Clustering-based ion chromatogram extraction and peak-picking for high-resolution LC-MS data

Martin Loos[1], Matthias Ruff[1], Heinz Singer[1], Juliane Hollender[1]

[1] *Eawag Dübendorf, Überlandstrasse 133, 8600 Dübendorf, Switzerland*

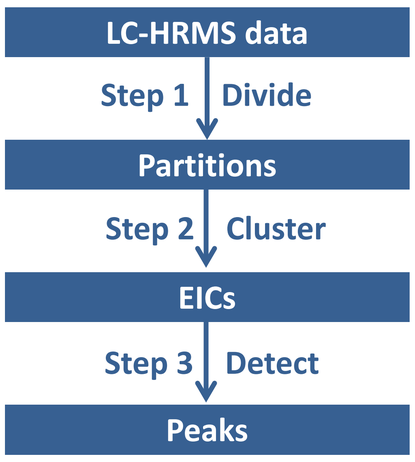


Figure 1: *Peak picking with enviPick is done in three consecutive steps.*

Motivation

Peak picking joins individual LC-HRMS data points into a single mass signal. As a critical step of data reduction and analysis, peak picking must deal with:

- Noisy, heterogeneous and large data sets as shown in Figure 2.
- Chromatographic peaks varying strongly in shape and width.
- Isobaric compounds and chemical baselines.
- Data preprocessed at acquisition (baseline-correction, centroidization).
- Interference of (un)resolved masses even at high resolutions.
- High priority and confidence in data points of high intensity.

Existing peak picking strategies do not cover all of the above aspects. Hence, a novel approach is presented to extract ion chromatograms (EICs) and to detect individual signal peaks, based on three steps (cp. Figure 1).

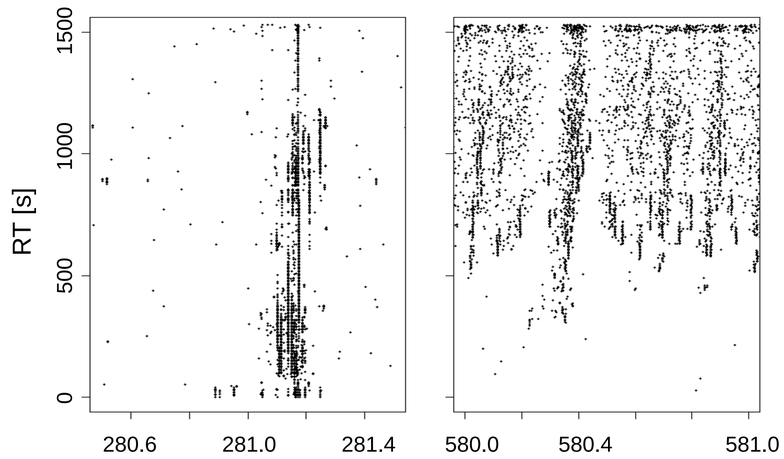
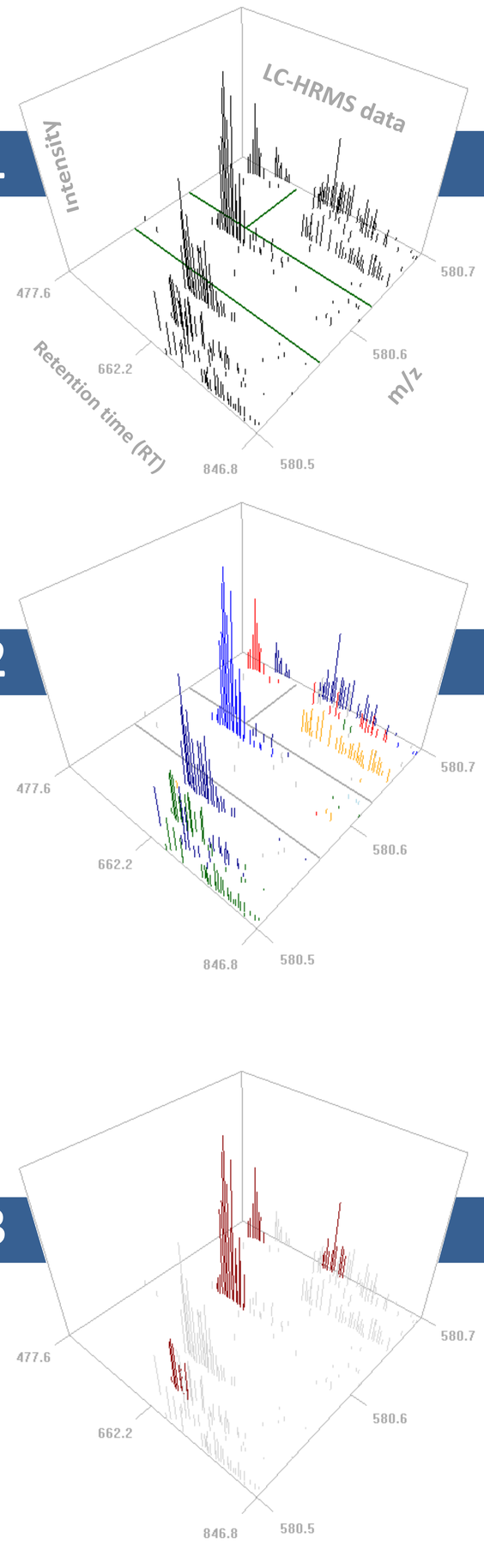


Figure 2*: Data from the same LC-HRMS experiment differs strongly depending on the m/z-region (Orbitrap XL Velos Pro, R60K@m/z400).*



Step 1 Data partitioning

Given: set of centroided and baseline-corrected LC-HRMS data points *m*

Divide data into unrelated regions to acccelerate and parallelize computation of step 2:

- Link each data point to its neigbours found within large tolerances of ΔRT_L and $\Delta m/z_L$
- Group all directly and indirectly linked data points into a single partition P_x

The partition size (number of data points) often increases with m/z.

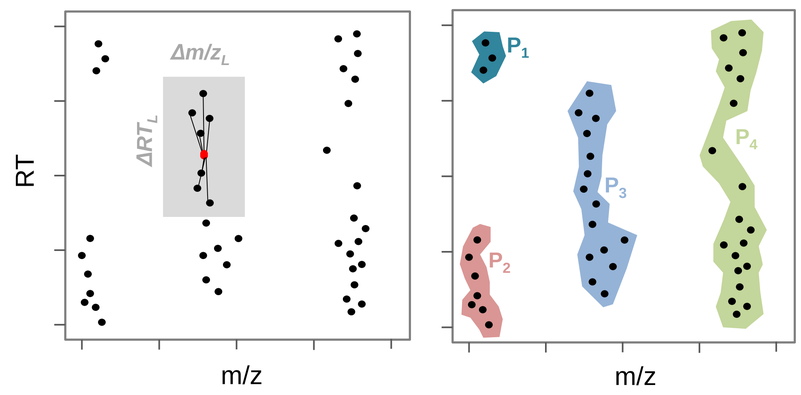


Figure 3*: LC-HRMS data points during the partitioning process*

Step 2 EIC clustering

Given: *n* **partitions** $P=\{P_{x=1},\ldots,P_{x=n}\}$ **with** *k* **data points** $m_x=\{m_{x,1},\ldots,m_{x,k}\}$ *in each.*

(A) Cluster data points in each partition.

(1) Start with an empty cluster set $C_x = \{\}$.
(2) Select data point $m_{x,1}$ from intensity-ranked data m_x.
(3) Is $m_{x,1}$ assignable to any of *w* existing cluster $C_x=\{C_{x,y=1},\ldots,C_{x,y=w}\}$ within
(a) $\Delta m/z_{x,y}$, $\Delta RT_{x,y}$ and **(b)** no other clustered $m_{x,y,i\leq p}$ with identical *RT*?
No: Assign $m_{x,w+1,1} = m_{x,1}$. Add new cluster $C_{x,w+1}=\{m_{x,w+1,1}\}$ to C_x.
Set $\Delta m/z_{x,w+1}$ = 2*(*m/z* range of mass precision).
Set $\Delta RT_{x,w+1}$ = *RT search window* << RT_L.
Yes: Add measurement to the cluster *y* with smallest difference in mean *m/z*, i.e., assign $m_{x,y,p+1} = m_{x,1}$. Adapt $\Delta m/z_{x,y}$. Extent $\Delta RT_{x,y}$.
(4) Remove $m_{x,1}$ from m_x.
Repeat (2) to **(4)** until $m_x=\emptyset$.

(B) Merge cluster in C_x:

(1) List all cluster pairs with data points **(a)** nested in each others $\Delta m/z_{x,y}$ and **(b)** not having duplicated *RT*.
(2) Merge cluster pair with smallest mean *m/z* difference.
(3) Update: **(a)** $\Delta m/z_{x,y}$ of merged cluster, **(b)** C_x and **(c)** list of cluster pairs.
Repeat (1) to **(3)** until list of cluster pairs is empty.

Each cluster in the final set C_x corresponds to an extracted ion chromatogram (EIC).

Step 3 Peak detection & filtering

Given: *w* **EICs** $C_x=\{C_{x,y=1},\ldots,C_{x,y=w}\}$ **with** *p* **data points** $m_{x,y}=\{m_{x,y,1},\ldots,m_{x,y,p}\}$ **.**

(A) Detect up to *q* peaks in each EIC*:*

(1) Order $m_{x,y}$ by RT and interpolate data gaps ≤ ΔRT_{gap}.
(2) Select most intense data point $m_{x,y,max}$ as candidate peak apex.
(3) $S_d(n)$ = sum of intensity decreases between $m_{x,y,max}$ and $m_{x,y,n}$.
(4) $S_i(n)$ = sum of intensity increases between $m_{x,y,max}$ and $m_{x,y,n}$.
(5) For *n>max,* set upper peak bound n_{UB} as $\underset{n}{\operatorname{argmax}} \; Sd(n) - \gamma S_i(n)$
(6) For *n<max,* set lower peak bound n_{LB} as $\underset{n}{\operatorname{argmax}} \; Si(n) - \gamma S_d(n)$
(7) Remove data points of peak $n_{LB} \leq n \leq n_{UB}$ from $m_{x,y}$
Repeat (2) to **(7)** at most *q* times

(8) If $m_{x,y}\neq\emptyset$, set baseline: interpolate gaps and smooth
(9) Subtract baseline intensity from peak intensities
(10) Set EIC noise *N* as median intensity deviation of $m_{x,y}$ from baseline

(B) Filter peaks by *S/N*, *S/B*, intensity threshold, # data points per ΔRT

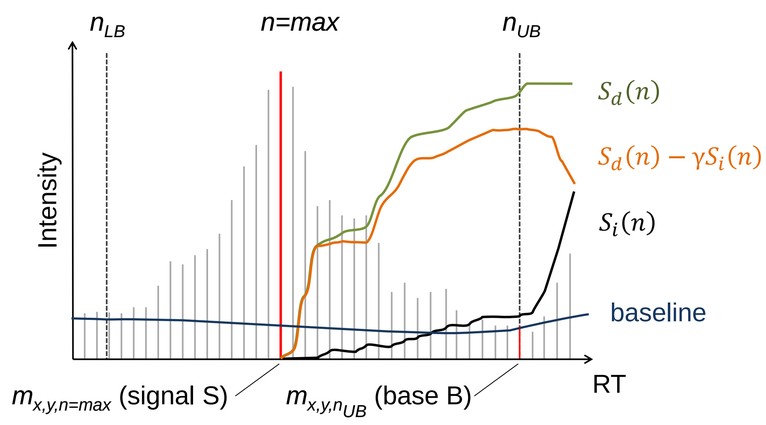


Figure 4*: LC-HRMS data points (grey) in an EIC subsection. Notations refer to the peak picking process of step 3.*

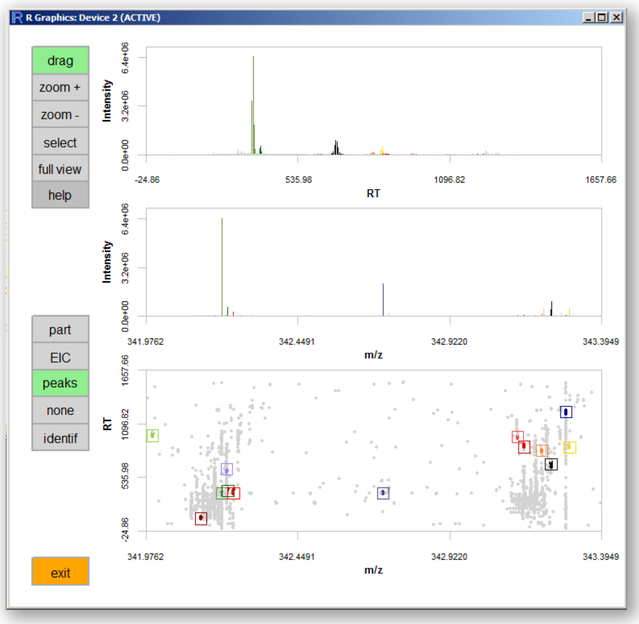


Figure 5*: The interactive data and result viewer of enviPick.*

enviPick R package

- **Raw data & result viewer**
- **Browser-based user interface**
- **Wrapper & step-wise functions**
- **Batch processing**
- **Freely available (GPL-3)**
- **Reads .mzXML data:**
- **Baseline-corrected**
- **Centroided**

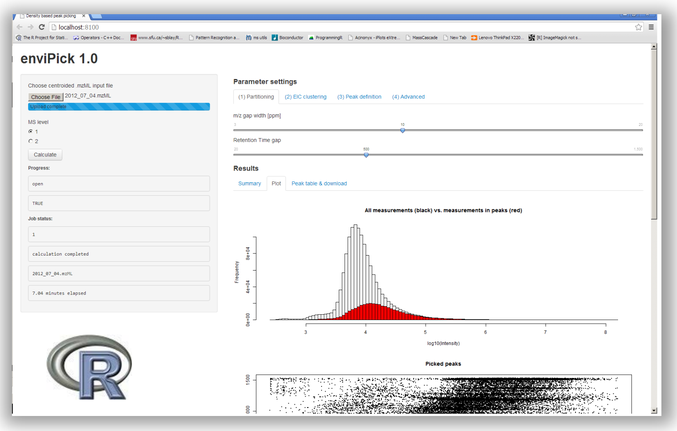


Figure 6*: enviPick provides a convenient user interface.*