

# **A Statistical Approach to Quantifying Uncertainties in the Accurate Mass and Time (AMT) tag Pipeline**

**Navdeep Jaitly, Joshua Adkins, Matthew E. Monroe,  
Angela D. Norbeck, Heather M. Mottaz, Alan R.  
Dabney, Mary S Lipton, Gordon A. Anderson,  
Richard D. Smith**

# Outline

---

- Analytical Pipelines using Liquid Chromatography coupled to High Resolution Mass spectrometry Experiments (the Accurate Mass and Time Tag approach)
- Need for a statistical foundation to characterize identifications obtained under such a paradigm
- Summarize a statistical method (Peptide Prophet) which tackles this problem for MS/MS based analysis pipelines
- Our approach to characterizing confidence in identifications from AMT tag pipeline
- Some results

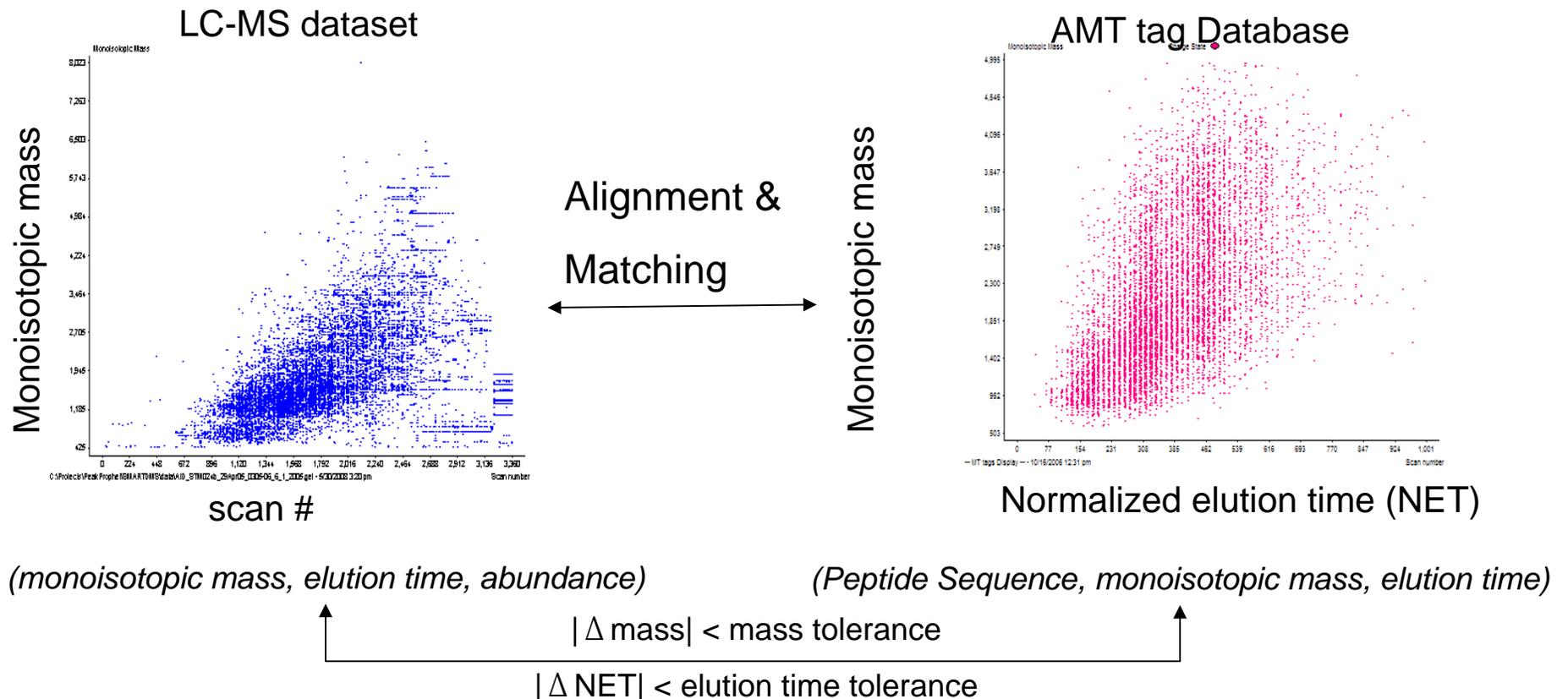
# LC-MS based Proteomics Analysis Pipelines

---

- Availability of High mass accuracy instruments has led to novel identification methods using LC-MS experiments
  - AMT, PePPER, MSInspect, SuperHirn, CRAWDAD
  - Liquid Chromatography retention time used in identification
- Generally higher sensitivity than MS/MS analyses; better for broad comparative proteomics studies
- Lower specificity – matching is performed using mass and LC elution time alone, not multidimensional fragmentation patterns.
- *An approach is required to estimate probability that individual identifications are correct to help separate good and bad results at user-specified false discovery rates*

# Accurate Mass and Time (AMT) tag pipeline

- Features in an LC-MS dataset are matched to a database of peptides previously identified by LC-MS/MS analyses using specified mass and elution time tolerances



# Current Method for Assessing and Controlling Rate of Random Matches

---

- Decoy database matching used to assess rate of random matching
  - Shift database (or features) by some mass, and re-match to database – the numbers of matches reflects the random rate of errors
- Rate of error controlled by
  - Reducing mass and elution time tolerances
  - Building more stringent LC-MS/MS databases (higher XCorr, hyperscore, Mascot score, Peptide Prophet score, etc)
- Overall method consists of iteratively controlling and assessing error to choose “optimal” parameters

# Challenges with current method

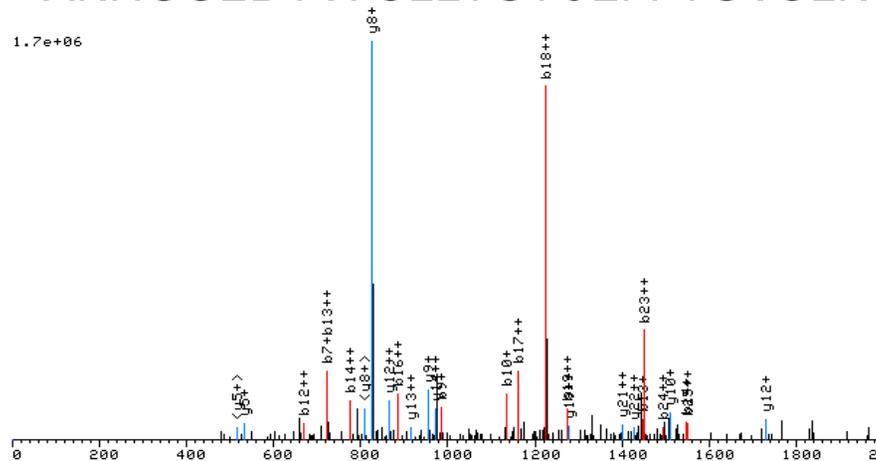
---

- Balancing false positives and false negatives is a tricky game!
  - Building more confident LC-MS/MS database decreases background false positives, but, increases false negatives
  - Reducing Mass and Elution time tolerances has a similar effect
  - Manually chosen parameters may look suitable, but in reality be sub-optimal
- Each identification is either accepted or rejected
  - In reality some identifications are better than others – higher MS/MS confidence, lower mass and elution time differences, etc

# Statistical Method for MS/MS Identifications

- Peptide Prophet – A Statistical Model to estimate probability that an MS/MS spectrum is correctly identified
  - Uses a Linear function (F-Score) of result metrics from SEQUEST to calculate an overall value representing the confidence of identifications

ARHGGEDYVFSLLTGYCEPPTGVSLR



$c_1$	$c_2$	$c_3$	$c_4$
XCorr	DelCN	SpRank	$d_M$
4.015	0.325	1	0.14

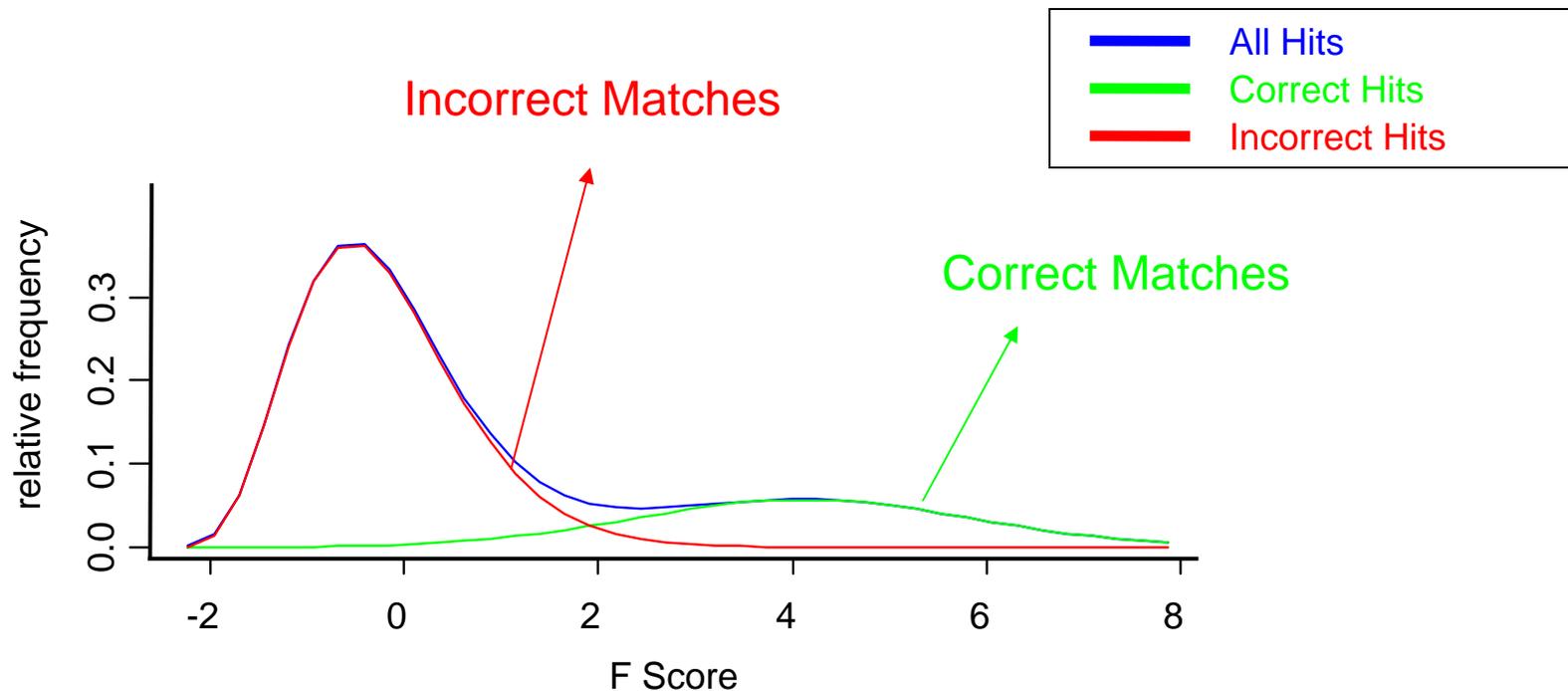
Parameters Used

$$F(x_1, x_2, \dots, x_s) = c_0 + \sum_{i=1}^s c_i x_i$$

Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R. *Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.* Anal Chem, 2002, 74, 20, pg. 5383-92

# Peptide Prophet F-Score Distributions

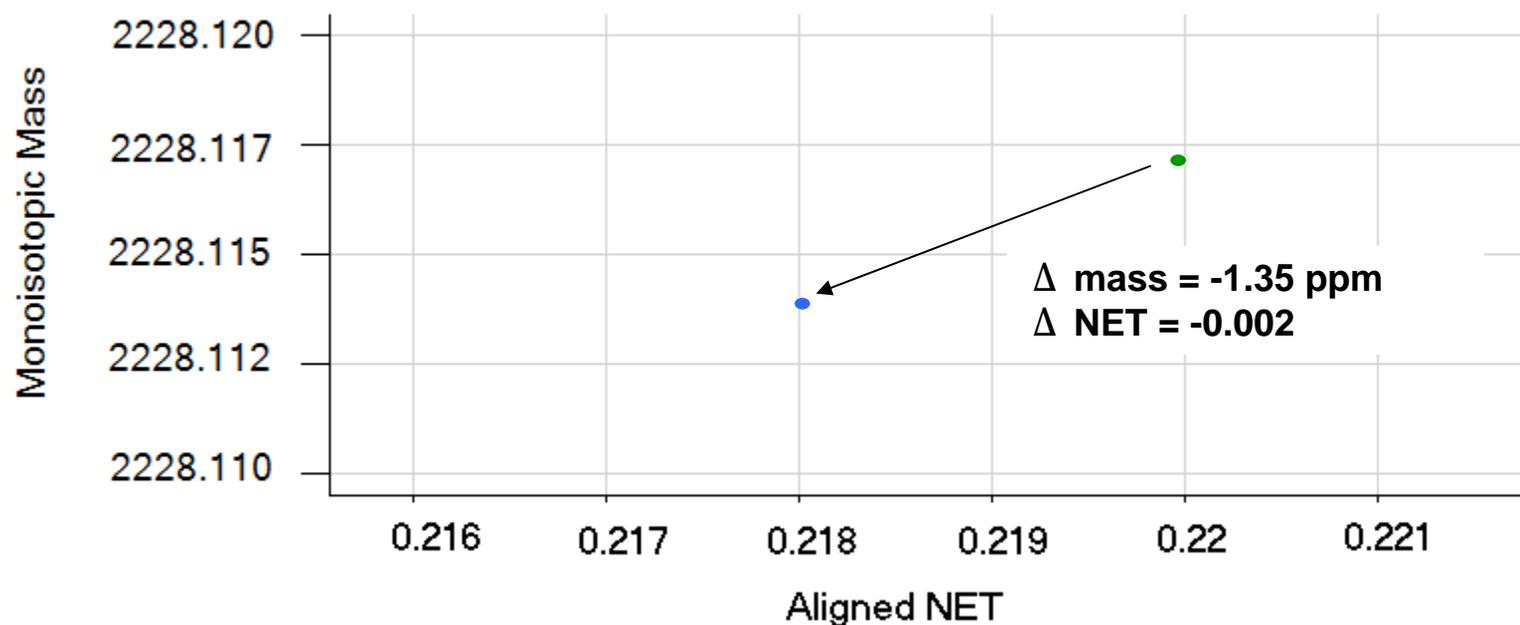
- Overall F-Score distribution is bimodal, with distinct distributions for correct and incorrect matches
- Probability that an identification is correct can be computed from relative probabilities of coming from correct or incorrect distribution



# Metrics Associated with a Candidate Identification from AMT Tag Pipeline

- Each match between an LC-MS feature and a peptide AMT tag is described by a mass error and an LC NET error, and metrics about the AMT Tag, related to the MS/MS searches from which it was identified

Mass	Scan	Aligned NET	Peptide	NET	Mass	Discriminant Score (Peptide Prophet)
2228.114	1097	0.218	TETQEKNPLPSKETIEQEK	0.220	2228.117	3.1



*Would like to use these metrics to separate results into correct and incorrect matches*

# Statistical *Method* for Assignment of *Relative Truth (SMART) Score*

---

- A *SMART* score combines the mass and LC NET error of a peak match with the probability the MS/MS identification\* (AMT tag) was correct, to estimate the probability that a feature was correctly assigned

$$p(+_{match} | \delta m, \delta net, Fscore, peptide) = \frac{p(\delta m, \delta net | +_{match}) p(Fscore | +_{peptide}) p(+_{peptide})}{p(\delta m, \delta net | +_{match}) p(Fscore | +_{peptide}) p(+_{peptide}) + p(\delta m, \delta net | random_{match}) p(Fscore | random_{match}) p(random_{match})}$$

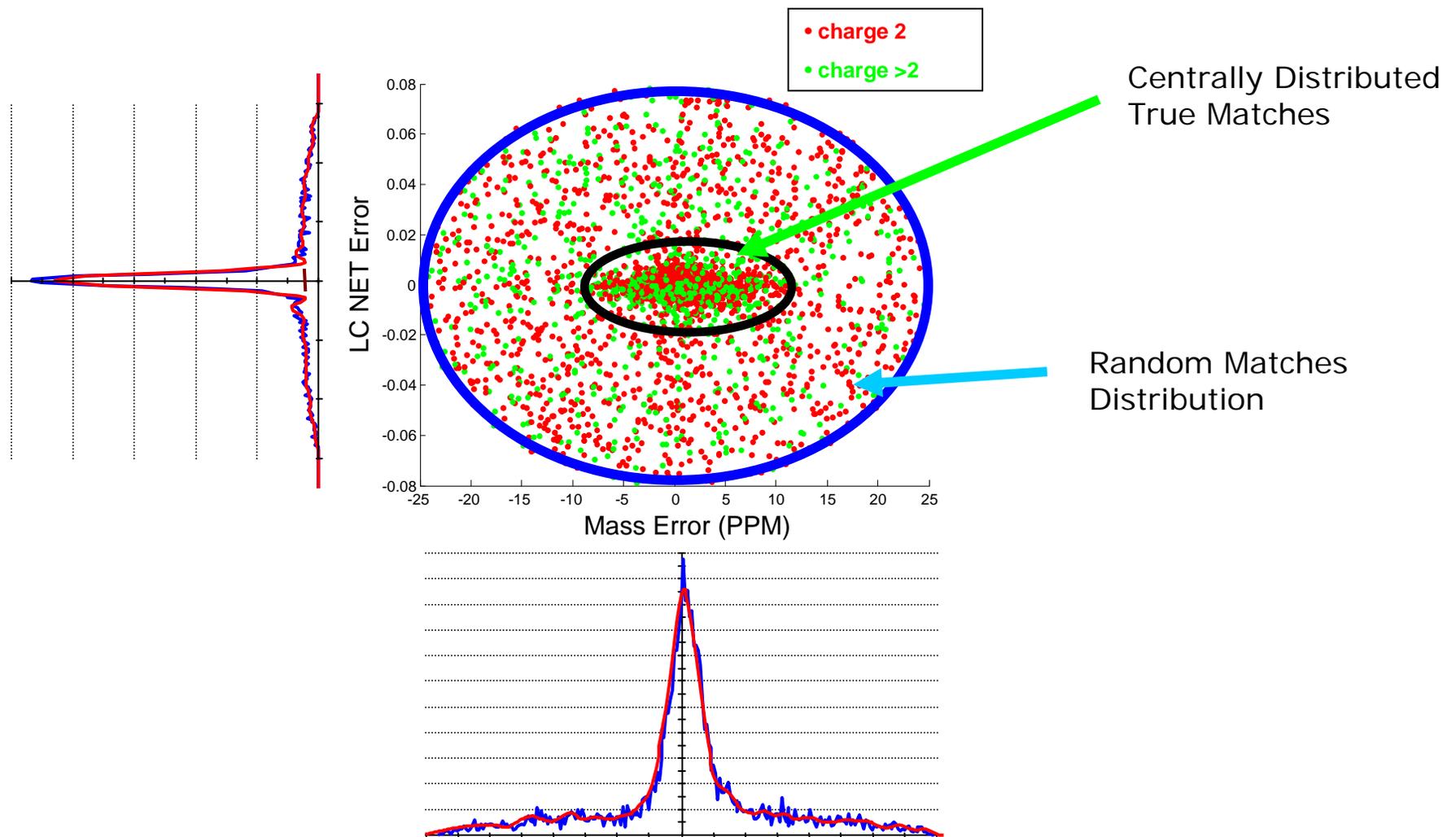
Assumes Independence of mass error, net error and Fscore

The SMART score is a q-value for the matching process.

\* Probability that a peptide in a database can be computed with existing methods such as Peptide Prophet based on properties of an MS/MS identification

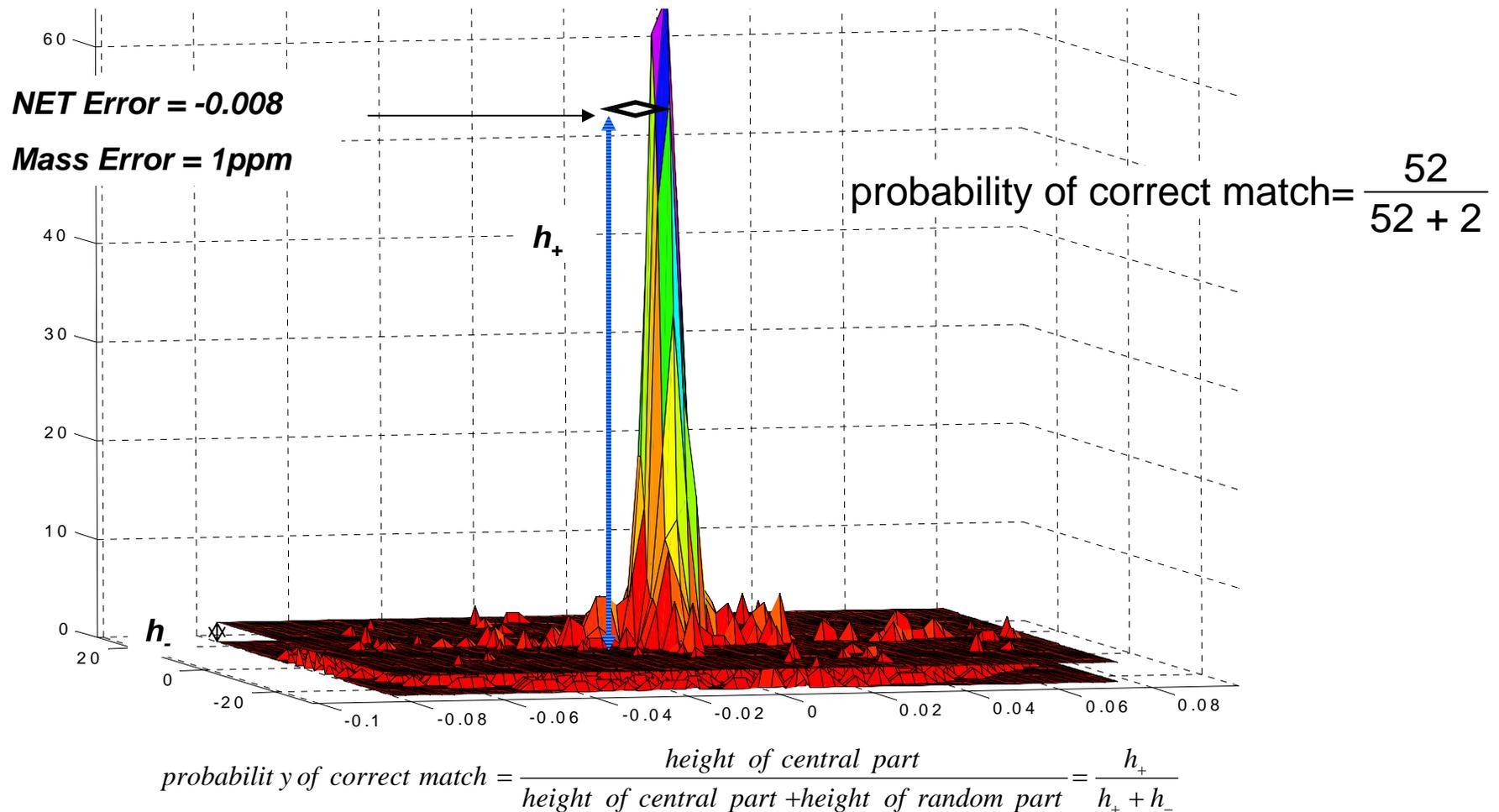
# Distribution of Mass and LC Normalized Elution Time (NET) differences

- True and False matches resulting from peak matching display different Mass and LC NET error distributions



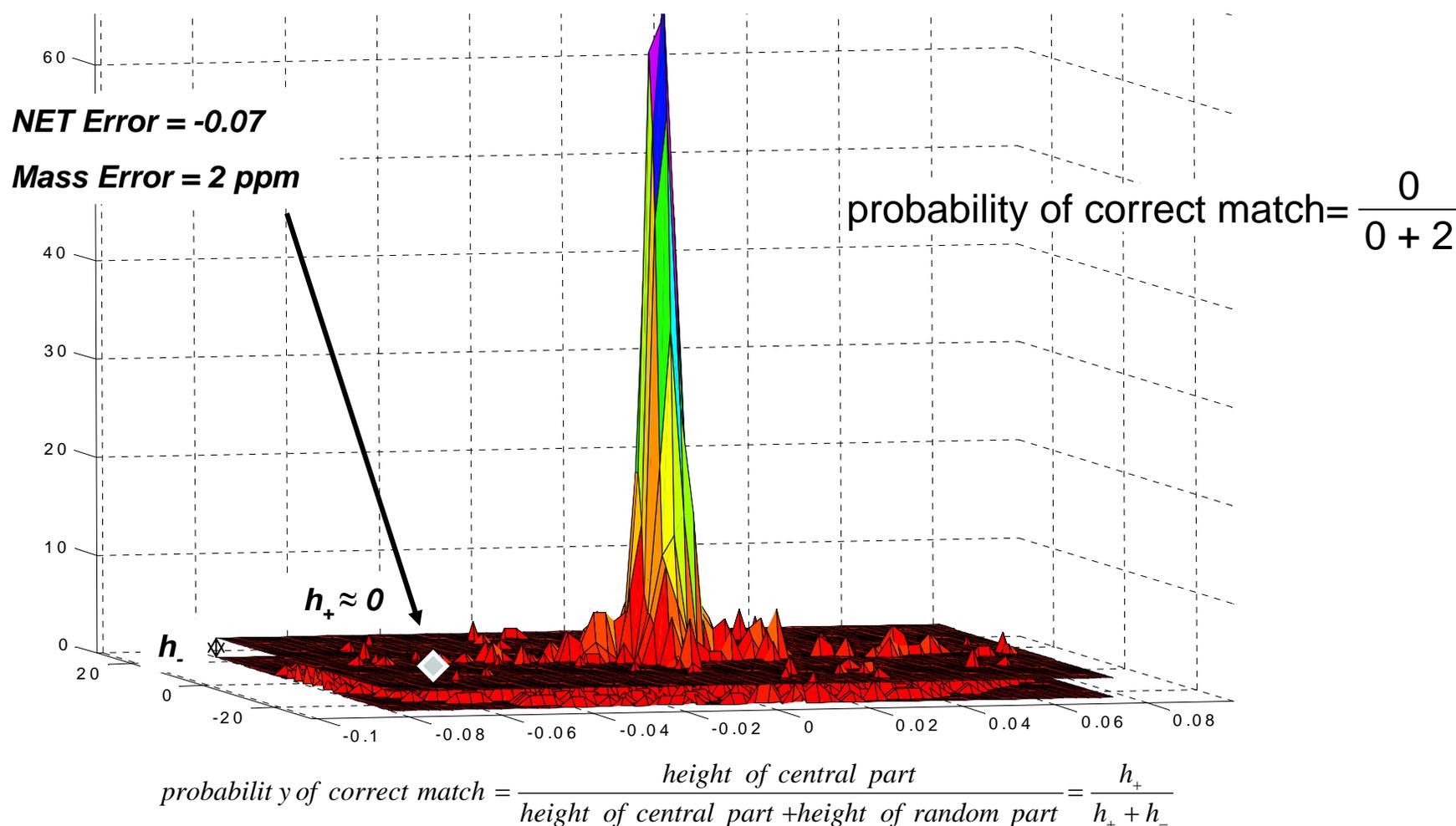
# Estimating the Probability a Match is Correct

- The probability that a peak match is correct depends on where its mass and LC NET error value lies on the two-dimensional distribution



# Estimating the Probability a Match is Correct

- The probability that a peak match is correct depends on where its mass and LC NET error value lies on the two-dimensional distribution



# Data Model and Model Fitting

- What distributions describe the observation vectors appropriately for positive and negative matches and can be used in the Bayes Formula ?

$$p(+_{match} | \delta m, \delta net, Fscore, peptide) =$$

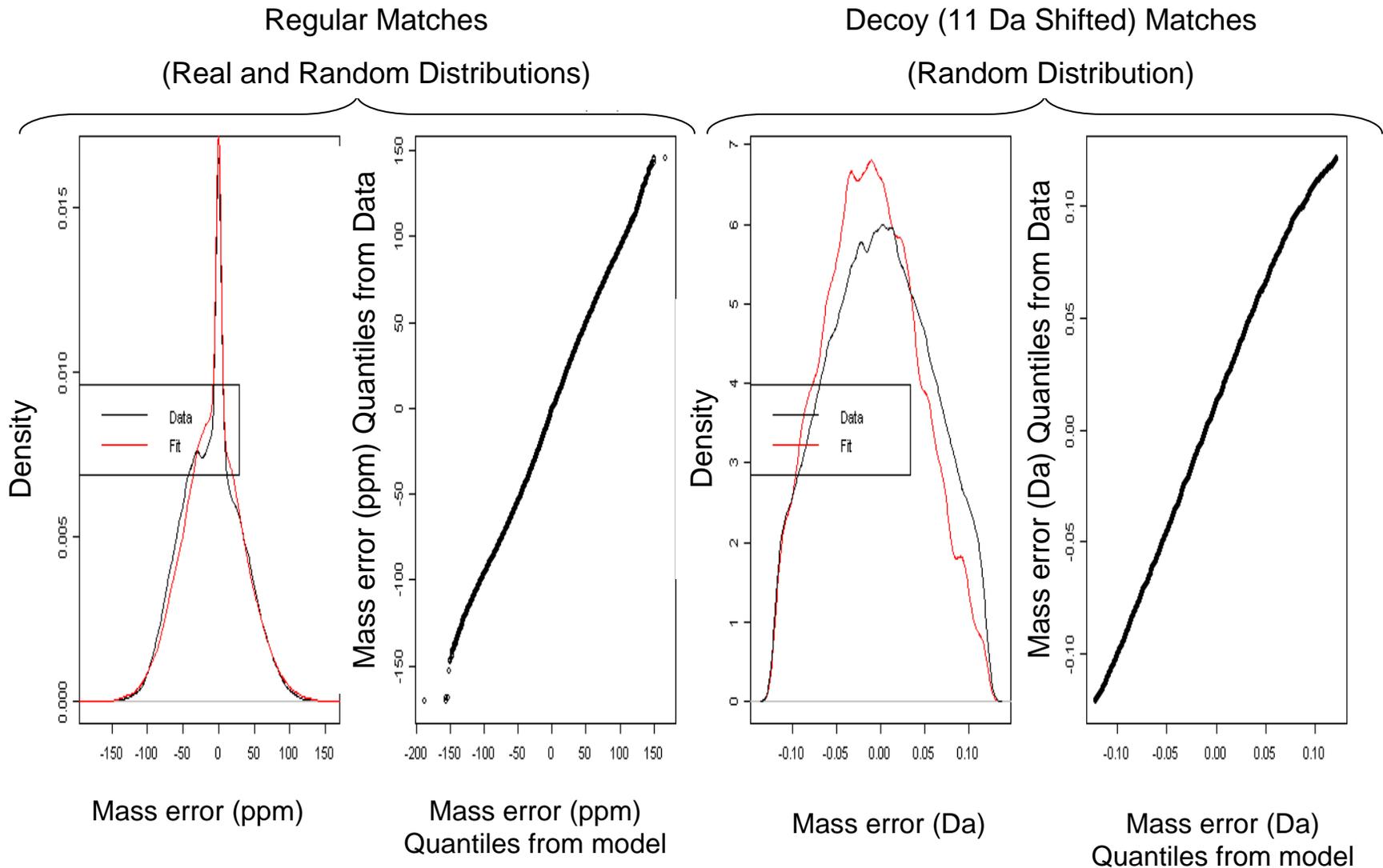
$$p(\delta m, \delta net | +_{match}) p(Fscore | +_{peptide}) p(+_{peptide})$$

$$p(\delta m, \delta net | +_{match}) p(Fscore | +_{peptide}) p(+_{peptide}) + p(\delta m, \delta net | random_{match}) p(Fscore | random_{match}) p(random_{match})$$

- Expectation Maximization Algorithm used to find optimal parameters for the distributions
- Data to Model: Mass Errors, NET Errors, F-Score distribution

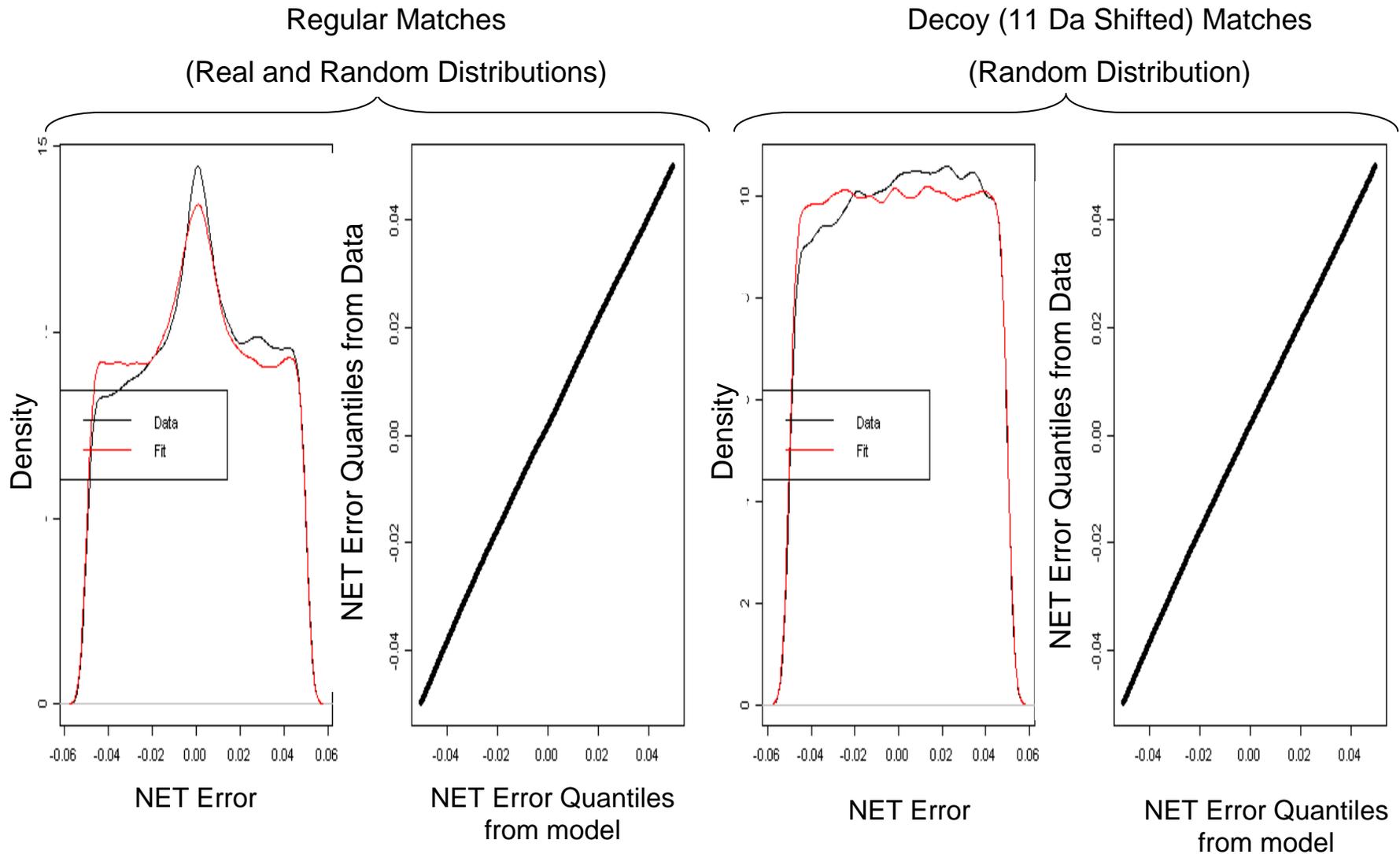
Real/Random	Type	Distribution
Real	Mass Errors	ppm errors distributed normally (truncated)
Random	Mass Errors	Da Errors distributed normally (truncated)
Real	NET Errors	Distributed normally
Random	NET Errors	Uniform Background
Real	F Scores	FScores Normally distributed as a function of mass
Random	F Scores	Gamma distribution

# Data Model - Example



Mass Error Distribution for *Salmonella Typhimurium* protein extract sample

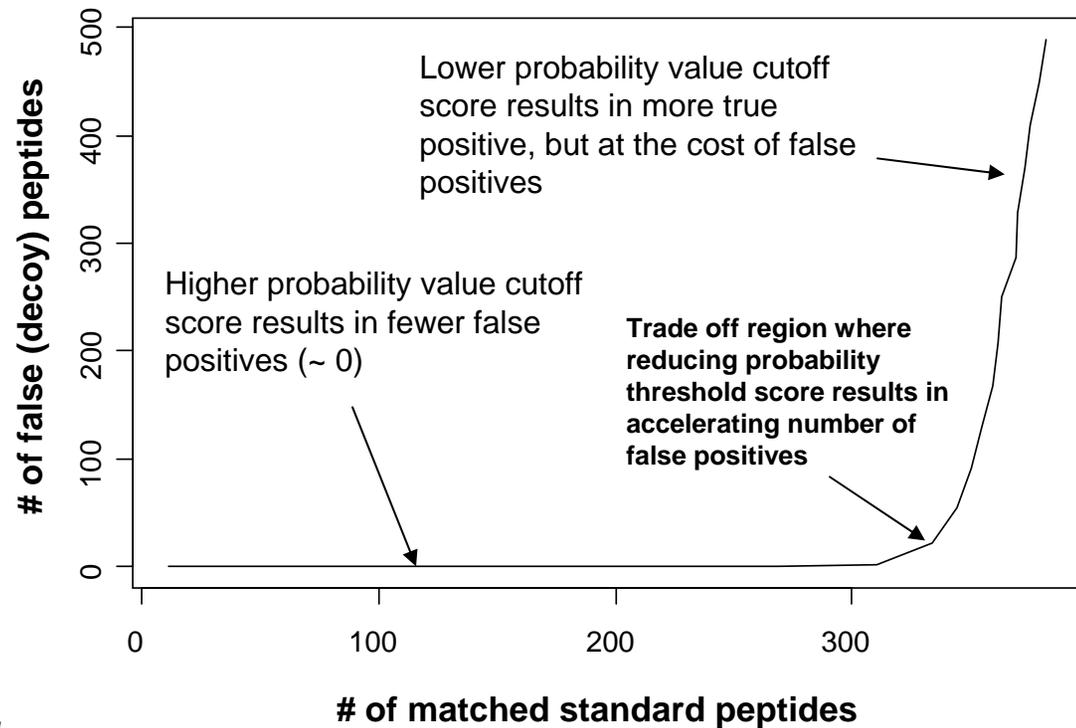
# Data Model - Example



NET Error Distribution for *Salmonella Typhimurium* protein extract sample

# Performance Curves

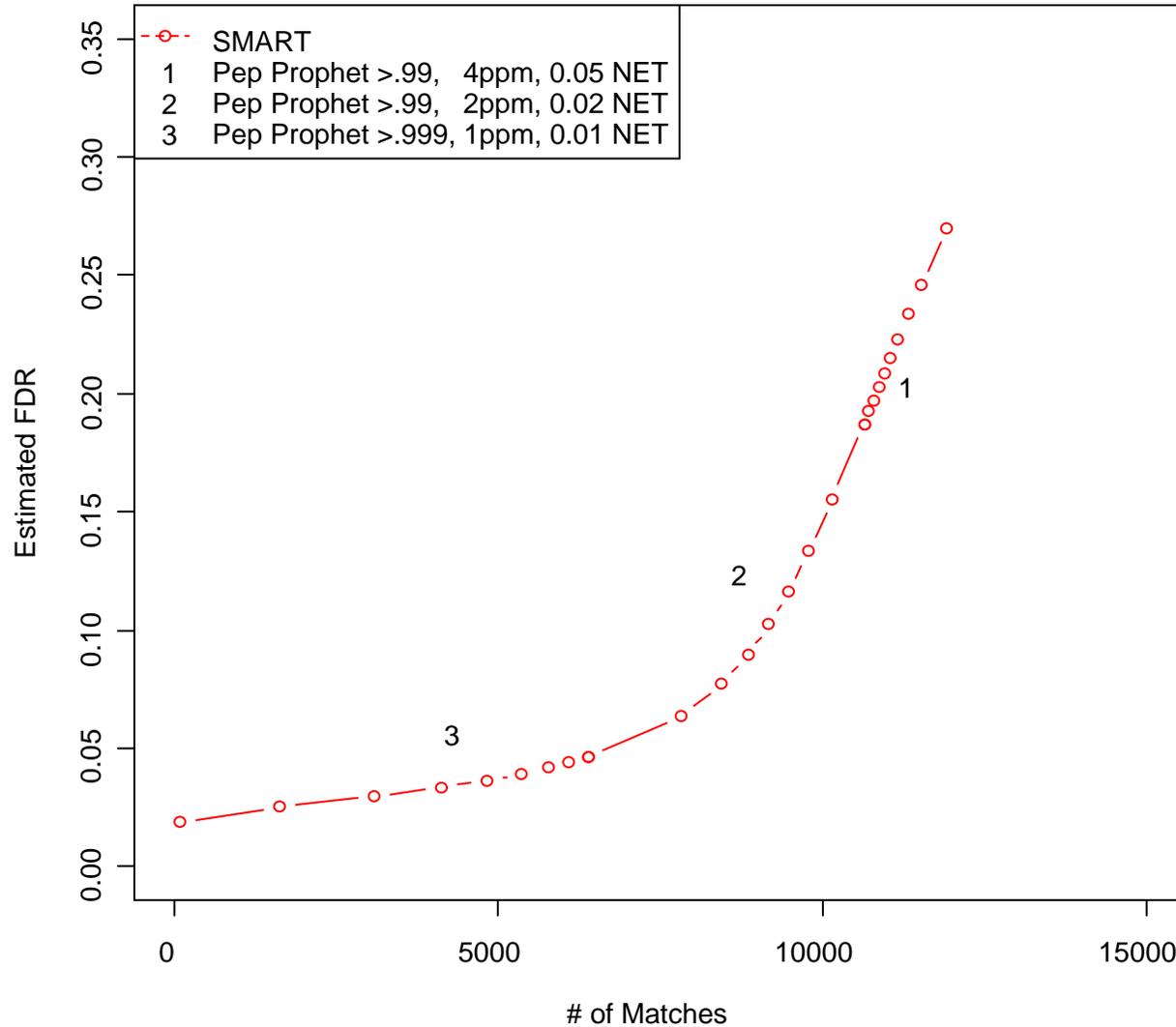
- Using different thresholds for SMART score, allows us to get results at different levels of error



***High number of true positives achieved with very few false positives!***

QC\_05\_3\_13Jan06\_Andromeda\_05-10-03\_1\_18\_2006  
Matched to database with 8000 decoy proteins.

# Performance Curves – Example 2



Performance Curve for *Salmonella Typhimurium* protein extract sample compared to typical criteria

# Summary

---

- Developed a model to estimate confidence of peak matching
- The *Statistical Method for Assignment of Relative Truth* (SMART) score provides a measure to prioritize acceptable matches using one number, by defining a probability score combining disparate information
- Allows calculation of FDR for identifications and estimates the tradeoff between false negatives and false positives
- Initial evaluation: shows good correlation with observed number of correct answers
- Making Implementation available as free software downloadable from <http://omics.pnl.gov/>

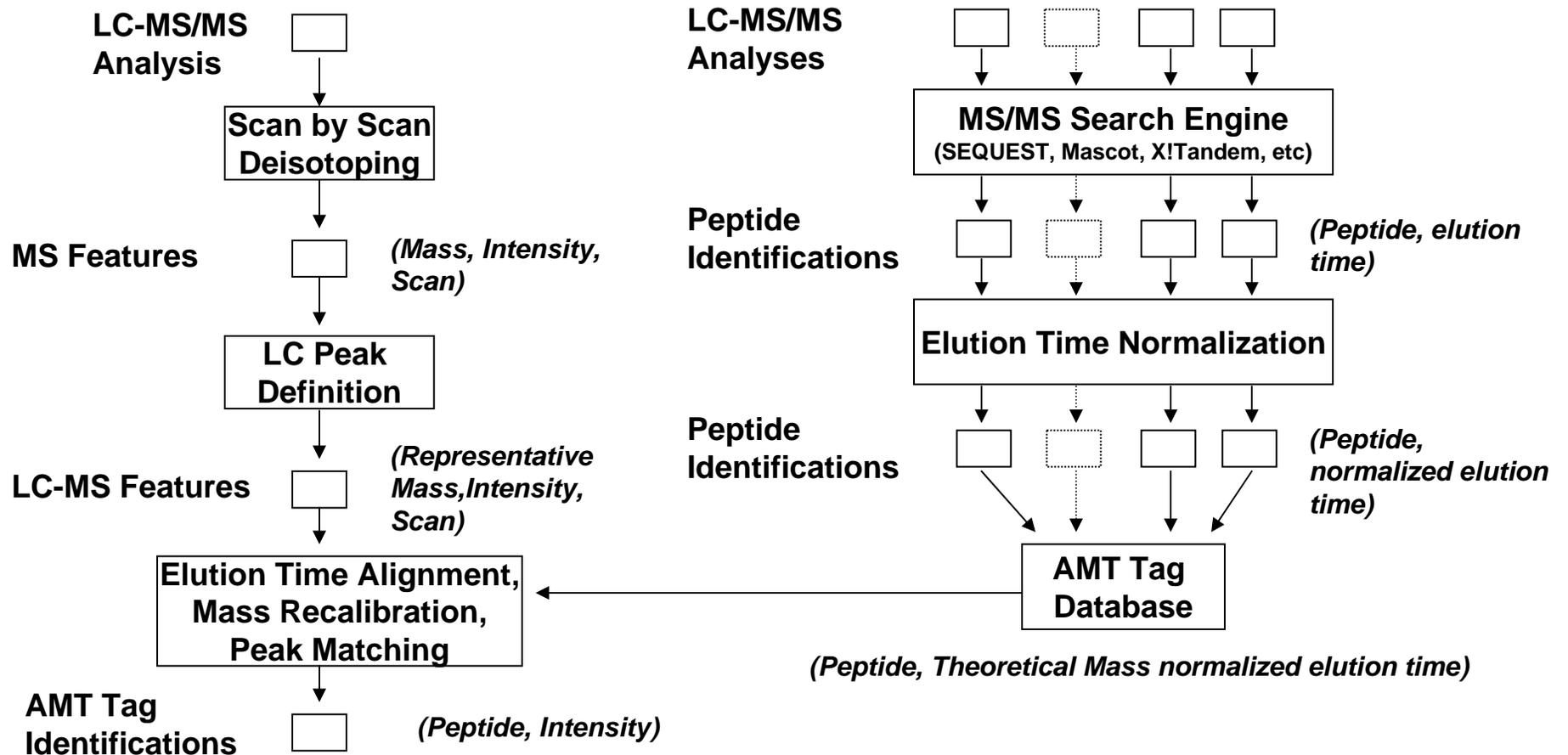
# Acknowledgements

---

- Co-authors
- Samuel Purvine (for acronym)
- Dick Smith
  
- Support
  - DOE Office of Biological and Environmental Research
  - NIH National Center for Research Resources
  - National Institute of Allergy and Infectious Diseases

# Accurate Mass and Time (AMT) Tag Pipeline

- Uses database of LC-MS/MS results to identify features discovered in LC-MS analyses<sup>1</sup>



1. <http://ncrr.pnl.gov/training/workshops/2007HUPO/LCMSBasedProteomicsDataProcessing.pdf>

# Decoy Vs Model

