NATIONAL RESEARCH UNIVERSITY

High-dimensional generative probabilistic models for peptide-spectrum matching in tandem mass spectrometry

Pavel Sulimov and Attila Kertész-Farkas

School of Data Analysis and Artificial Intelligence

Faculty of Computer Science,

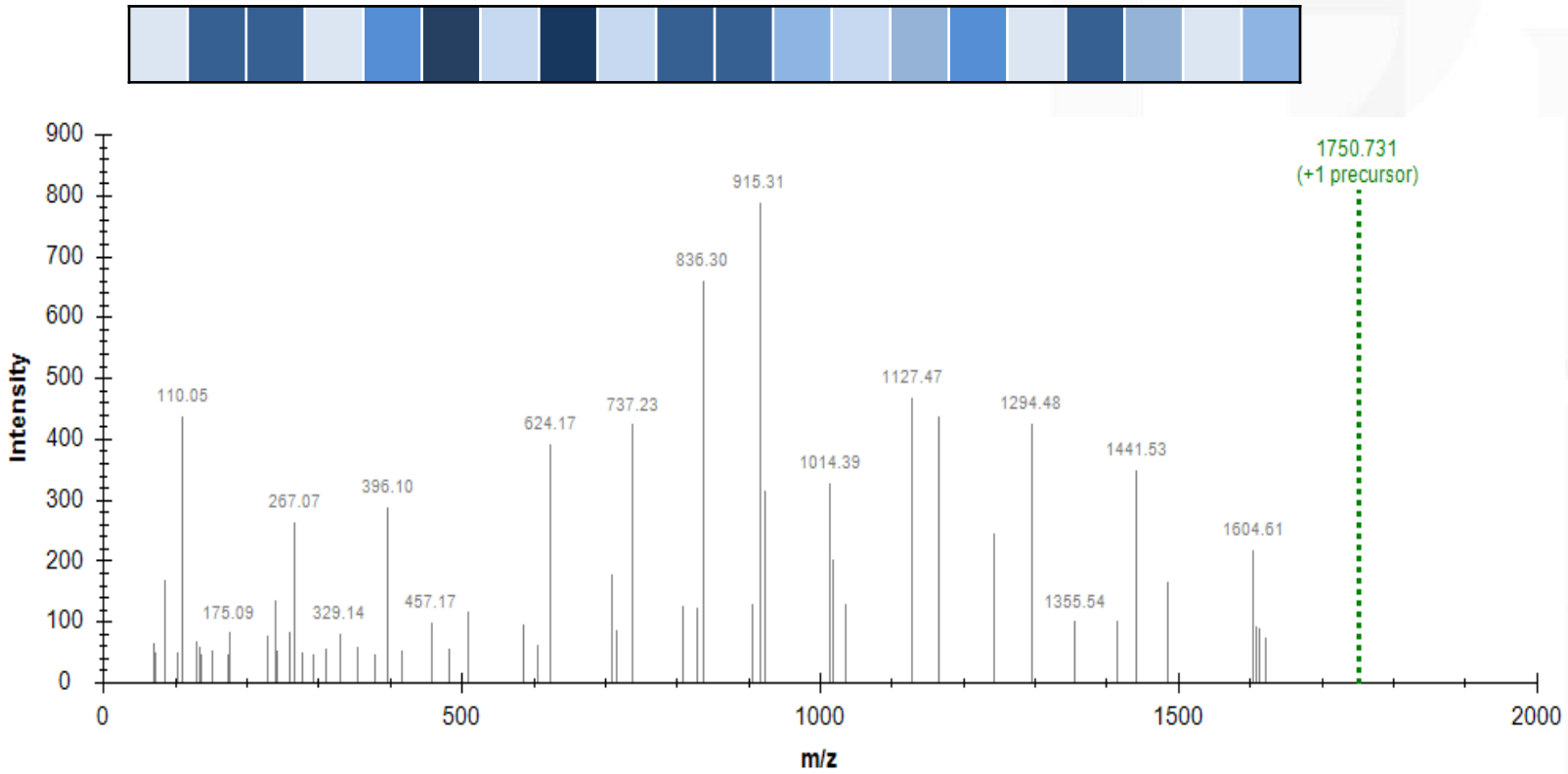
National Research University Higher School of Economics (HSE)

Room 308A, 3 Kochnovskiy Proezd, Moscow, 125319, Russia

Problem statement

Input: data, obtained from spectrum (below)

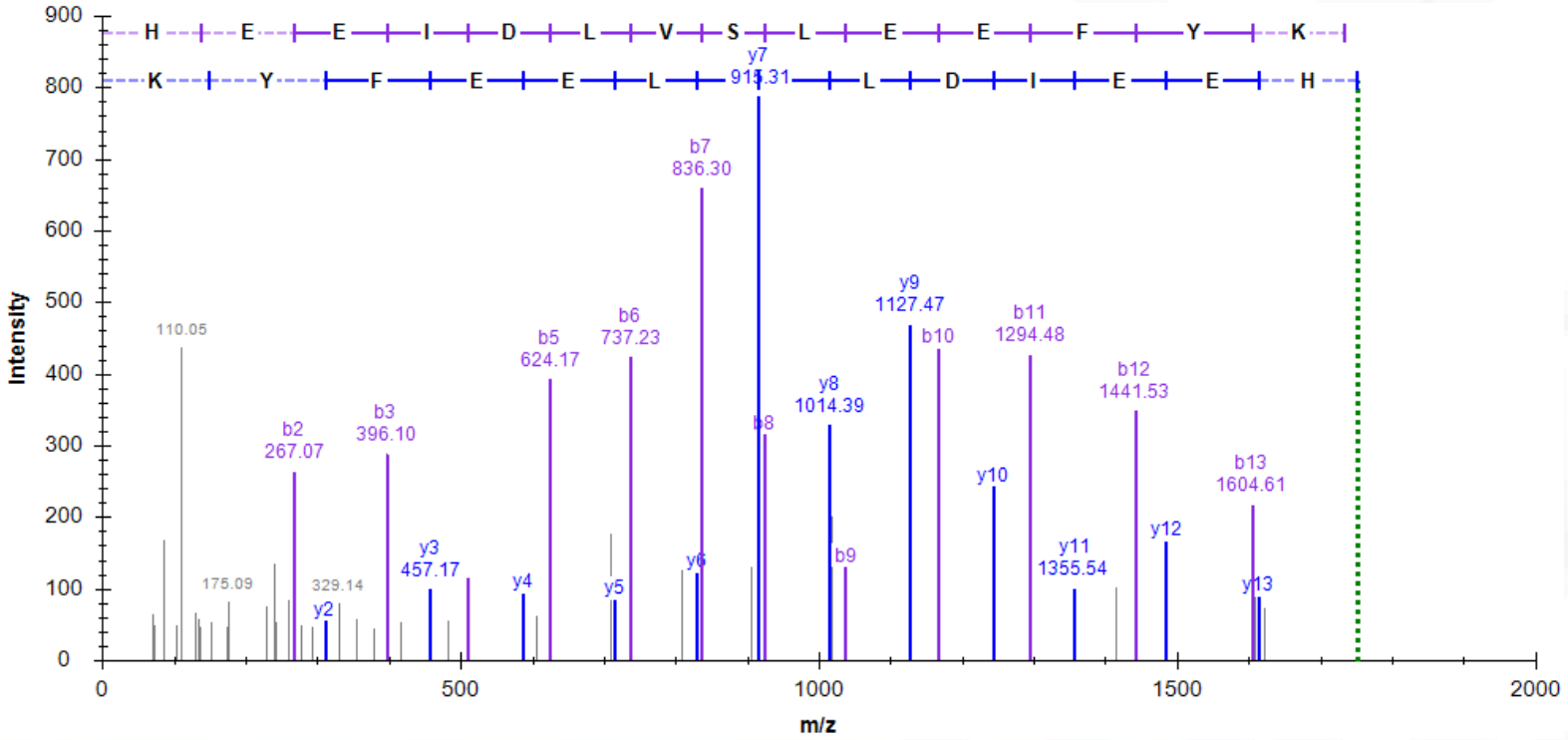
Target: identify the molecule which generated this observed spectrum



Problem statement

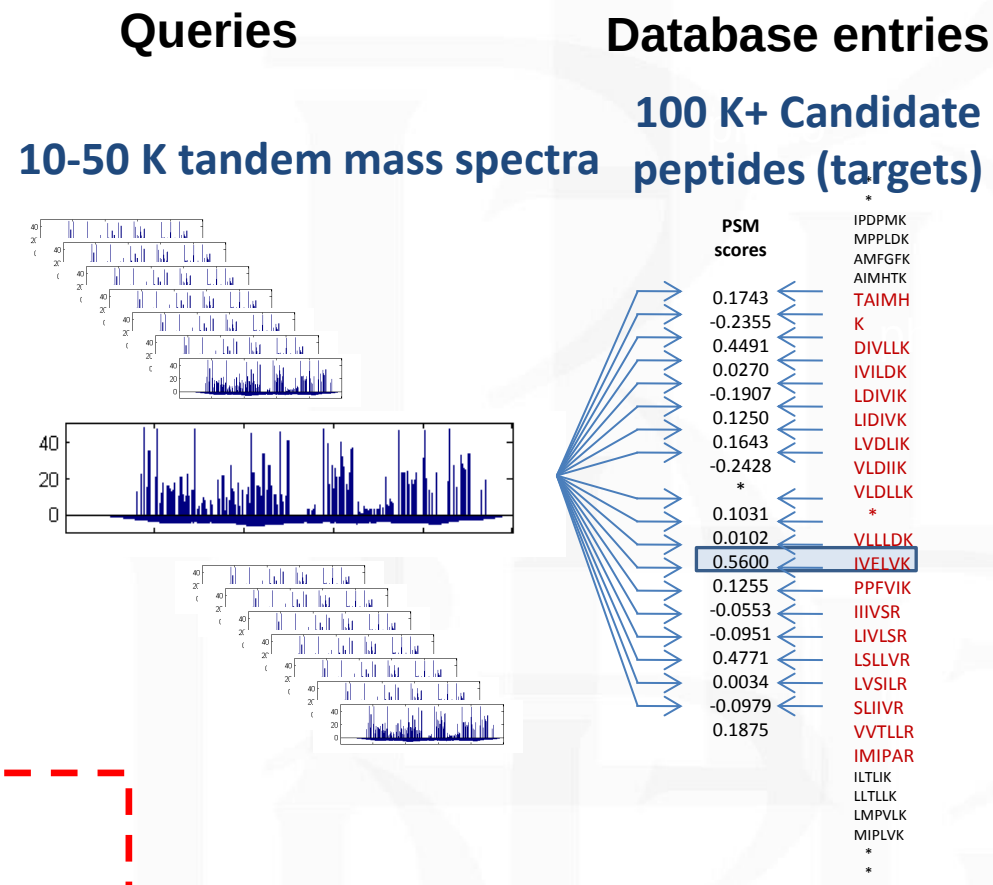
What is important for the identification?

- Location of the peak
- The distance between peaks



How annotation goes?

An experimental spectrum is iteratively matched against a large database, which contains a possible annotations.
The spectrum is annotated with the best scoring database entry.



Matching is based on simple scalar product scoring!

Dataset

Spectra type	Quantity	Volume (TB)
Experimental	9 092 380	7
Theoretical	19 874 734	1.9

The identification process is hampered by:

1. noise peaks
2. missing peaks
3. modifications in the data (correct annotations not included to the reference database)

50-80% of the data cannot be annotated with high confidence!

Expanding the reference dataset would decrease the statistical confidence scores (due to multiple testing correction) => we need a scoring function robust to multiple testing corrections!

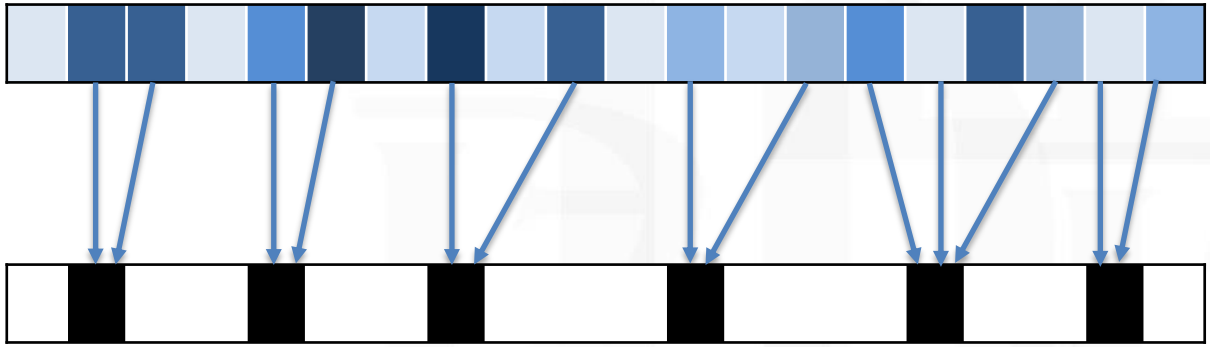
Our approach (under development)

Modeling method: Restricted Boltzmann Machines
Low resolution (old instruments): 2 000 bins
High resolution (new instruments): 1 000 000 bins

Experimental spectrum

W

Theoretical spectrum



$$P(v,h) = \frac{1}{Z}\exp(v^T W h)$$

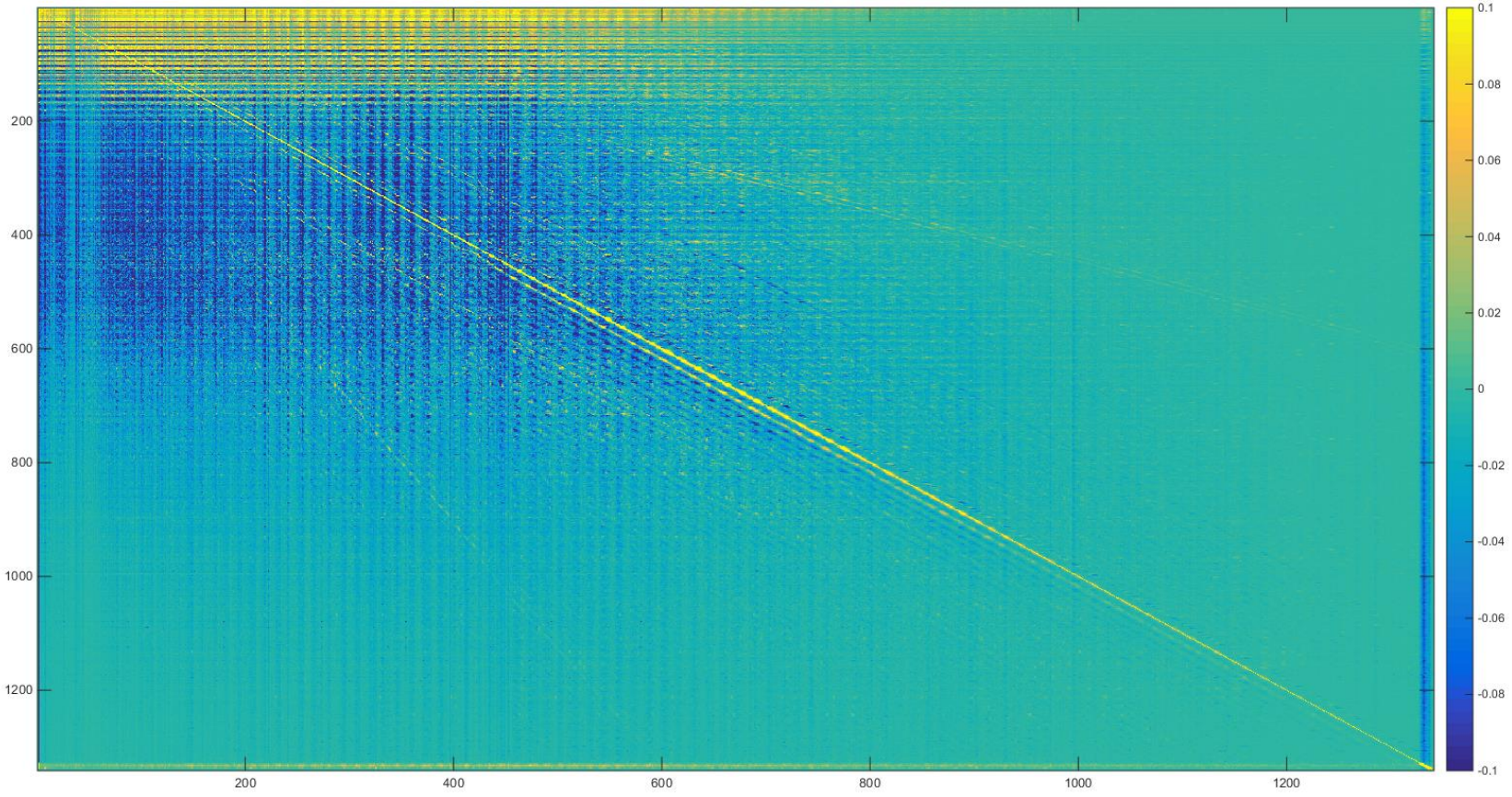
W (100K x 100K) will be very big to learn!

Challenges

- Large dimensions (for high resolution data)
- Traditional training methods do not work: they use sampling of h with MCMC, but then h will be garbage as usual weights update rule $\Delta w_{ij} = \eta(E[v_i h_j]^0 - E[v_i h_j]^\infty)$ => we approximate h from database
- For the training we approximate $P(v,h) \sim P(v,v)$
- Therefore we got a fully observed dataset so the training is straightforward
- Ubiquitous peaks they appear in every data and cause high correlation with all other data

Results

The weight matrix W learnt:



Future work

1. Suppress the effect from ubiquitous peaks using some regularization methods
2. Run experiment on high resolution data
3. Speeding up the training (there is a *log-sum-exp* in the training, so it would be nice to eliminate it with some accurate approximations)

NATIONAL RESEARCH UNIVERSITY

Thank you for your attention!

National Research University Higher School of Economics (HSE)
20, Myasnitskaya str., Moscow, Russia, 101000
Tel.: +7 (495) 628-8829, Fax: +7 (495) 628-7931
www.hse.ru