	E	5	660.3675	330.6874
973.4989	487.2531	8	Q	4	531.3249	266.1661
1101.5575	551.2824	9	Q	3	403.2663	202.1368
1229.6525	615.3299	10	K	2	275.2078	138.1075
		11	K	1	147.1128	74.0600

Peptide Identification from MS/MS spectra

1: Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics. 2010 Oct 10;73(11):2092-123. doi: 10.1016/j.jprot.2010.08.009. Epub 2010 Sep 8. Review. PubMed PMID: 20816881; PubMed Central PMCID: PMC2956504.

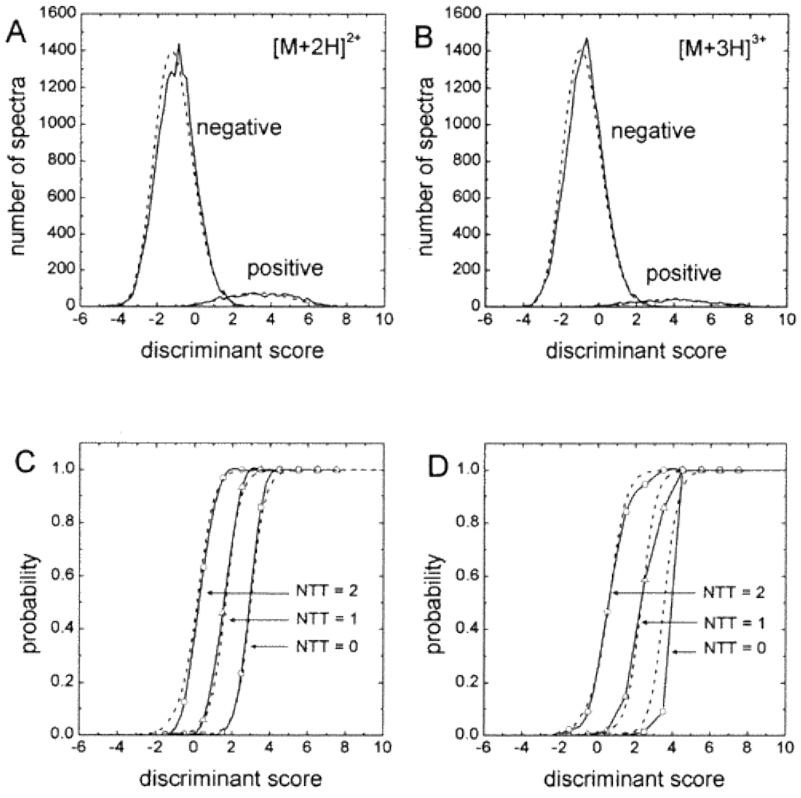


Problems with PSMs

- Different search engines give different results
- Different scoring methods
- Need to ‘unify’ results
- ➔ answer:
  - Modelling of distribution of best hits factoring other effects
  - NTT: number of tryptic termini
  - NMC: number of missed cleavages
  - Mass accuracy
  - Mixture modelling and Bayes

Peptide Validation from MSMS Search Engine

1: Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem. 2002 Oct 15;74(20):5383-92. PubMed PMID: 12403597.



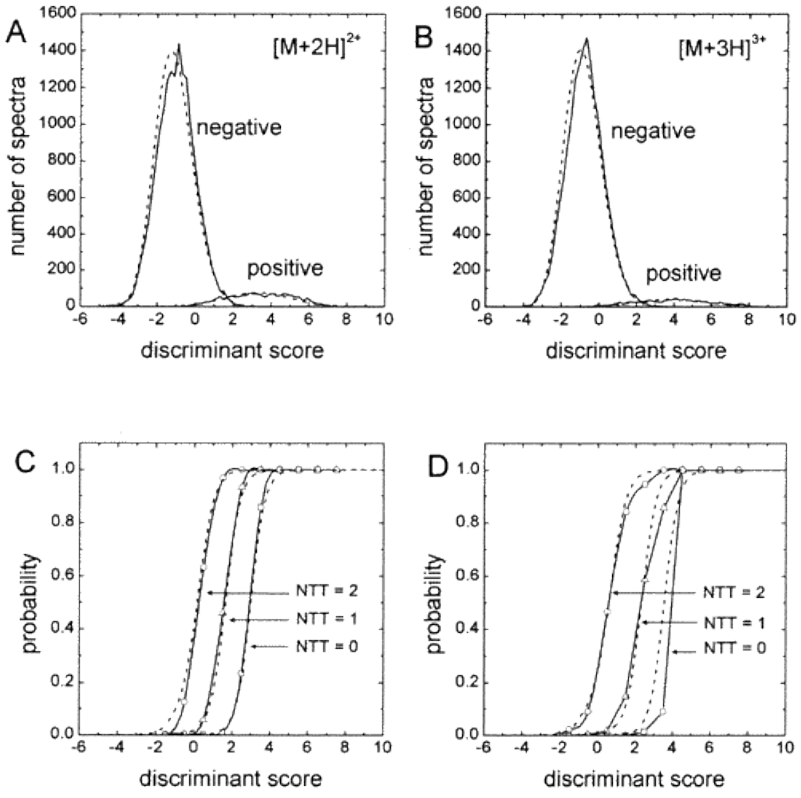
$$p(+|F,\mathrm{NTT}) = \frac{p(F|+)p(\mathrm{NTT}|+)p(+)}{p(F|+)p(\mathrm{NTT}|+)p(+) + p(F|-)p(\mathrm{NTT}|-)p(-)} \quad (8)$$

Models optimum discriminant score

- Parameters include mass, # amino acids, search engine scores
- Separate models for charge state

Peptide Validation from MSMS Search Engine

1: Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem. 2002 Oct 15;74(20):5383-92. PubMed PMID: 12403597.

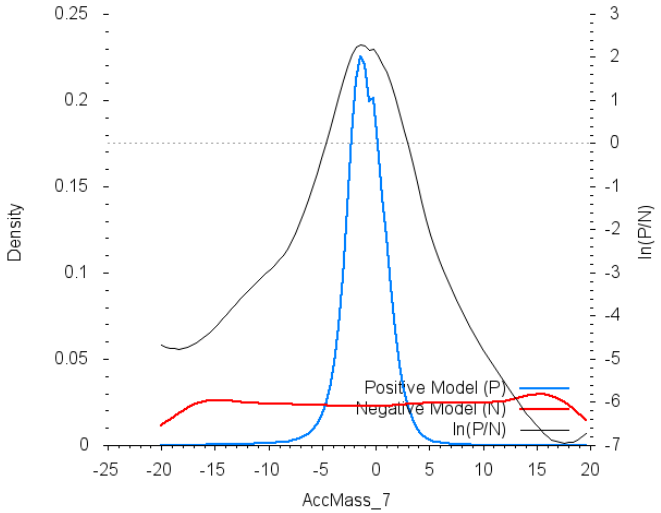


$$p(+|F,\mathrm{NTT}) = \frac{p(F|+)p(\mathrm{NTT}|+)p(+)}{p(F|+)p(\mathrm{NTT}|+)p(+) + p(F|-)p(\mathrm{NTT}|-)p(-)} \quad (8)$$

Goal: ease the burden of manually validating 100,000's spectra

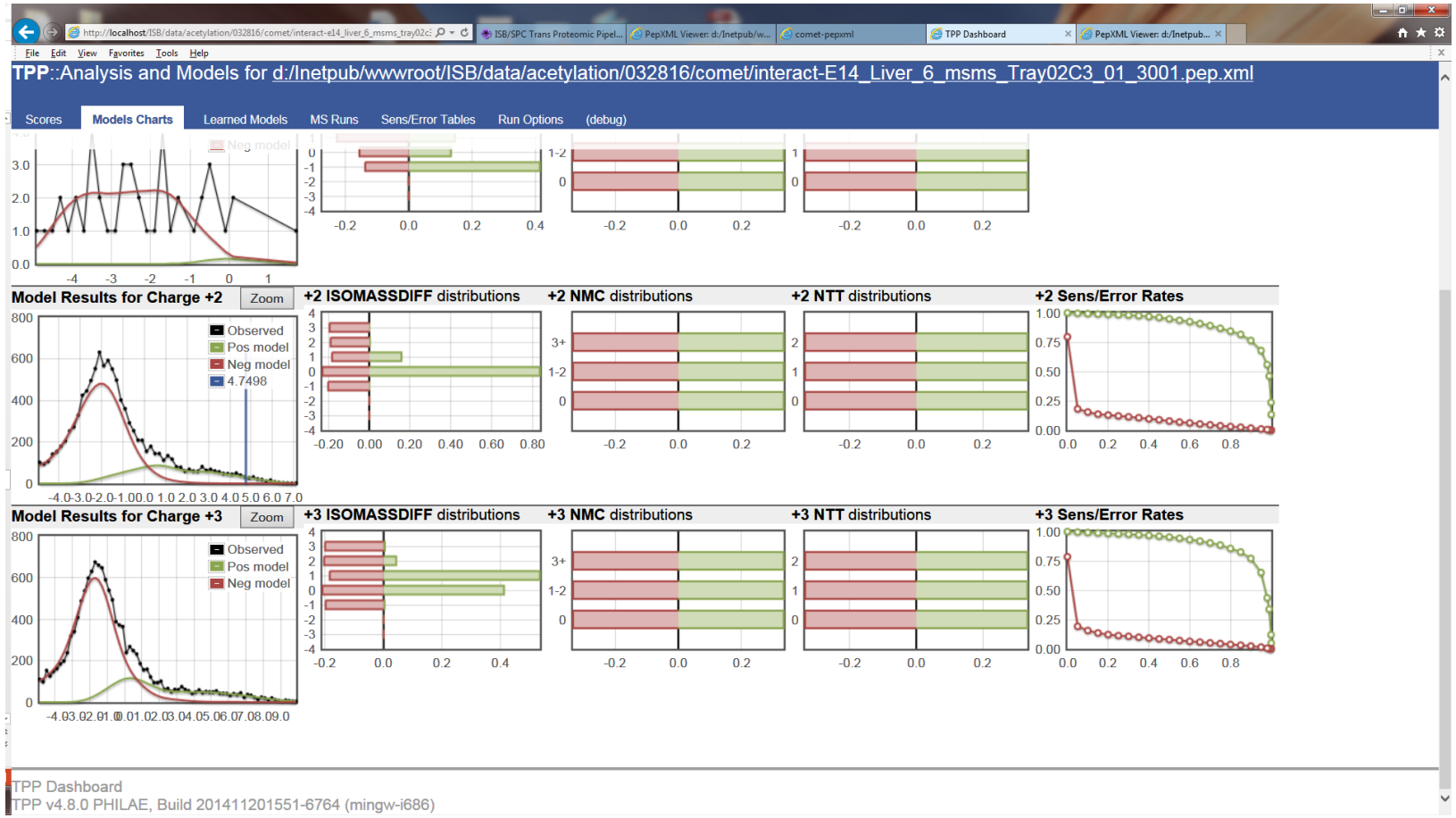
Typical Modelling Results

GOOD model “fit”?



Modelling mass accuracy

Typical Modelling Results



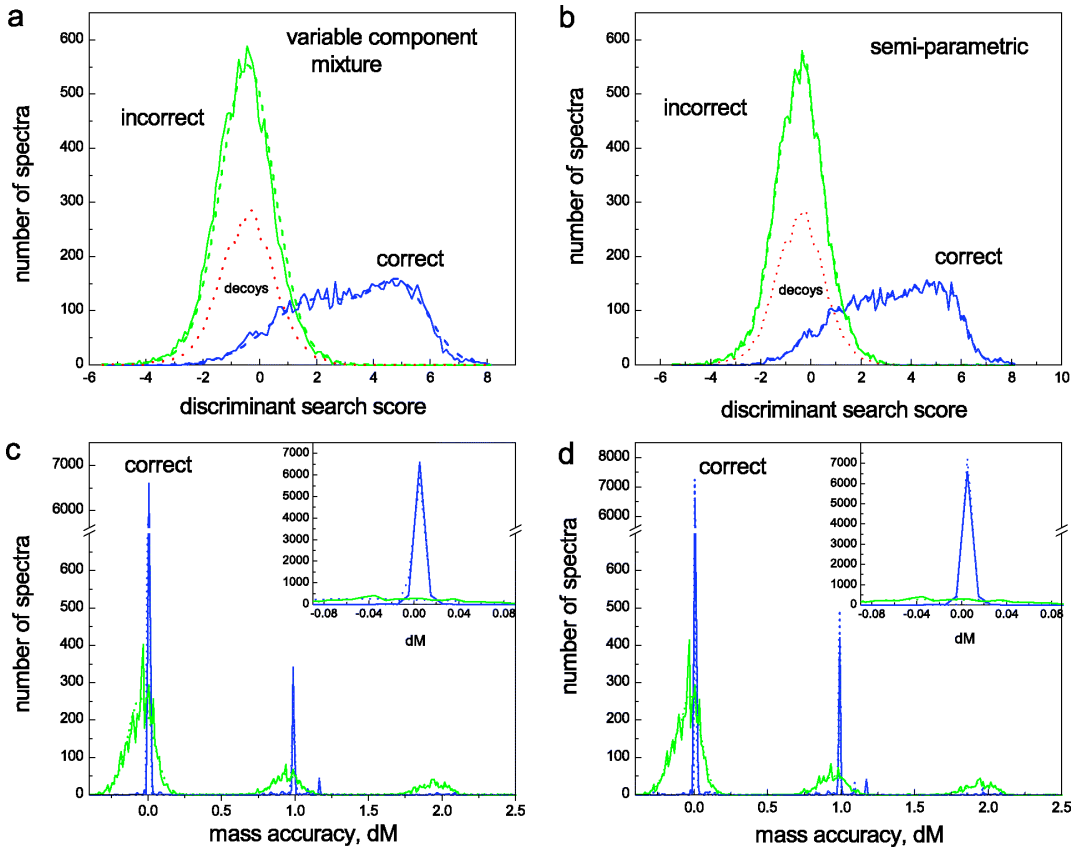
Familiarization with the parameters for modelling is always beneficial.

Error rate estimation of PSMs

Choi H, Ghosh D, Nesvizhskii AI. Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. J Proteome Res. 2008 Jan;7(1):286-92. Epub 2007 Dec 14. PubMed PMID: 18078310.

Decoy databases

- **"Synthetic" sequences not native to the proteome.**
- **Typically "reversed" and added to the search database**
- Lower quality spectra likely to randomly match to "false" sequence.
- Allows an estimate of error rate of PSM



FDR:

		The MS/MS spectrum comes from a peptide sequence in the database	
		True	False
Search reports a match to the correct sequence	True	True positive	False positive
	False	False negative	True negative

www.matrixscience.com

Or model the distributions:

$$P(+|S_i, E_i) = \int_\Theta P(+|\Theta, S_i, E_i)\, dF(\Theta|S_i, E_i)$$
$$= \sum_d p(d) \int_{\Theta^{(d)}} P(+|\Theta^{(d)}, S_i, E_i)\, dF(\Theta^{(d)}, S_i, E_i)$$
$$\approx \frac{1}{I} \sum_{k=1}^{I} 1(+|\Theta_k, S_i, E_i)$$

Error rate estimation of PSMs

1: Käll L, Storey JD, MacCoss MJ, Noble WS. Posterior error probabilities and false discovery rates: two sides of the same coin. J Proteome Res. 2008 Jan;7(1):40-4. Epub 2007 Dec 4. Review. PubMed PMID: 18052118.

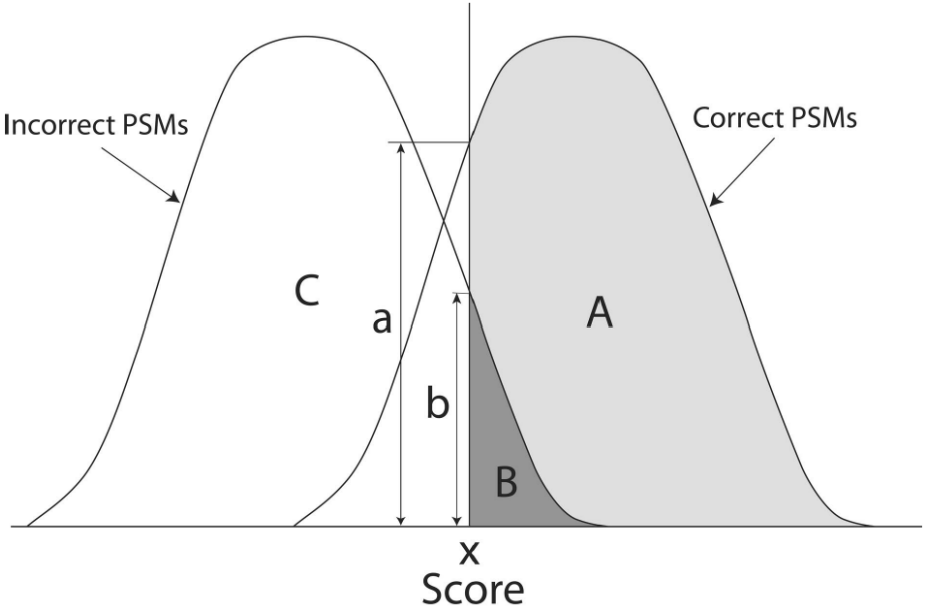
Decoy databases

- Lower quality spectra likely to randomly match to “false” sequence.
- Allows an estimate of error rate of PSM

FDR:

		The MS/MS spectrum comes from a peptide sequence in the database	
		True	False
Search reports a match to the correct sequence	True	True positive	False positive
	False	False negative	True negative

www.matrixscience.com



FDR = B/(A + B)

PEP = b/(a + b)

Figure 1. Two complementary methods for assessing statistical signiffcance.

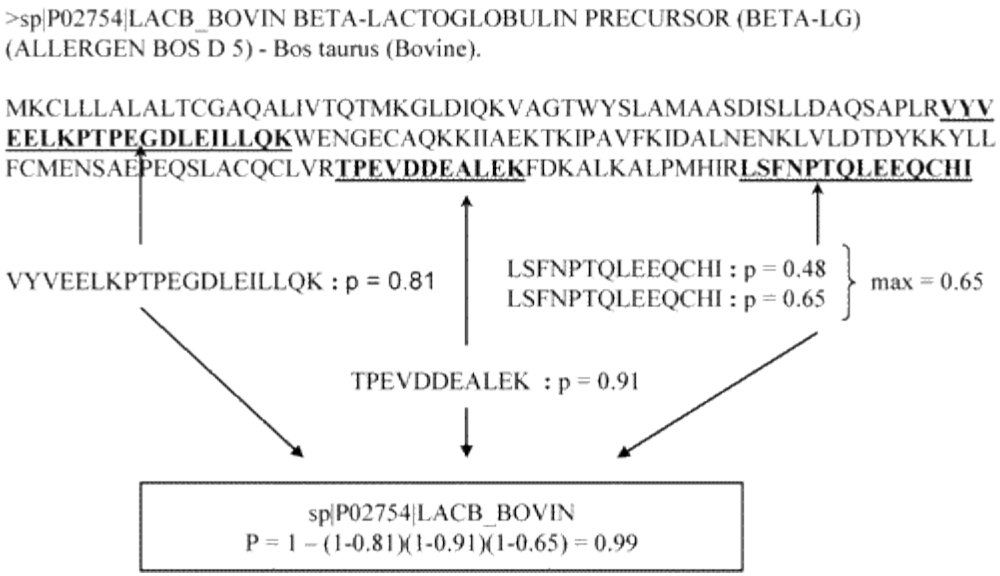
Problems with PSMs for protein inference

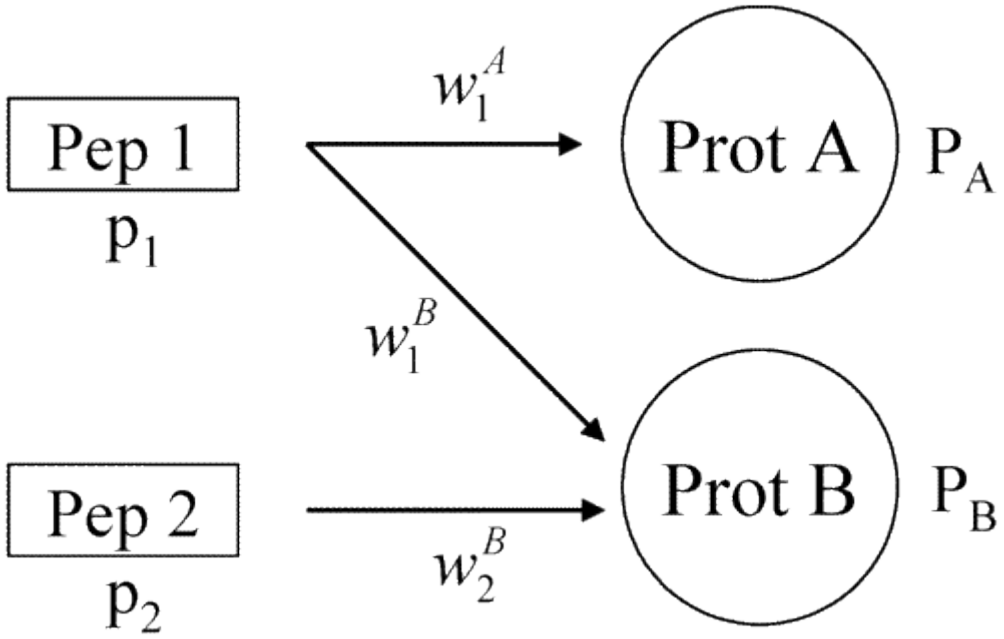
- Sequences are redundant
- Isoforms and peptide protein mapping
- Replicate PSMs acquired (multiple samples).....which peptide probability is best?
- What's an appropriate threshold?

- ANSWER→
  - NSP…Number sibling peptides model..is the sequence unique?
  - Best observable peptide
  - Protein probabilities
  - "decoy" databases

1: Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem. 2003 Sep 1;75(17):4646-58. PubMed PMID: 14632076.

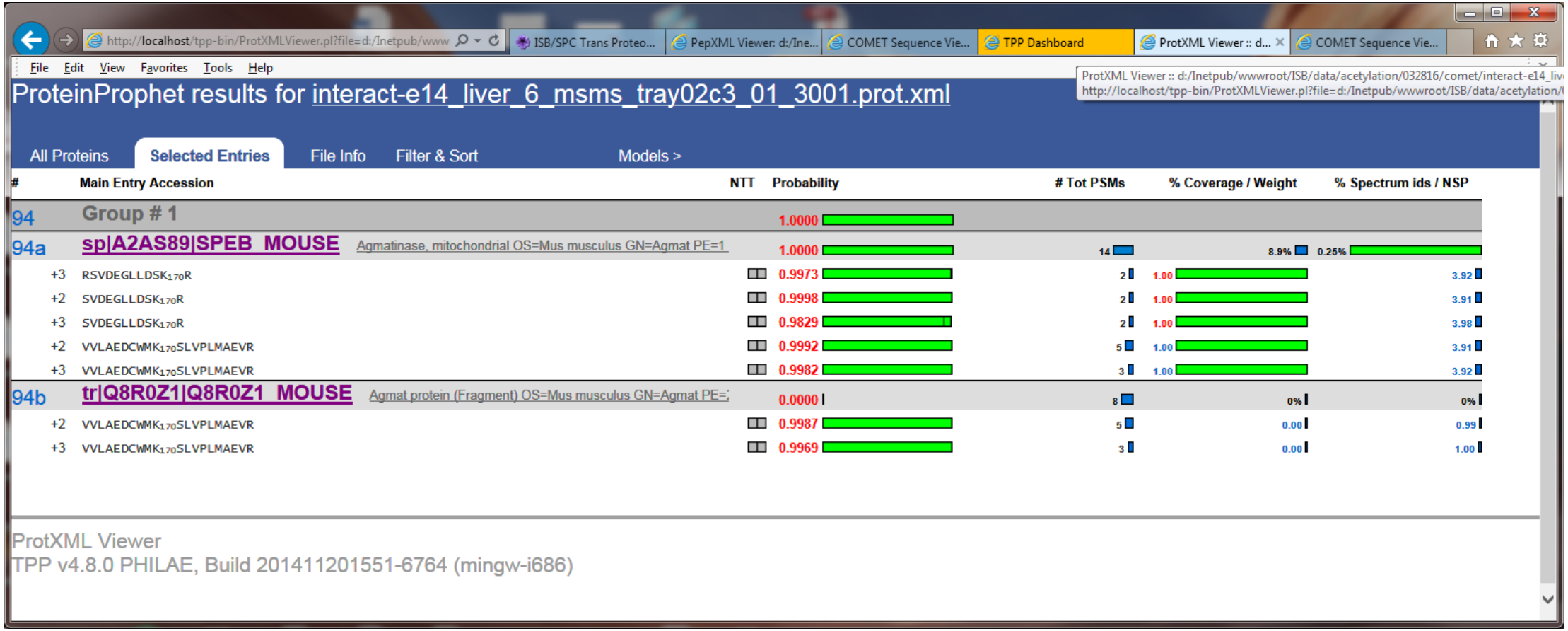
Statistical Model for Protein Inference





$$p(+|D,\mathrm{NSP}) = \frac{p(+|D)p(\mathrm{NSP}|+)}{p(+|D)p(\mathrm{NSP}|+) + p(-|D)p(\mathrm{NSP}|-)} \quad (5)$$

Protein Identification from Peptides
Shared peptides



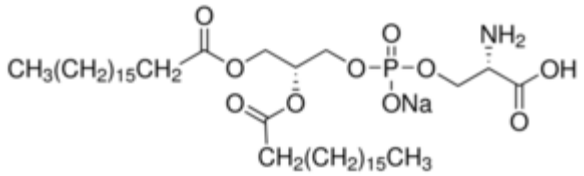
Small Molecule MS/MS

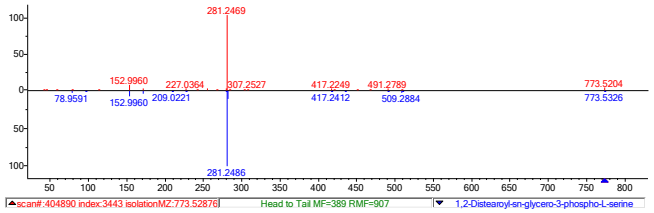
- Unknown predictable patterns for electrospray ionization- MS/MS (ESI-MS/MS)

- Empirical knowledge versus predictive methods
  - Empirical patterns, 1970's, 1990's
    - More accurate
    - Dependent on identifying and storing validated spectra
  - Predictive: 2010's

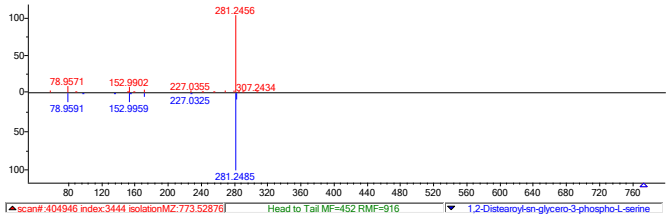
MS^2 at work:

PS(18:0/18:0)
1,2-Distearoyl-sn-glycero-3-phospho-L-serine
$C_{42}H_{82}NO_{10}P$

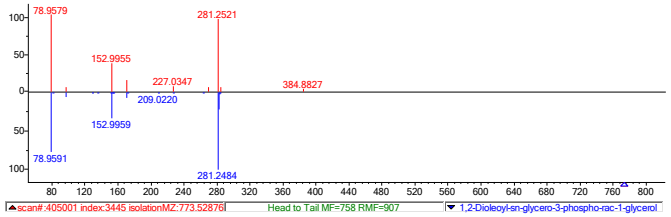




40V [M-H-]- : 773.52876



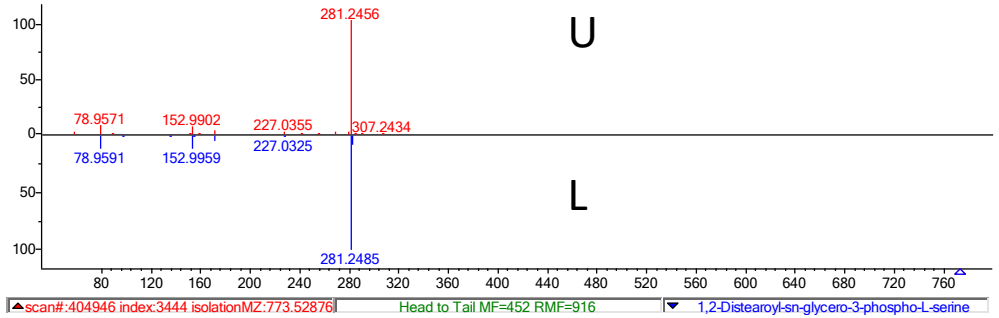
60V [M-H-NH3]- : 791.567635



80V [M-H-]- : 773.52876

Early work

Optimization and testing of mass spectral library search algorithms for compound identification, Stein and Scott (1994)



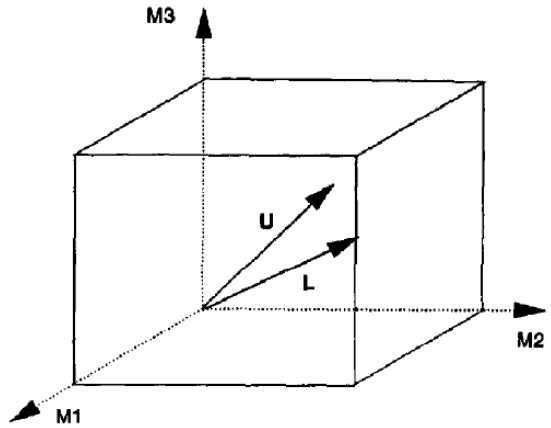


Figure 1. Vector representation of a hypothetical three-peak unknown (U) and library (L) mass spectrum in three-dimensional space (peaks have mass M1, M2, and M3).

Table 1. *Search algorithms investigated*[a]

Euclidean distance

$$\left(1 + \frac{\sum (W_L - W_U)^2}{\sum W_U^2}\right)^{-1}$$

Absolute Value Distance

$$\left(1 + \frac{\sum |W_L - W_U|}{\sum W_U}\right)^{-1}$$

Hertz et al. [9]

$$\frac{\text{average of weighted peak intensity ratios}}{[1 + \text{fraction of unmatched intensities}]}$$

Dot-product (cosine), F_d

$$\frac{\left(\sum W_L W_U\right)^2}{\sum W_L^2 \sum W_U^2}$$

Probability-based matching (PBM) [5b, 5d, 10c]
Uses probability that, by chance, peaks match within an abundance window (W value) by using uniqueness values for mass (U value) and abundance (A) along with a variety of rules and correlation tables.

Composite:

$$\frac{N_U F_D + N_{L\&U} F_R}{N_U + N_{L\&U}}$$

F_D = Dot-Product Term Above
F_R = Ratio of Peak Pairs (below)
W = [Peak Intensity]^n [Mass]^m = Weighted Intensity
N = Number of peaks

$$F_R = \frac{1}{N_{L\&U}} \sum_{i}^{L\&U} \left(\frac{W_{L,i}}{W_{L,i-1}} \frac{W_{U,i-1}}{W_{U,i}}\right)^n$$

where $n = 1$ or -1 when the term in parentheses is less than or greater than unity, respectively

[a] Subscripts: L = library, U = unknown, L & U = peak in both library and unknown spectrum.

Early work

Optimization and testing of mass spectral library search algorithms for compound identification, Stein and Scott (1994)

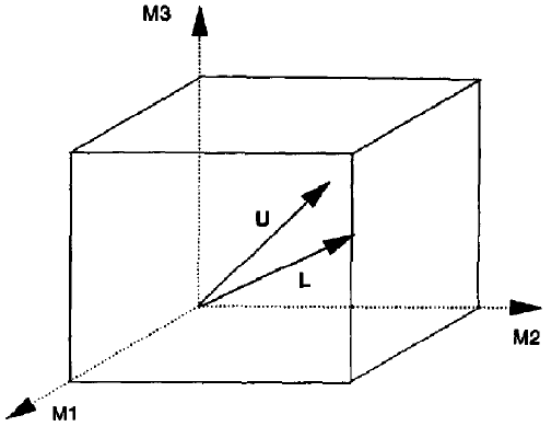


Figure 1. Vector representation of a hypothetical three-peak unknown (U) and library (L) mass spectrum in three-dimensional space (peaks have mass M1, M2, and M3).

Table 2. Performance of various algorithms

Algorithm	% Correct at Rank			Scaling /Comments	
	1	1–2	1–3	Mass Power	Intensity Power
Dot-product	72.9	85.9	90.8	1	0.5[a]
	73.2	86.3	91.0	1	0.5[b]
	72.8	85.9	90.8	1	0.5
Optimized	74.9	86.9	91.7	3	0.6
Euclidean distance	65.8	79.3	84.9	0	0.5
	69.9	82.9	88.2	1	0.5
Optimized	71.9	83.9	88.9	2	0.6
Absolute distance	61.4	74.9	81.2	0	1
	66.8	79.4	85.1	1	1
Optimized	67.9	80.3	85.5	2	0.9
PBM	57.1	71.5	78.5	"*k* value"[c]	
	64.0	77.7	84.3	Reliability[c]	
	64.7	78.4	84.8	Complete[c]	
Hertz et al.	59.9	73.9	81.1	0	See ref 9
Optimized	64.4	77.2	83.2	2	0.5
Composite	75.7	88.0	92.5	3, 0[f]	0.5, 1[f]

[a] Used local and global "normalization" [6].
[b] Used local "normalization" [6].
[c] Based on "reverse-search" overall spectral match factor [10].
[d] All recommended features except "quadratic scaling" [5b, d].
[e] All recommended features [5b, d].
[f] The first value is for the dot-product term and the second value is for the peak ratios term. The second and third power of the mass were equally effective for the first term.

Small Molecule MS/MS

- Still predominantly use DOT product for library- unknown matching.
  - Comprehensive limited by lack of library entries….eg not enough spectra.

- New predictive methods attempt to produce theoretical spectra from known structures
  - Then attempt to match experimental spectra to these theoretical entries.

Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification, Allen et al (2015)

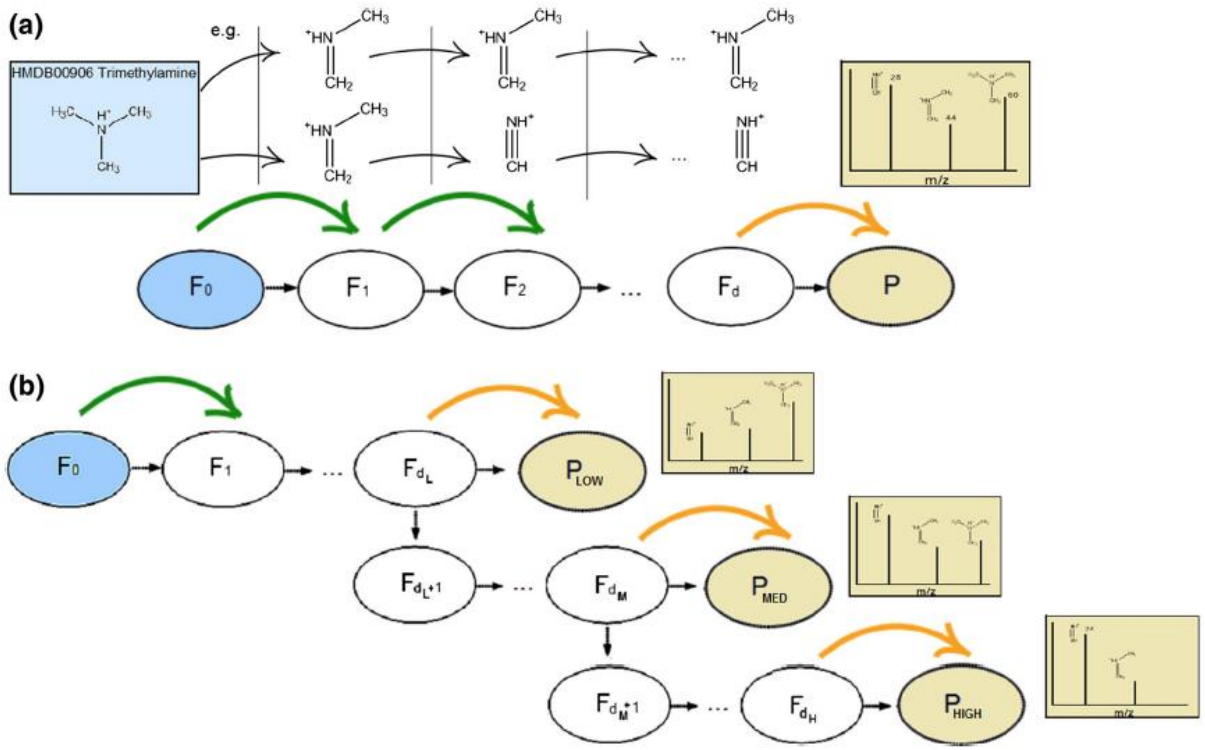


Fig. 1 **a** Single energy competitive fragmentation model (SE-CFM): a stochastic, Markov process of state transitions between charged fragments. **b** Combined energy competitive fragmentation model (CE-CFM): an extension of SE-CFM that combines information from multiple collision energy spectra into one model

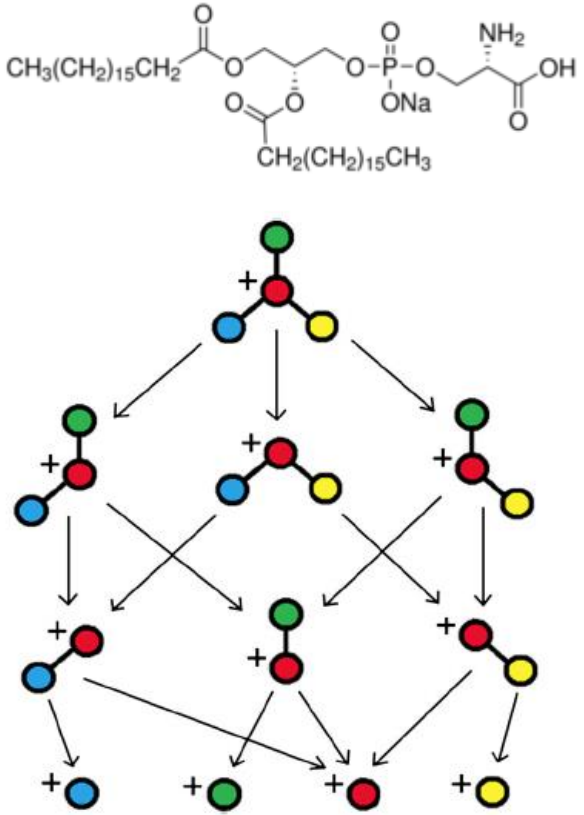
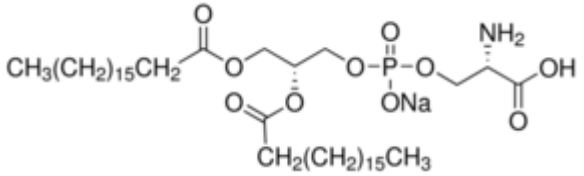


Fig. 2 An abstract example of a fragmentation graph, showing a directed acyclic graph of all possible ways in which a particular charged molecule may break to produce smaller charged fragments

Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification, Allen et al (2015)



- Very complex problem
  - Competing moieties for fragmentation
  - Fragmentation tendency
  - Sequential transition state models

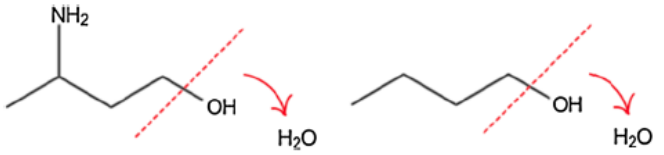


Fig. 3 Two similar breaks, both resulting in an H_2O neutral loss. The *right case* should be assigned a higher probability, as in the *left case*, the NH_3 is also likely to break away, reducing the probability of the H_2O loss

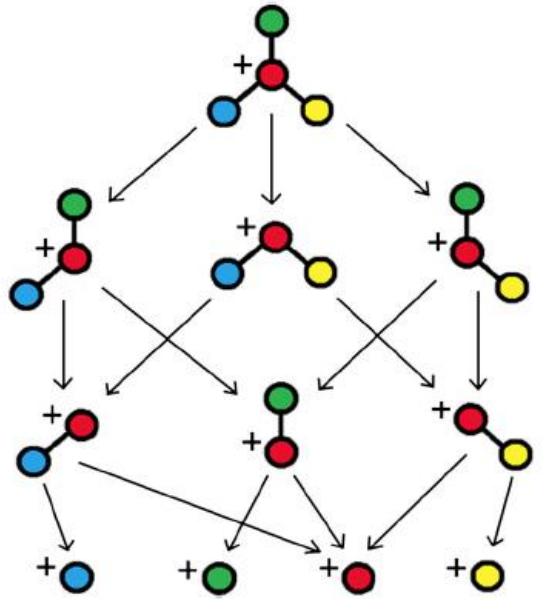
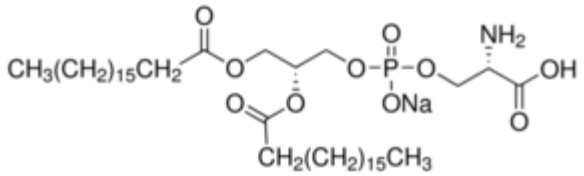
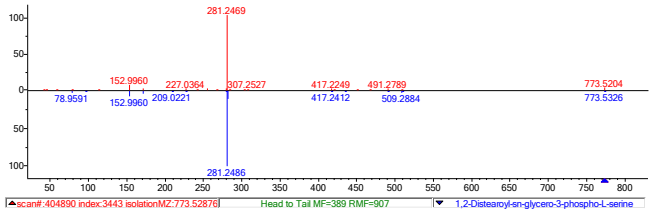


Fig. 2 An abstract example of a fragmentation graph, showing a directed acyclic graph of all possible ways in which a particular charged molecule may break to produce smaller charged fragments

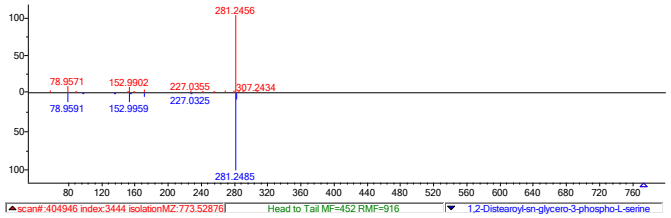
MS^2 at work:

PS(18:0/18:0)
1,2-Distearoyl-sn-glycero-3-phospho-L-serine
$C_{42}H_{82}NO_{10}P$

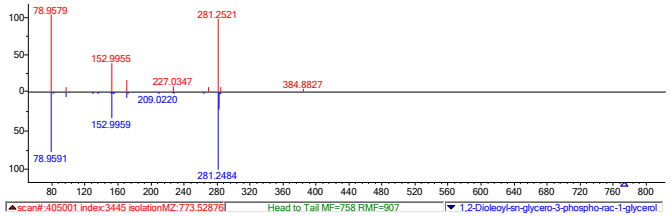




40V [M-H-]- : 773.52876



60V [M-H-NH3]- : 791.567635



80V [M-H-]- : 773.52876

Small Molecule MS/MS

- Old reliable methods remain in use
  - Labor intensive to produce search libraries
  - Not comprehensive enough
  - Not platform / laboratory specific

- New methods promising
  - Comprehensive / high specificity theoretical databases
  - Not truly benchmarked