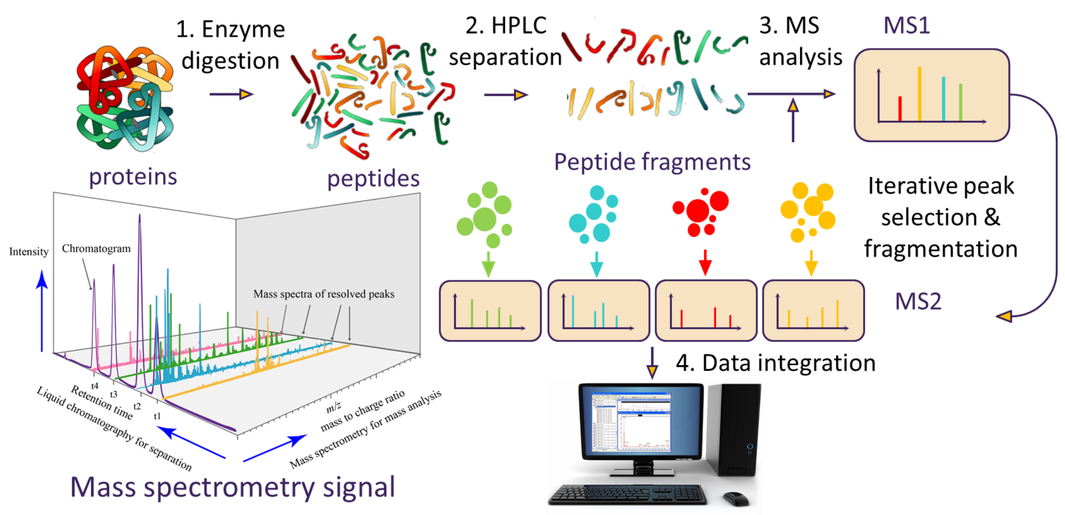
Scalable computational method for big proteomics data

Olga PERMIAKOVA, Thomas BURGER
Interdisciplinary Research Institute of Grenoble (IRIG), Exploring the Dynamics of Proteomes (EDyP)
olga.permiakova@cea.fr, thomas.burger@cea.fr

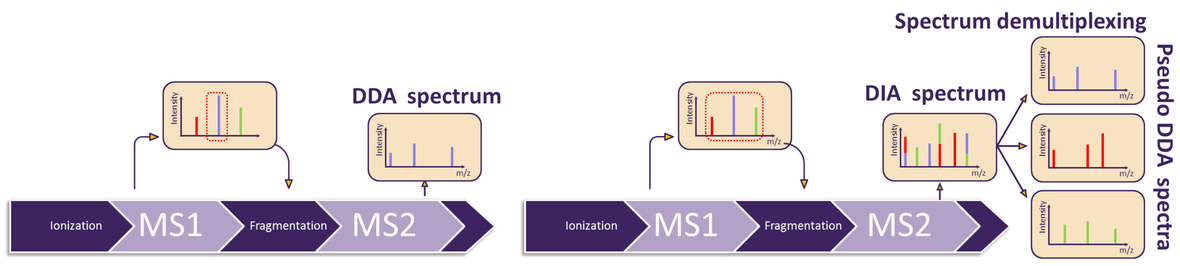
1. Biological context

The main goal of proteomics research is to explore the proteome composition so as to understand protein functions. Tandem mass spectrometry (MS/MS), coupled with liquid chromatography (LC) is a powerful and the most widely used technique for high throughput identification and quantification of proteins in complex biological mixtures.

LC-MS/MS pipeline



Two modes of analysis



☺ Straightforward identification

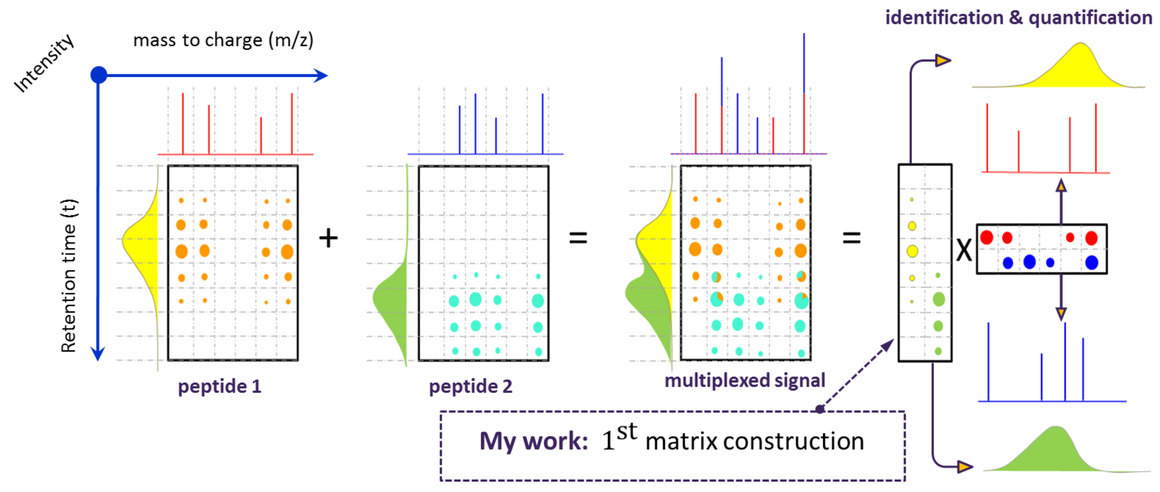
☹ Incomplete analysis

☺ More comprehensive analysis

☹ Multiplexed spectrum

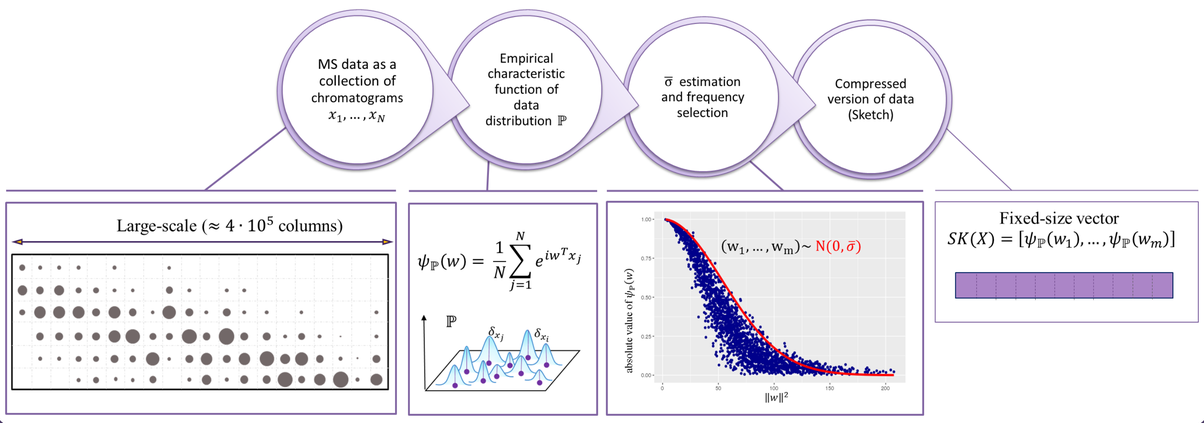
2. From biochemistry to mathematics

Reveal-MS project: Spectrum demultiplexing can be achieved via Non-Negative Matrix Factorization (NMF). The matrix built from data is decomposed into two factors: **dictionary** (collection of chromatograms) and **sparse code** (pseudo DDA spectra).



3. Data compression procedure

Mass spectrometry data is voluminous and processing them is a challenging task. A possible way to tackle this problem is to construct a fixed-size data summary, and then find out a matrix decomposition relied on the compressed data.



4. Compressive Dictionary Learning

Algorithm 1 Stochastic Compressive Dictionary Learning

Require: $SK(X) \in \mathbb{C}^m$ data sketch, K dictionary size, $S_R \in \{0,1\}^{N \times n}$ random column sampling matrix, $n \ll N$

$D_0 \leftarrow \emptyset$, $r \leftarrow SK(X)$

while $|D_0| \leq K$ **do** ▷ Dictionary initialization

$$D_0 = \left\{ D_0, d_{new} \leftarrow \arg\min_{d \in \mathbb{R}^s} \Re\left\langle \frac{SK(d)}{\|SK(d)\|^2}, r \right\rangle \right\}$$

$$\beta^* \leftarrow \arg \min_{\substack{0 \leq \beta \\ \beta \in \mathbb{R}^{|D_0|}}} \left\| \sum_{k=1}^{|D_0|} \beta_k \cdot SK(d_k) - SK(X) \right\|^2$$

$$r = SK(X) - \sum_{k=1}^{|D_0|} \beta_k^* \cdot SK(d_k) \in \mathbb{C}^m$$

end while

for $t \leftarrow 1$ to T **do** ▷ Stochastic optimization w.r.t. code A

$A_{t-1} \leftarrow \arg \min_{A \in \mathbb{R}^{K \times n}} \|D_{t-1} \cdot A - X \cdot S_R\|_2^2$ ▷ Code initialization

$D_t \leftarrow \arg \min_{\substack{D \in \mathbb{R}^{s \times K} \\ A \in \mathbb{R}^{K \times n}, A \geq 0}} \|SK(D \cdot A) - SK(X)\|^2$ ▷ Dictionary update

end for

return D_T

5. Results

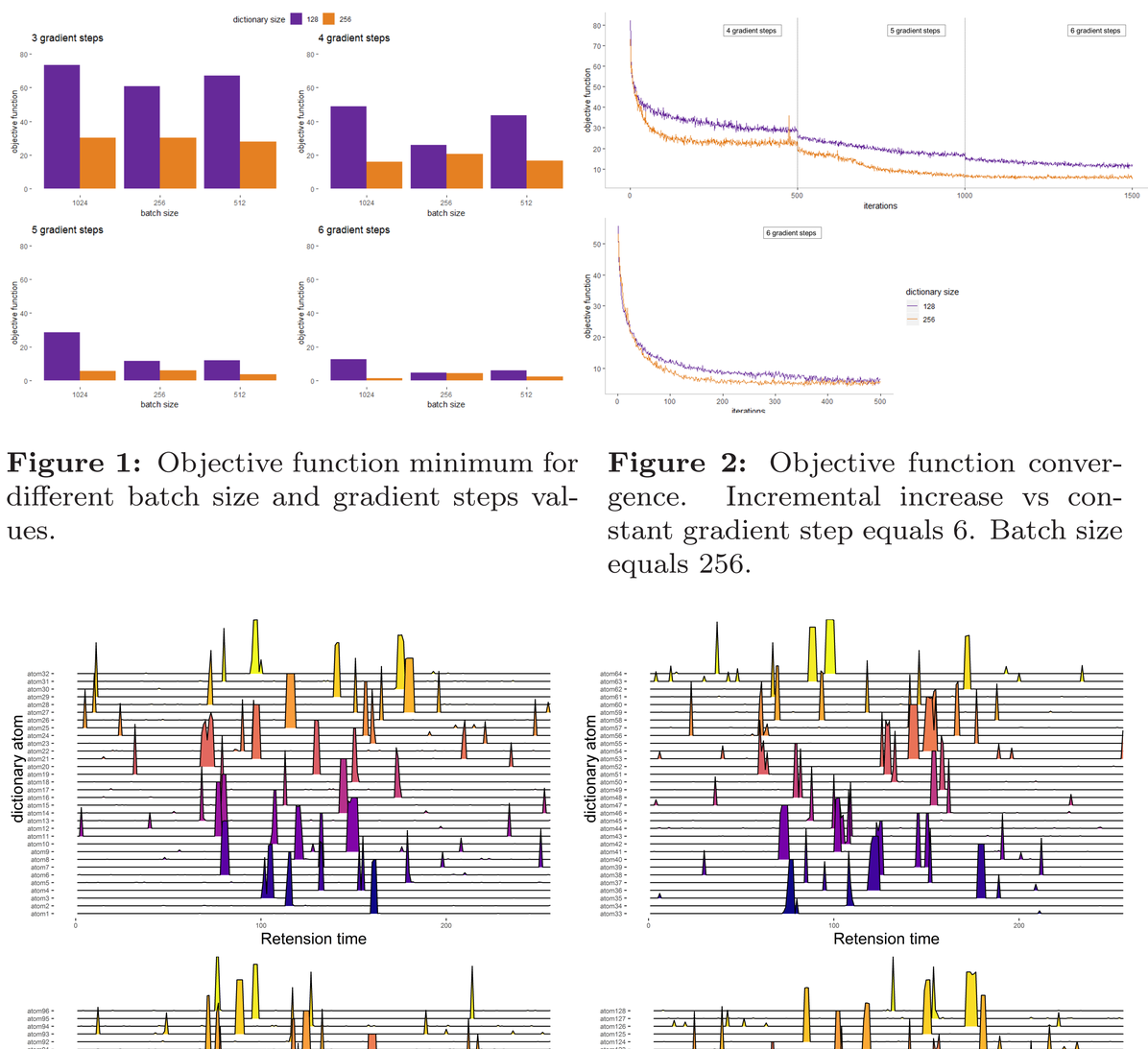


Figure 1: Objective function minimum for different batch size and gradient steps values.

Figure 2: Objective function convergence. Incremental increase vs constant gradient step equals 6. Batch size equals 256.

Figure 3: Obtained dictionary

6. Conclusions

- ☑ A new method for learning dictionary from sketch of data is developed and implemented.
- ☑ Proposed approach is computationally efficient and it allows to construct big size dictionary.
- ☑ Obtained dictionary atoms are consistent with the real chromatographic profiles

7. References

[1] N. Courty, X. Gong, J. Vandel, and T. Burger. Saga: Sparse and geometry-aware non-negative matrix factorization through non-linear local embedding. *Machine learning*, 97(1-2):205–226, 2014.

[2] N. Keriven, A. Bourrier, R. Gribonval, and P. Pérez. Sketching for large-scale learning of mixture models. *CoRR*, abs/1606.02838, 2016.