# Bayesian approach to peak deconvolution and library search for high resolution gas chromatography-Mass spectrometry

Barcaru, A.; Mol, H.G.J.; Tienstra, M.; Vivó-Truyols, G.

**DOI**
10.1016/j.aca.2017.06.044

**Publication date**
2017

**Document Version**
Other version

**Published in**
Analytica Chimica Acta

# Supporting information

# Bayesian approach to peak deconvolution and library search for high resolution gas chromatography – mass spectrometry

A. Barcaru[a,*], H.G.J. Mol[b], M. Tienstra[b], G. Vivó-Truyols[a]

a Analytical Chemistry Group, van't Hoff Institute for Molecular Sciences, University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, The Netherlands

b RIKILT – Wageningen University & Research, Akkermaalsbos 2, 6708 WB, Wageningen, The Netherlands

## 1. Theory

### 1.1 The channel-wise model fitting

The objective of this step in the pipeline is to find – in a probabilistic fashion – the most optimal parameters (i.e. retention time, peak width) for each mass-to-charge channel. Prior to this step, the data was binned and included in a matrix form of *QxJ* elements, to ease the implementation of the algorithm. For information about the binning method, see Supporting Information. The channel-wise modelling is based on work of Sivia et al. [11] which proposed a Bayesian probabilistic solution to the calculation of the potential number of Gaussian models in spectral data. The strategy is based on the following modelling of the data:

$$S_j = \sum_{i=0}^{N} A_{i,j} e^{\left(-\frac{\left(t-t_{i,j}\right)^2}{2{\sigma_{i,j}}^2}\right)} \quad (1)$$

Where $N$ is the number of components, $S_j$ is the modeled signal (i.e. the chromatogram for one fixed m/z channel $j$), $t$ is the time axis (containing Q elements), $A_{i,j}$ is the amplitude (or peak height) for the i[th] peak at the j[th] channel, $t_{i,j}$ is the elution time for the analyte $i$ at channel j, and $\sigma_{i,j}$ is the peak width for the analyte $i$ at the channel j. When fitting the model from the Eq. (1), the objective is to find the values of the parameters $A_{i,j}, t_{i,j}$ and $\sigma_{i,j}$ that minimize the $\chi_j^2$ function between the model and the data from channel $j$:

$${\chi^2}_j = \sum_{q=1}^{Q} \frac{\left(D_{jq} - S_{jq}\right)^2}{{\sigma_n}^2} \quad (2)$$

Where $D_{jq}$ is the q[th] element (i.e. time registry) of the Dj data vector (of Q elements), corresponding to the chromatogram of channel j. ${\sigma_n}^2$ is the standard deviation of the noise of the signal. Eq. (1) can be solved probabilisitically in a Bayesian framework, calculating the posterior probability of the number of

components (N) [11]. However, in this work [11] , such methodology is not extended to two-dimensional data.

A simple extension of the ideas of [10] would be to fit each channel independently using Eqs. (1) and (2). The main problem of such approach is that the fitted retention times for the same compound will be highly different from channel to channel. In fact, by performing independent fittings for each channel using the above equations, we are assuming that the channels in High Resolution GC-Orbitrap are statistically independent. This is not true, since the chromatographic profile of a compound (and therefore the value of $t_{i,j}$ and $\sigma_{i,j}$ in Eq. (1)) is the same, independently for the channel j. Solving this issue directly causes a major computational burden, forcing all the channels to be fitted simultaneously to Eq. (1), using common values of of $t_i$ and $\sigma_i$ per compound. As this is unfeasible from a practical perspective, we have decided to modify the objective function to include a part of the contribution of other channels when finding the fitted parameters. Our proposal in solving these issues is to include a penalty on the objective function that will penalize the differences in $t_{i,j}$ and $\sigma_{i,j}$ across channels:

$$\widetilde{\chi^2}_j = \sum_{q=1}^{Q}\left[\frac{\left(D_{jq} - S_{jq}\right)^2}{{\sigma_n}^2} + \lambda\sum_{j}^{J}\left(\widetilde{\boldsymbol{D}} - \boldsymbol{S}\right)^2\right] \quad 3$$

Where the matrix $\widetilde{\boldsymbol{D}}$ has each column normalized (i.e. each m/z channel has the intensities between 0 and 1) and **S** is the matrix in which each column is equal to the fitted signal $S_j$, also normalized. Note that, with this particular construction of **S**, the left-hand side part of the equation estimates the goodness of fit of the proposed for all the channels at the same time. The penalization parameter $\lambda$ governs how common the chromatographic profiles should be across channels. Its value is subjected to a best estimation. A large value of $\lambda$ will lead to extremely high bias towards the first moment of the TIC peak. A very small value of the $\lambda$ will assume almost independence between the mass channels. This second option, given the noise in the data, results in a high dissimilarities between the channel-wise optimized values of the fitted retention time and peak width. We propose the value of $\lambda = \frac{Q}{J}$ as recommended in [13]. This value was found optimal in this work after several empirical trials.

The question of the number of components can be solved probabilistically simply by computing the posterior probability of a particular number of components in the mixture using the Bayesian equation [11]:

$$p(N|D) = \frac{p(D|N) \times p(N)}{p(D)} \quad 4$$

Where $p(N|D)$ is the posterior probability of the number of components (N) after the data has been taken into account, $p(D|N)$ is the so-called likelihood (the probability of obtaining the data given a proposed number of compounds), $p(N)$ is the prior probability of the number of compounds and p(D) is a normalization factor (not calculated explicitly). By applying a uniform prior on N and marginalizing over the parameters of the model from the equation (1) we can rewrite the equation 4:

$$p(N|D)_j \propto \int_{A,t,\sigma} p(D|A,t,\sigma)_j \, p(A)\, p(t)\, p(\sigma)\, dA\, dt\, d\sigma \quad 5$$

Where $p(D|A,t,\sigma)_j$ is the likelihood function that can be obtained from Eq. (2) or (3) if Gaussian noise is assumed. The integral expressed in eq. (5) might be solved using MCMC techniques, but this can be time consuming. An alternative is to solve it by proposing a Taylor expansion of the integrand (i.e. of the expansion of the Chi-square distribution from eq. (2)) [10]:

$$\chi^2 \approx \chi_0^2 + \frac{1}{2}(\mathrm{X}-\mathrm{X}_o)^T \nabla\nabla\chi_0^2 (\mathrm{X}-\mathrm{X}_o) + \cdots \qquad 6$$

Where *X* is any set of parameters of the model and $X_o = \{A, t, \sigma\}$ is the set of parameters yieldng the minimum value of $\chi_0^2$ of the objective function from the eq. (3). In this equation, $\nabla\nabla\chi_0^2$ is the Hessian matrix of the objective function. By imposing flat priors on the nuisance parameters ($p(A)\ p(t)\ p(\sigma)$) we obtain:

$$\int_{A,t,\sigma} p(D|b,A,t,\sigma)_j\ p(a)p(t)\ dA\ dt\ d\sigma \approx \frac{\int_{A,t,\sigma} e^{\left[-\frac{1}{4}(\mathrm{X}-\mathrm{X}_o)^T \nabla\nabla\chi_0^2 (\mathrm{X}-\mathrm{X}_o)\right]}\ dA\ dt\ d\sigma}{((t_{max}-t_{min})(A_{max}-A_{min})(\sigma_{max}-\sigma_{min}))^N} \qquad 7$$

Where the $[t_{max}, t_{min}]$, $[A_{max}, A_{min}]$, and $[\sigma_{max}, \sigma_{min}]$ are the limits of the (flat) prior distributions of the nuisance parameters. Note that the solution, although coming from flat priors, is biased in our case since we take into account the inter-dependence of the mass channels. The integral on the right hand side of the Eq. (7) is well known integral of a multivariate Gaussian distribution. By using this solution and assuming that the different N models are indistinguishable and interchangeable, we can now express Eq. (5) using Eq. (7) as follows [10]:

$$p(N|D)_j \propto \frac{(4\pi)^{N/2} N!}{((t_{max}-t_{min})(A_{max}-A_{min})(\sigma_{max}-\sigma_{min}))^N} \frac{e^{-\frac{\chi_0^2}{2}}}{\sqrt{\det(\nabla\nabla\chi_0^2)}} \qquad 8$$

Note that the denominator and numerator containing the parameter N in the eq. (8), serves as the so-called “Occam factor” and decreases the posterior probability if the complexity of the model increases.

We will be further interested in the *“retention time - peak width”* (*RT-PW*) space that is obtained from the first step of the pipeline. Basically the values of fitted $t_{i,j}$ and $\sigma_{i,j}$ for the different values of N are of interest. The values of p(N|D) might be different per channel. Instead of deducing the true number of compounds from p(N|D), we use this value as a threshold (i.e. only the data points $p(N|D) > 10^{-5}$ are considered). All the fitted $t_{i,j}$ and $\sigma_{i,j}$ values (for all channels) generating p(N|D) above a threshold are used to populate the RT-PW space.

The number of clusters found in this space should give a hint on the number of compounds. In other words, for each analyte we would expect to have a high agglomeration of points in RT-PW space around the retention time and peak width of such analyte. The penalty λ used in Eq. (3) reduces the sparsity of

the clusters in the *RT-PW* space. However, the noise is still partially affecting the parameters obtained at this step.

We assume that the values of retention time and peak width follow a multivariate Gaussian distribution (in the RT-PW space) for each compound. From here, the idea of fitting a Gaussian mixture model (MM) for all points found in this space arises.

## 1.2 The Mixture Model classification

In order to fit the Mixture of Gaussians to the data, the Expectation maximization (EM) algorithm is employed. The EM algorithm has been extensively described in the literature [14] [15] and only a brief description is given here. The algorithm starts with a proposal for the center and the variance of each cluster i, $\mathbf{c}_i$, and an estimate of the variance for each cluster, $\mathbf{\Sigma}_{ii}$. $\boldsymbol{c_i} = [\overline{t_r}, \overline{\sigma}]_i$ is the expected value of the centroid of the cluster $i$ (i.e. the expected value of the retention time and peak width of such chromatographic peak) At the next step, the latent parameter $\omega_{m,i}$ is calculated. This latent parameter is found for each of the M data points and for each cluster i. This latent parameter (called "responsibility") is defined as the likelihood, $\omega_{m,i} = p(\boldsymbol{x_m}|\boldsymbol{c_i}, \sum \boldsymbol{i})$, where $\mathbf{x}_m$ is a data point (containing a pair value in the RT-PW space). Note that we have defined the whole data points using $\boldsymbol{x} = [t_{r1}, \sigma_1; \ldots t_{rm}, \sigma_m; \ldots t_{rM}, \sigma_M;]$. This likelihood is assumed to be multivariate Gaussian, i.e. $p(\boldsymbol{x_m}|\boldsymbol{c_i}, \sum \boldsymbol{i}) = (2\pi)^{-\frac{K}{2}} \det(\boldsymbol{\Sigma_i})^{-\frac{1}{2}} exp\left(-\frac{1}{2}(\boldsymbol{x_m} - \boldsymbol{c_i})^T \boldsymbol{\Sigma_i}^{-1}(\boldsymbol{x_m} - \boldsymbol{c_i})\right)$ where K=2 in our case since the RT-PW is a bidimensional space. In the next iteration, an update of the $\mathbf{c}_i$ and $\sum \boldsymbol{i}$ is obtained using the new values of $\omega_{m,i}$. For the case of ci, the update is calculated as follows:

$$\boldsymbol{c_i} = \frac{\sum_{m=1}^{M} \omega_{m,i} \boldsymbol{x}_m}{\sum_{m=1}^{M} \omega_{m.i}} \qquad 10$$

For $\boldsymbol{\Sigma_i}$, we force the matrix to be diagonal, since we assume in our case that parameters RT and PW do not influence each other. Each diagonal element (i.e. the k-th element, k={1,2}) is thus computed as follows:

$$(\Sigma_i)_{k,k} = \frac{\sum_{m=1}^{M} \omega_{m,i}(x_{m,k} - c_{i,k})(x_{m,k} - c_{i,k})}{\sum_{m=1}^{M} \omega_{m,i}} \qquad 11$$

These new parameters of $\mathbf{c}_i$ and $\mathbf{\Sigma}_i$ are then used to update the responsibilities $w_{m,i}$ (Eq. 9), and the algorithm continues up to convergence. The values of $\mathbf{c}_i$ and $\mathbf{\Sigma}_I$ define the center of each cluster and its bandwidth. In other words, they describe the features (retention time and band broadening) of the i-th eluting peak.

The main factors influencing the final (converged) values of $\mathbf{c}_i$ and $\mathbf{\Sigma}_I$ are: the density of each cluster (i.e. number of points concentrated in one cluster) and the total number of the data points (M). It is important to stress here that the number of mass channels of the compounds that are in the dataset can severely influence the fitting of $\mathbf{c}_i$ and $\mathbf{\Sigma}_I$. If a compound (i.e. a cluster in RT-PW space) has a larger number of channels compared to the neighboring compounds, it is more likely that the centers of other clusters (eq. 10) will be biased (i.e. "attracted") towards the center of this cluster, which can unbalance the final result. One solution to this handicap is to limit the value of $(\Sigma_i)_{11}$ to a maximum possible value. This is done by imposing a threshold during the fitting. If, during an iteration, the value of $(\Sigma_i)_{11}$ found in Eq. 11 is above

a threshold (in our case the value of the threshold is 0.064 s), the threshold value is assigned to $(\Sigma_i)_{11}$instead of the value calculated with Eq. (11). In practice, this means limiting the influence of a centroid by imposing a maximum value on the sparsity of the clusters in the RT-PW space. Another important role of $(\Sigma_i)_{22}$ is to control the uniqueness of the retention time. In other words, we assume that there are no analytes eluting at exactly the same retention time. To impose this condition, we follow a similar strategy compared to the method used in $(\Sigma_i)_{11}$. However, in this case the threshld for $(\Sigma_i)_{22}$ defines its minimum, not its maximum. In other words: if, during an iteration, the value of $(\Sigma_i)_{22}$ is below a threshold (i.e. 0.64 s), the value of the threshold is used, instead of the value given by Eq. 11. The values of both thresholds (0.064 and 0.64) were found empirically. More explanation about these values is found in the results and discussion section.

One of the most important problems in clustering in general is how to answer to the question on how many clusters should be used to describe the data. Note that the number of cluster expresses the number of eluting analytes within the region of interest. In this work we propose the use the well-known Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) to evaluate the number of clusters. Both BIC and AIC are forms of penalization of the log-likelihood (i.e. $L = \sum_{m=1}^{M} \ln\left[\sum_{i=1}^{N} p(\boldsymbol{x_m}|\boldsymbol{c_i},\sum_{\boldsymbol{i}}) \right]$):

$$BIC = -2L + \beta \ln M \quad (12)$$

$$AIC = 2\beta - L \quad (13)$$

With the number of free parameters $\beta = (3N-1)^K$ , M is the number of data points in the *RT-PW* space, *K* is the dimension of the multivariate distribution (2 in our case). The value of BIC and AIC is used to select the optimal value of N. The value of N yielding the lowest value of BIC (or AIC) is considered optimal. Further, the BIC can be used to calculate the posterior probability for N: [14]

$$p(N|BIC) = \frac{e^{\frac{1}{2}BIC_N}}{\sum_N e^{\frac{1}{2}BIC_N}} \quad (14)$$

Where $BIC_N$ represents the value of BIC (Eq. 12) using N components.

## 1.3 Compound identification and spectral retrieval

This step makes use of the posterior probability of the number of components (Eq. 14) resulted from the previous step and the corresponding values of the centroids, $\mathbf{c}_i$. Note that, in fact, the values of $\boldsymbol{\Sigma}_i$ are not of interest for the next steps of the algorithm. This is because $\mathbf{c}_i$ describes at full the peak feature of the $i^{th}$ component: it describes its position (RT) and its band broadening (PW). A simple optimization algorithm is employed to find the intensities $A_{i,j}$ for each compound at each j channel. We simply introduce the value of RT (i.e. $t_i$ in Eq. 1) and PW ($\sigma_i$ in Eq. (1)) for each compound $i = 1,2, \dots N$ and solve (via least squares) the value of $A_{i,j}$ for each channel and compound. Note that step 2 of the algorithm made the values of $\mathbf{c}_i$ common for each channel, and therefore $\boldsymbol{t_i}$and $\boldsymbol{\sigma_i}$ do not depend on $j$. The intensities of $A_{i,j}$ found via least-squares estimation of Eq. (1), have the imposed condition of non-negativity. At the end of

the process, a matrix (i.e. $\boldsymbol{A}_{i,j}$) of $A_{i,j}$ intensities for each $i$ ($i = 1,2, \dots N$) compound and $j$ channel ($j = 1,2, \dots J$) is obtained. For a fixed value of $i$, the vector $\boldsymbol{A}_{i,*}$ (for all $j = 1,2, \dots J$ values) constitutes the retrieved spectrum for the $i^{th}$ compound. It is likely to observe channels that are selective for only one compound. In this case, the algorithm will simply assign null intensities for this channel for the other compounds. We consider in this work an estimation of the uncertainty (calculated from the variance-covariance matrix of the fitting [12]) of the fitting. These values (standard deviation, $\delta_{i,j}$) are used in the next step.

In order to compare the retrieved spectra $\boldsymbol{A}_{i,*}$ with a spectra available in the database, we make use of the correlation coefficient as follows. First, the query spectra $\boldsymbol{s}_\tau$ from the database is interpolated to the same m/z values as the unknown spectra $\boldsymbol{A}_{i,*}$ retrieved from the previous step (for information about the interpolation method, see supporting information). The value of $\tau$ defines an arbitrary element of the library containing T spectra. To simplify the terminology, let's define **a** as a generic $i^{th}$ spectrum, retrieved from the data, and containing J channels, $\boldsymbol{a} \equiv \boldsymbol{A}_{i,*} = [\, A_{i,1}, A_{i,2}, \dots, A_{i,j}, \dots, A_{i,J}]$. Let's define $\boldsymbol{\delta}$ as the vector of standard deviations found in the previous step, $\boldsymbol{\delta} \equiv \boldsymbol{\delta}_{\boldsymbol{i},*} = [\delta_{i,1}, \delta_{i,2}, \dots, \delta_{i,j}, \dots, \delta_{i,J}]$ . The values of **a** are then perturbed drawing from a normal distribution with mean $A_{i,j}$ and standard deviation $\delta_{i,j}$, obtained a perturbed $\boldsymbol{a}_r$ spectrum. The correlation coefficient between the perturbed spectrum and $\boldsymbol{s}_\tau$ is calculated. By performing this calculation $R$ times (i.e. randomly drawing from the distribution centered at $\boldsymbol{a}$ and standard deviation $\boldsymbol{\delta}$ ) and calculating the mean of all these correlations, an average correlation is found, which includes the uncertainty in the estimation of **a**:

$$\rho_{\tau,i,N} = \frac{1}{R} \sum_{r=1}^{R} \left[ \frac{\boldsymbol{s}_\tau \cdot \boldsymbol{a}_r}{\|\boldsymbol{s}_\tau\| \|\boldsymbol{a}_r\|} \right]^2 \qquad 15$$

A value of $R$ of $500$ was found appropriate. $\tau$ indicates the id of a compound in the database ($\tau = 1, .. \mathrm{T}$ with T total number of the spectra in the library). The output will be a value of $\rho_{\tau,i,N}$ (i.e the average correlation with the spectrum $\tau$ of the database) assigned to every element in the database. Note that we made explicit that this correlation depends on i, i.e. it is obtained for each of the i=1,2,…, N spectra retrieved from the previous step.

In this step of the algorithm, N is also a variable. This is because we have tried to solve the deconvolution with a different number of proposed compounds. Hence, in order to evaluate objectively the identity of the analytes within the deconvolved data, we will calculate the coefficient $\rho_{\tau,i}$ from the Eq. 15 for all the $i$ compounds $i = 1, \dots N$ (i.e. for all the centroids $\boldsymbol{c}_{\boldsymbol{i}}$ from the Eq. 10) and for each value of proposed $N$, $N = 1, .., \mathcal{H}$. In this work, we have limited the maximum number of compounds to 12, hence $\mathcal{H}$ =12. Note that the models are not nested: for each of the values of N (i.e. supposing a different number of clusters in the RT-PW space), we may obtain a different map of the $\boldsymbol{c}_{\boldsymbol{i}}$ values, and therefore a different collection of retrieved $\boldsymbol{A}_{i,*}$ spectra (with $i = 1,2, \dots N$). We make this explicit in the definition of $\rho_{\tau,i,N}$.

There are two ways to explore the results given by the correlation: one is to find the $\tau$ element in the database that yields the maximum value of the correlation coefficient, $\max(\rho_{\tau,i,N})$ (referred onwards as "max ranked") and hence the identity of the identified compound in the database is:

$$\tau_{i,N}{}' = \underset{\tau}{argmax}(\rho_{\tau,i,N}) \qquad 16$$

Where the $i$ index means that this "max rank" is calculated for each of the N components ($i = \{1,2, \dots, N\}$) of the solution provided by equation 15. As explained earlier, the models are not nested. This means that this search performed also for all i elements for a variable N ($N = 1,..,\mathcal{H}$). Further, we define $\rho_{i,N,MAX} = \max_{\tau}(\rho_{\tau,i,N})$, i.e. the value of the maximum correlation found for all $\tau = 1,2, \dots, T$ compounds in the database for a given i and a given N. The list of ${\tau_{i,N}}'$ constitute an enumeration of the possible compounds from the database that could be present in the sample.

We evaluate the strength of the evidence of the presence of each of the ${\tau_{i,N}}'$ elements as follows. This evidence is defined as $\bar{\rho}'_{\tau}$. First, all $\tau$ elements in the database that are not listed in ${\tau_{i,N}}'$ (i.e. they were never yielding the maximum correlation) receive a value of $\bar{\rho}'_{\tau}$ of 0. For the rest of the $\tau$ values, we make use of the value found in $\rho_{i,N,MAX}$ . In this way we can define $\rho_{\tau,i,N,MAX}$ as follows:

$$\rho_{\tau,i,N,MAX} = \begin{cases} \rho_{i,N,MAX} for\ \tau = {\tau_{i,N}}' \\ 0 \qquad\qquad otherwise \end{cases} \tag{17}$$

Next, we make use of $p(N|BIC)$ as a probabilistic weight:

$$\bar{\rho}'_{\tau} = \sum_{N=1}^{\mathcal{H}} \left[ \left[ p(N|BIC) \frac{1}{N} \right] \sum_{i=1}^{N} \rho_{\tau,i,N,MAX} \right] \tag{18}$$

A second way is to interpret the values of the correlation from the Eq. 15 and use all values of $\rho_{\tau,i,N}$ opposed to consider only the maximum and assign a zero value to the others. In other words, we do not exclude other possibilities regardless of the magnitude of the correlation value. The reasoning behind this approach is the fact that the ground truth may be ranked the second or lower, with a very small or insignificant difference from the $\rho_{\tau,i,N_{MAX}}$. This method is called "all ranked" method. In this case, a similar computation to Eq. 18 is used, and the equation becomes:

$$\bar{\rho}_{\tau} = \sum_{N=1}^{\mathcal{H}} \left[ \left[ p(N|BIC) \frac{1}{N} \right] \sum_{i=1}^{N} \rho_{\tau,i,N} \right] \tag{19}$$

It is easy to identify in the last two equations a discreet form of integration over the centroids (i.e. over the retention times obtained with MM step) which essentially means that we lose the information about the retention times and the peak width of the identified compounds. The information about retention time of each identified compound (i.e. for each $\tau$) can be obtained when using "max ranked" approach as

in this case there is a link between the $\tau_i$, $c_i$ and $p(N|BIC)$. Hence, for one fixed $\tau' = \Theta$ (with $\Theta \geq 1$ and $\Theta \leq$ T) we can estimate a retention time using the median value of the centroids associated to this $\tau'$:

$$t_{r_{\tau=\Theta}} = median\left(c_{i_{\tau'=\Theta}}\right) \quad 20$$

## 1.4 Interpolation algorithm

Let $S$ be the spectra at the id $\tau$ in the database, $S = [mz_1, I_1;\ mz_2, I_2;\ \ldots mz_i, I_i; \ldots\ mz_I, I_I]$. Let $a$ be the vector of data at one retention time, $a = [mz_1, I_1;\ mz_2, I_2;\ \ldots mz_j, I_j; \ldots\ mz_J, I_J]$. And let $\sigma_s$ be the standard deviation of the mass-to-charge values of the $S$ calculated as follows.

$$\Delta_s = median(mz_i - mz_{i-1})$$

The interpolated spectra S' is obtained as follows:

$$S'_j = \sum_{i=1}^{I} I_i e^{\left(-\frac{(mz_i - mz_j)^2}{2\Delta_s}\right)}$$

# 2. Results and discussion

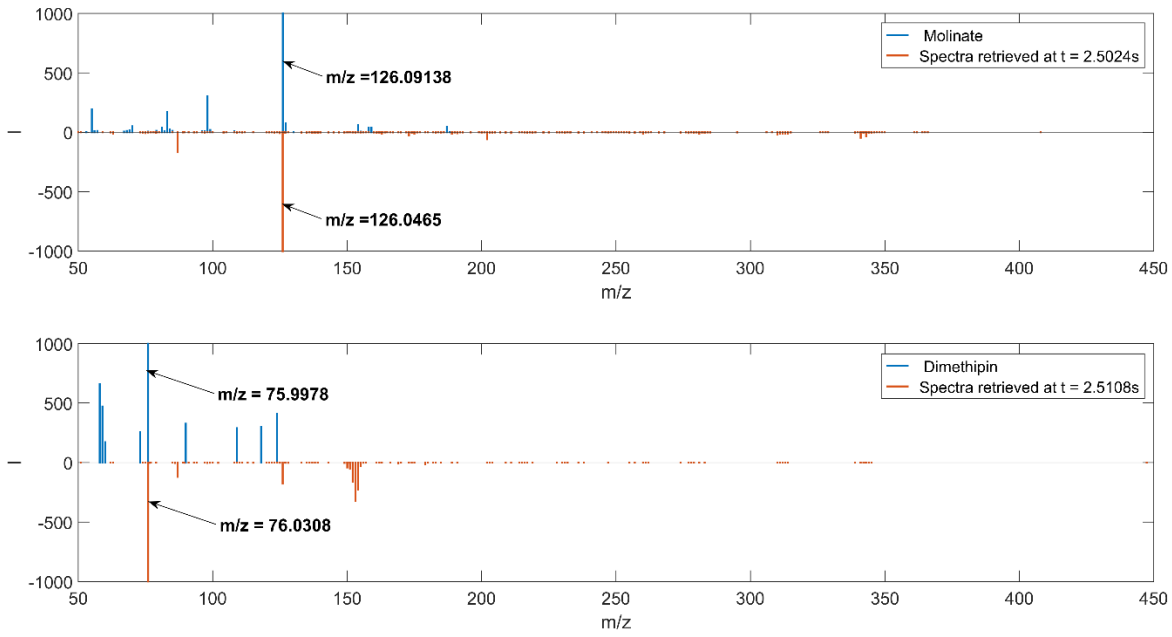



*Figure SI.1 Retrieved spectra at 2.5024s and 2.5108 with corresponding identified false positives for the Case 1*

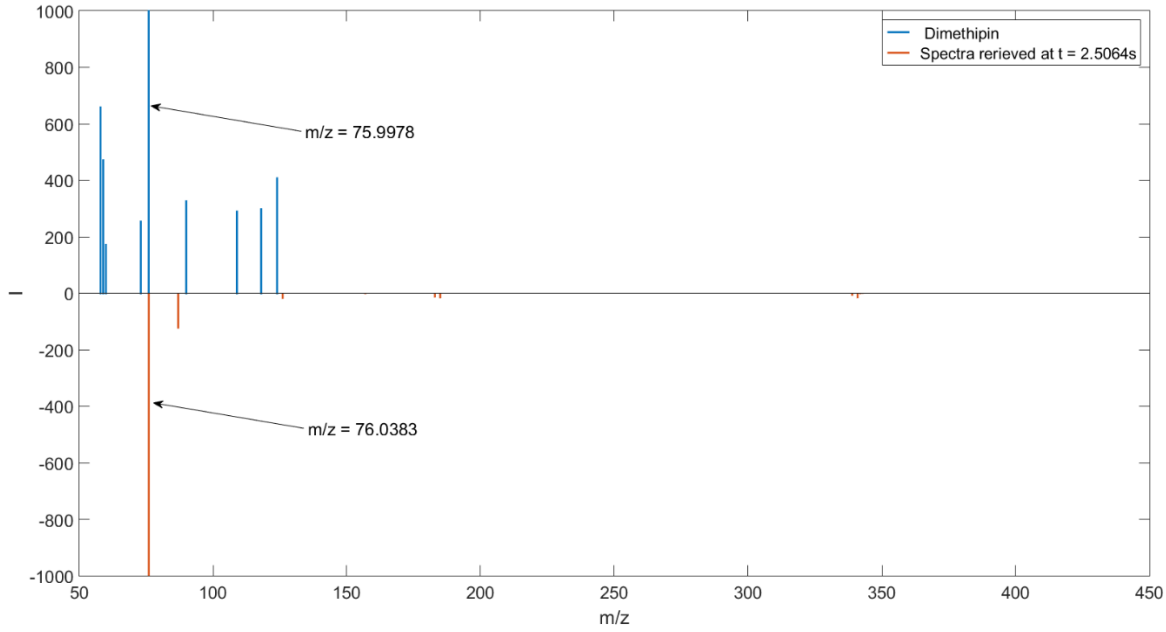



*Figure SI.2 Retrieved spectra at 2.5064s with corresponding identified false positive for the Case 2*

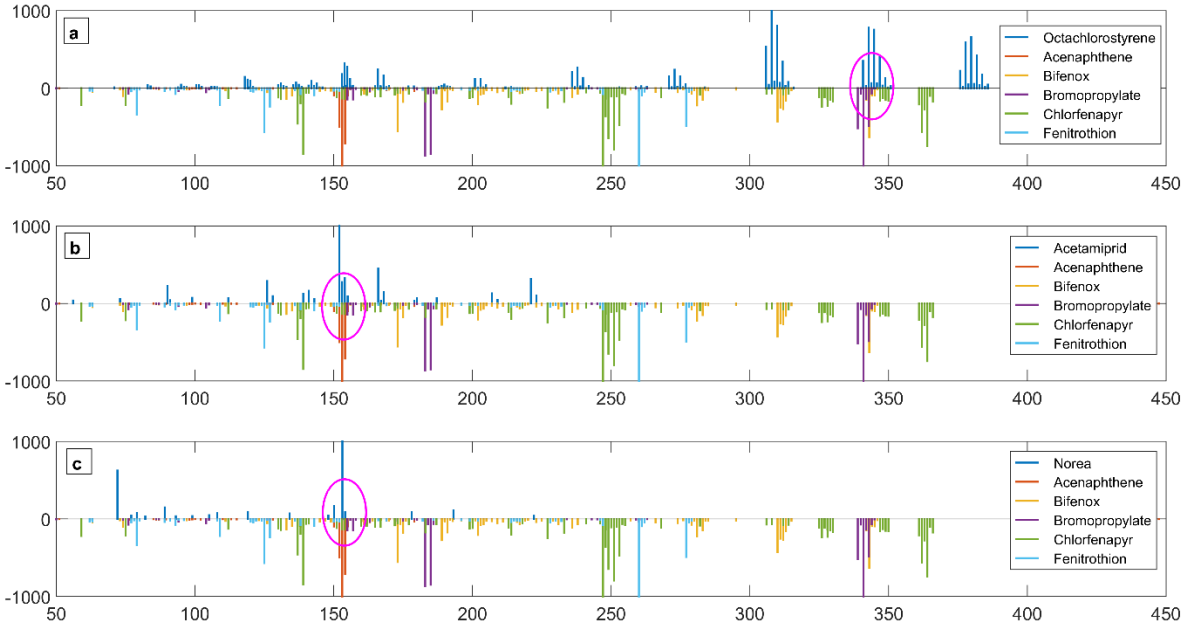



*Figure SI.3 (a) - Spectrum of Octachlorostyrene ($\tau = 92$), (b) - Spectrum of Acetamiprid ($\tau = 441$), (c) - Spectrum of Norea ($\tau = 447$) compared with the spectra of the compound used for the simulation. The magenta ellipse points to the mass channels that can cause high correlation.*

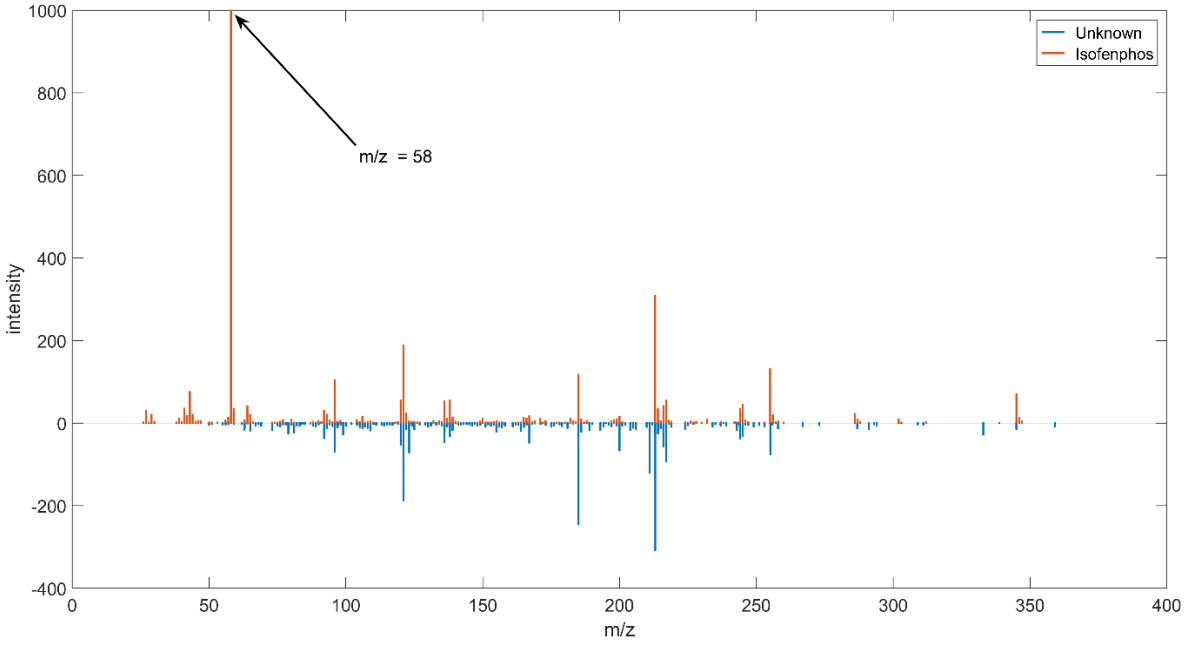



Figure SI.4 Spectra of the deconvolved compound (negative) and Isofenphos (positive).

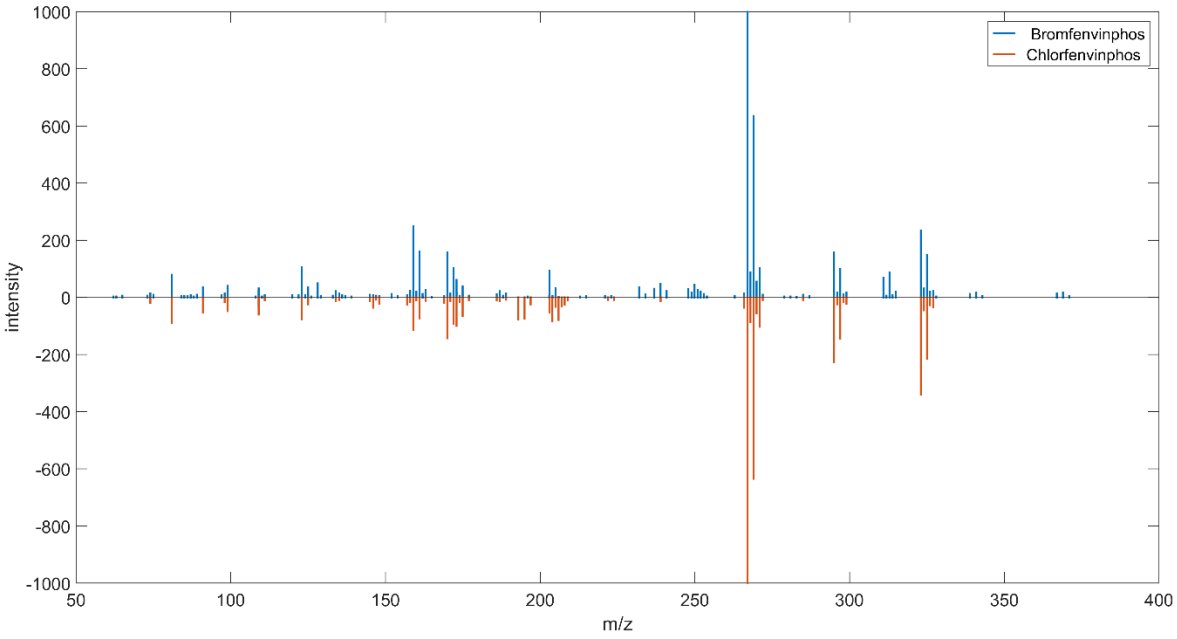



Figure SI.5 Spectrum of Bromfenvinphos and Chlorfenvinphos.

## References


[1] H. G. Mol, T. M. and P. Zomer, “Evaluation of gas chromatography – electron ionization – full scan high resolution Orbitrap mass spectrometry for pesticide residue analysis,” *Anal. Chim. Acta,* vol. 935, pp. 161-172, 2016.

[2] R. Tauler, “Multivariate curve resolution applied to second order data,” *Chemometr. Intell. Lab,* vol. 30, no. 1, pp. 133-146, 1995.

[3] R. Bro, “PARAFAC. Tutorial and applications,” *Chemometr. Intell. Lab,* vol. 38, pp. 149-171, 1997.

[4] X. Du and S. H. Zeisel, “Spectral Deconvolution for Gas Chromatography Mass Spectrometry-Based Metabolomics: Current Status and Future Perspectives,” *Comput. Struct. Biotechnol. J.,* vol. 4, no. 5, pp. 1-10, 2013.

[5] N. Pasadakis, V. Gaganis and P. Smaragdis, “Independent Component Analysis (ICA) in the Deconvolution of Overlapping HPLC Aromatic Peaks of Oil,” Mitsubishi Electric Research Laboratories, Inc., Cambridge, Massachusetts, 2003.

[6] M. Wasim and R. G. Brereton, “Determination of the number of significant components in liquid chromatography nuclear magnetic resonance spectroscopy,” *Chemometr. Intell. Lab,* vol. 72, pp. 133-151, 2004.

[7] G. Vivó-Truyols, J. Torres-Lapasió, M. García-Alvarez-Coque and P. Schoenmakers, “Towards unsupervised analysis of second-order chromatographic data: Automated selection of number of components in multivariate curve-resolution methods,” *J Chromatogr A.,* vol. 1158, no. 1-2, pp. 258-272, 2007.

[8] S. Peters, H.-G. Janssen and G. Vivó-Truyols, “A new method for the automated selection of the number of components for deconvolving overlapping chromatographic peaks.,” *Anal. Chim. Acta,* vol. 799, pp. 29-35, 2013.

[9] B. D. Fitz and R. E. Synovec, “Extension of the two-dimensional mass channel cluster plot method to fast separations utilizing low thermal mass gas chromatography with time-of-flight mass spectrometry,” *Anal. Chim. Acta,* vol. 913, pp. 160-170, 2016.

[10] B. Fitz, B. Reaser, D. Pinkerton, J. Hoggard, K. J. Skogerboe and R. E. Synovec, “Enhancing Gas Chromatography–Time of Flight Mass Spectrometry Data Analysis Using Two-Dimensional Mass Channel Cluster Plots,” *Anal Chem,* vol. 86, no. 8, pp. 3973-3979, 2014.

[11] D. Sivia and J. Skilling, Data Analysis A Bayesian Tutorial, Oxford: Oxford University Press, 2006.

[12] P. Gregory, Bayesian Logical Data Analysis for the Physical Sciences, Cambridge: Cambridge University Press, 2005.

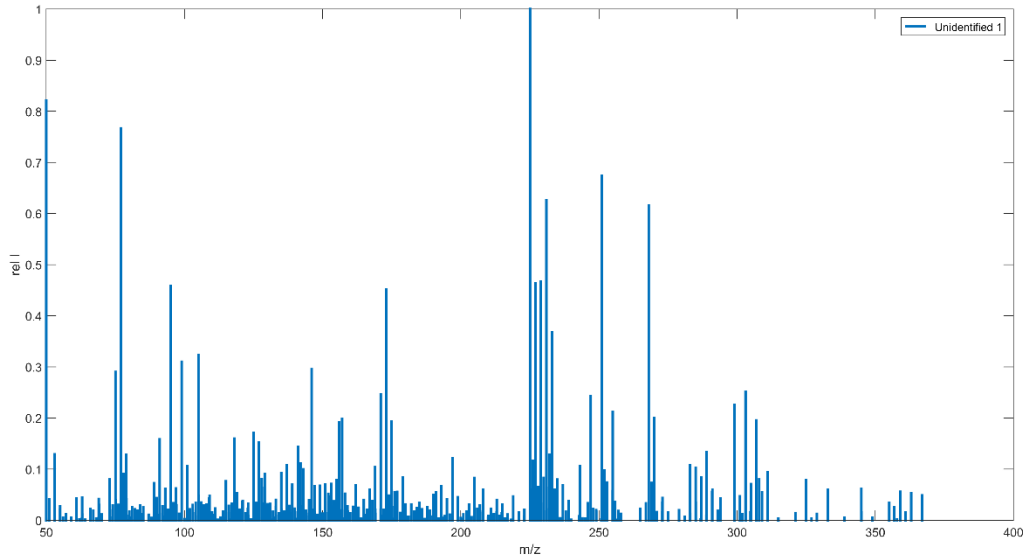
[13] P. Breheny, “The Estimation of Prediction Error: Covariance Penalties and Cross-Validation,” *J Am Stat Assoc,* vol. 99, no. 467, pp. 619-642, 2004.

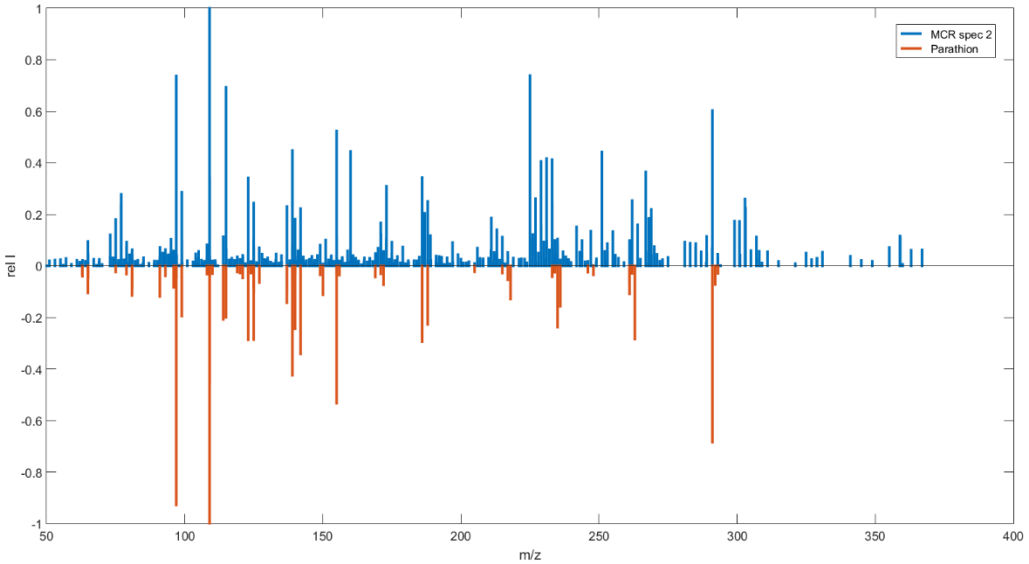
[14] T. Hastie, R. Tibshirani and J. Friedman, Elements of statistical Learning. Data Mining, Inference, and Prediction, New York: Springer, 2009.

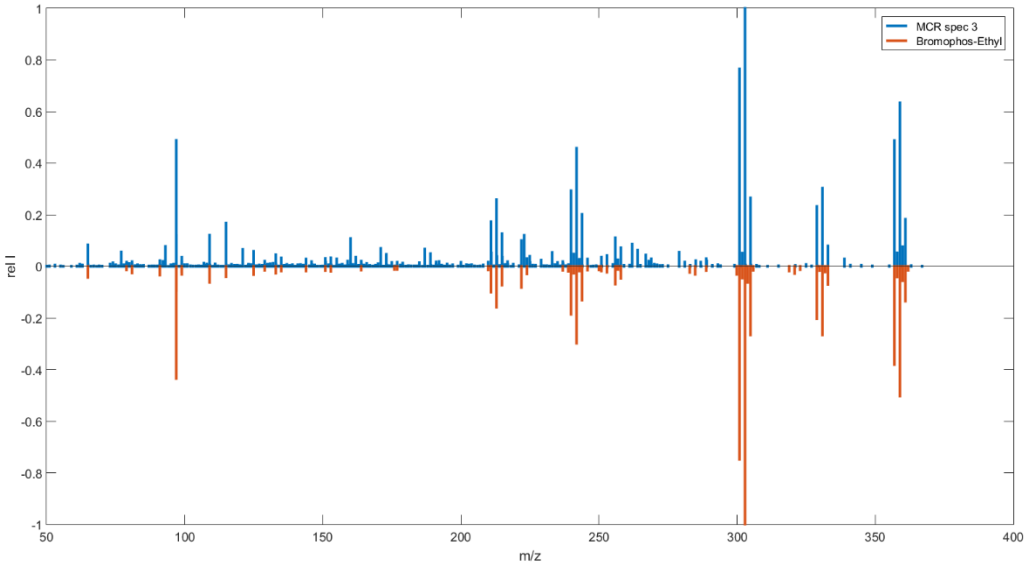
[15] C. Bishop, Pattern Recognition and Machine Learning, New york: Springer, 2006.

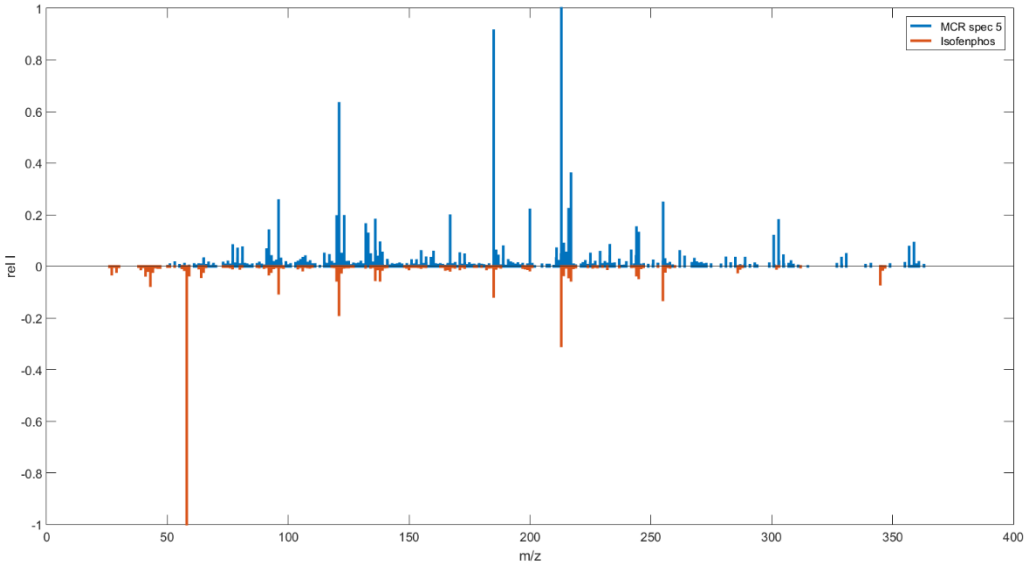
# Appendix A

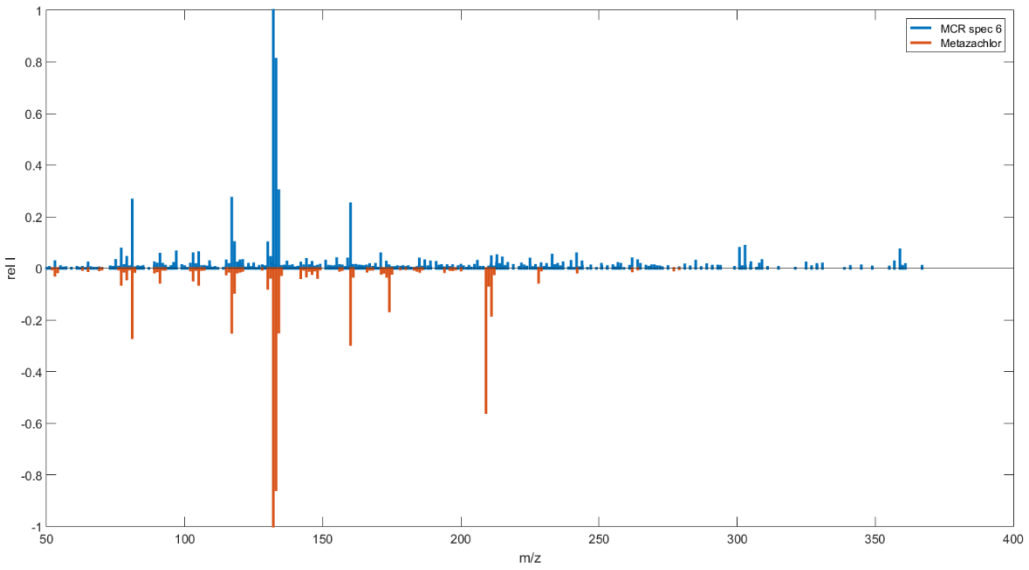
MCR output

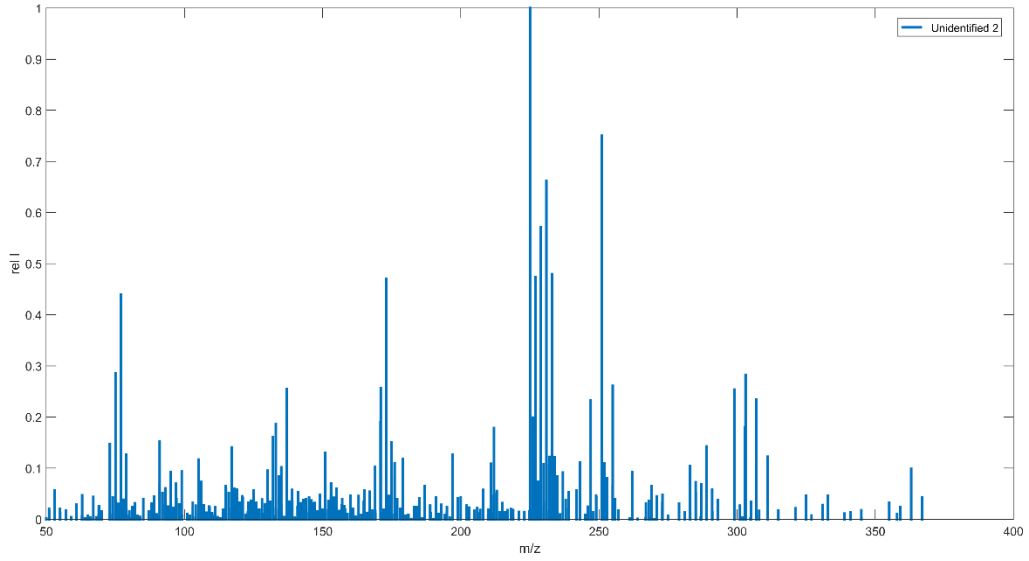


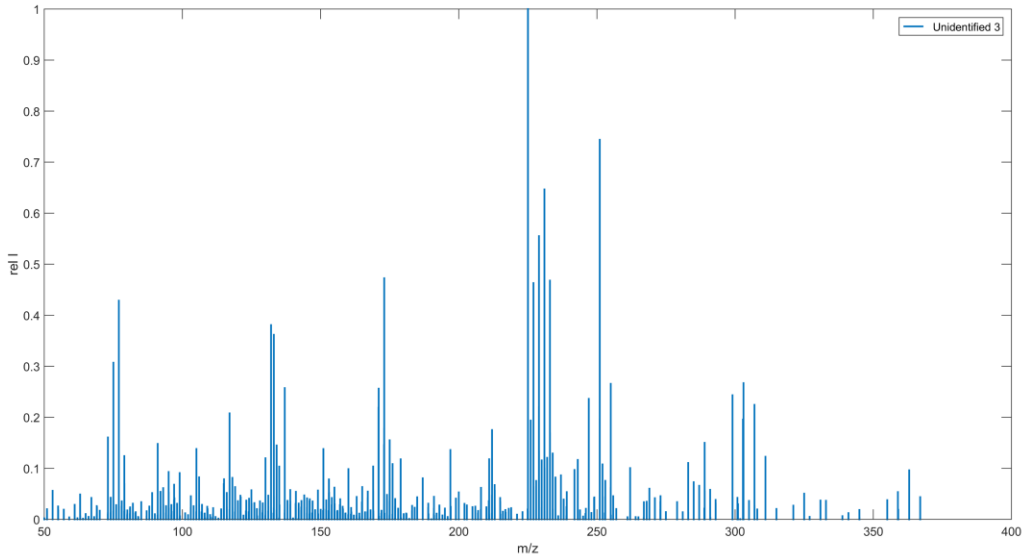




MCR spec 3
Bromophos-Ethyl
rel I
m/z


MCR spec 5
Isofenphos
rel I
m/z


MCR spec 6
Metazachlor
rel I
m/z
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
50
100
150
200
250
300
350
400


Unidentified 2
rel I
m/z
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
50
100
150
200
250
300
350
400




*Fitting error (lack of fit, lof) in % at the optimum = 4.1541(PCA) 6.0678(exp)*

*Percent of variance explained (r2)at the optimum is 99.6318*

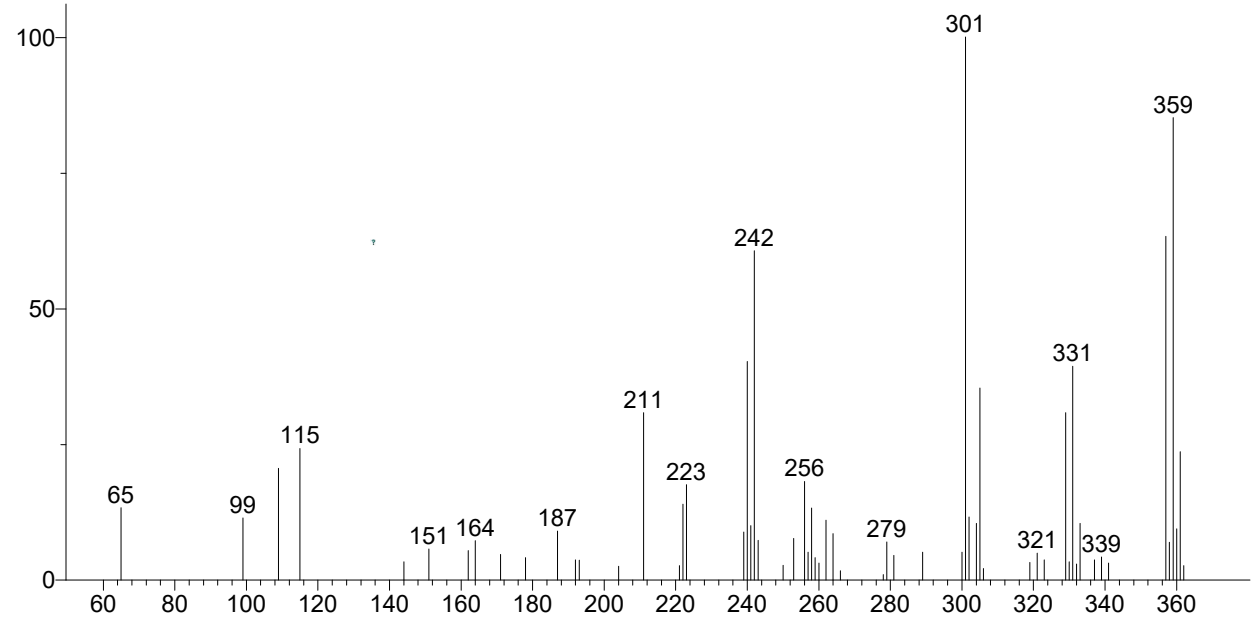
*Relative species conc. areas respect matrix (sample) 1 at the optimum*

*Plots are at optimum in the iteration 5*

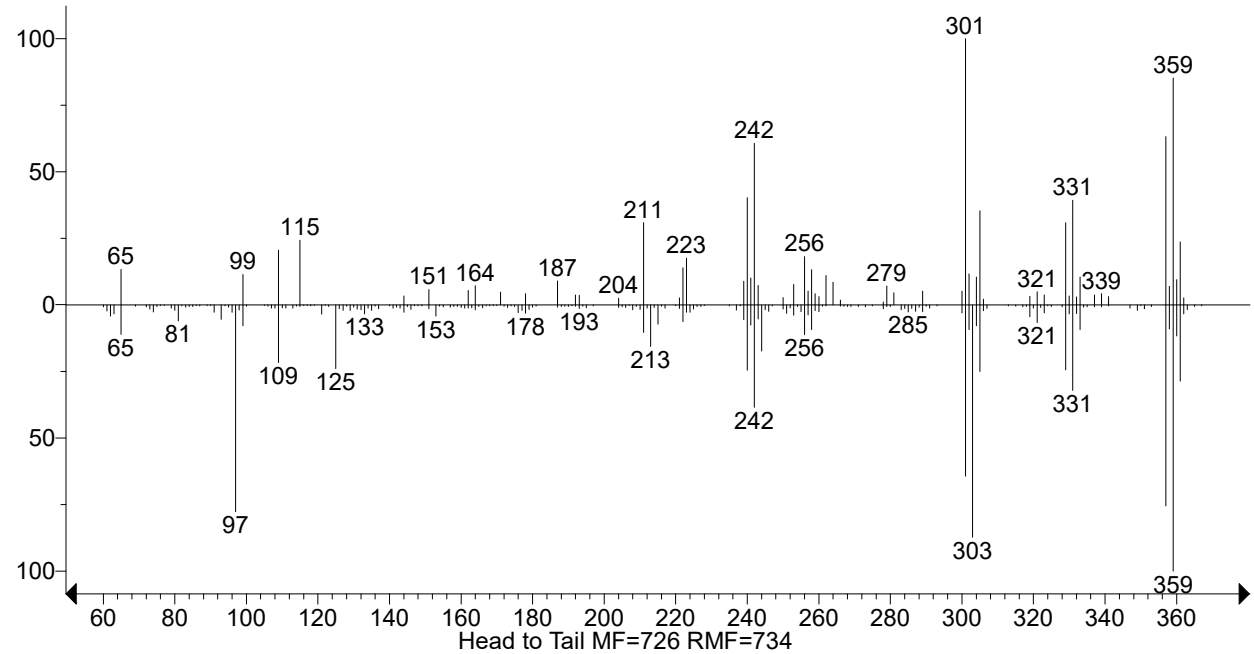
*Note: the compounds that were identified incorrectly (false positives) are labeled here with "Unidentified"*

Appendix B

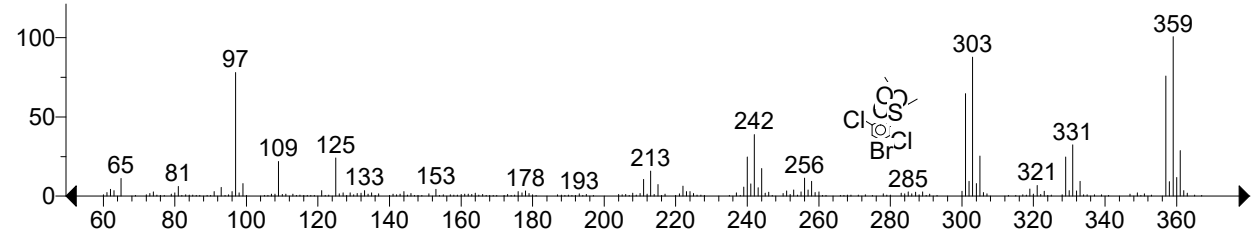
AMDIS deconvolution



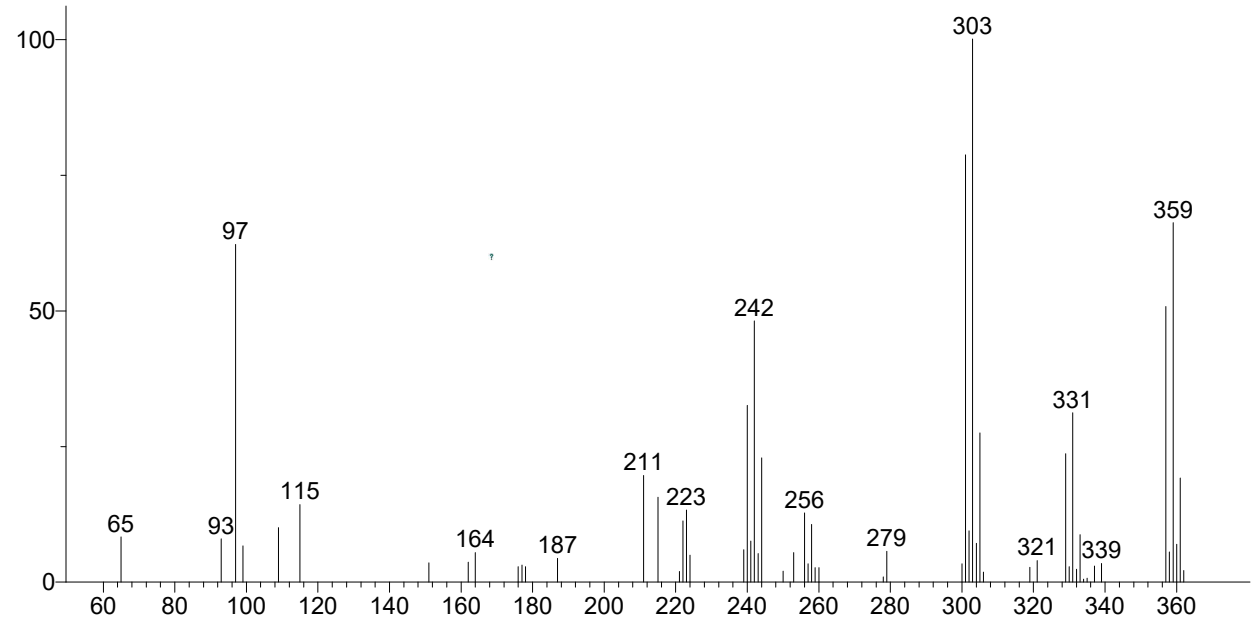

(Text File) Component at scan 2667 (16.354 min) [Model = +262u] in d:\codebank\phd4_deconvolution\160303_003l.cdf



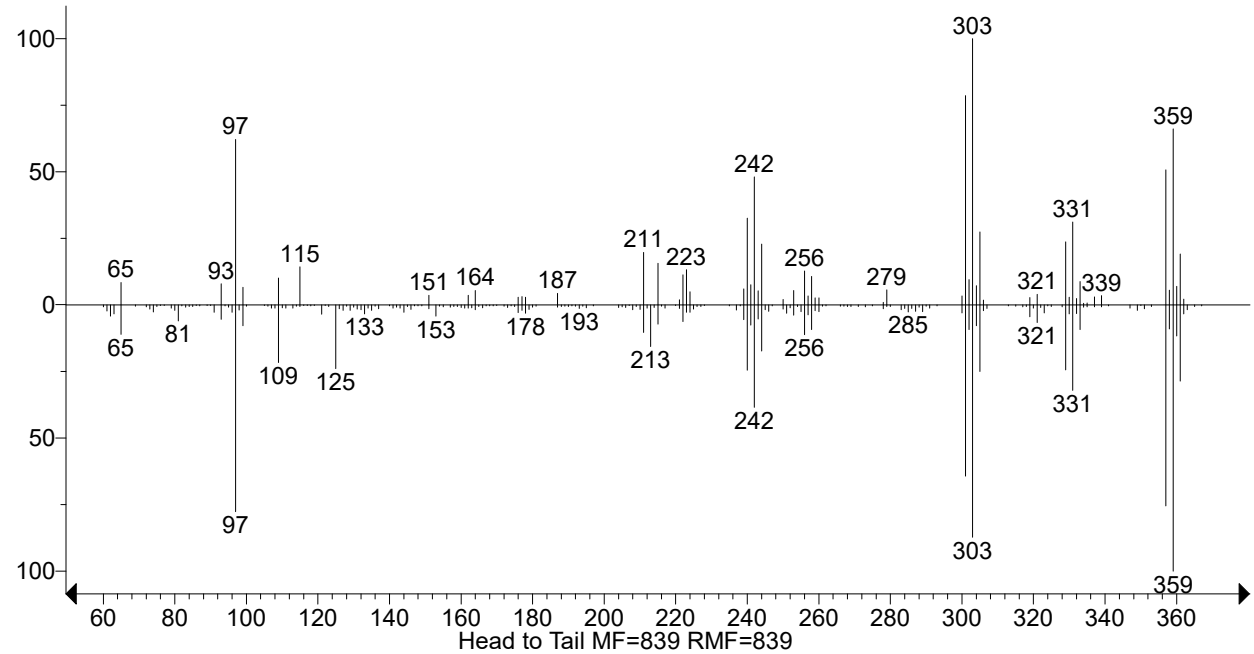

Head to Tail MF=726 RMF=734



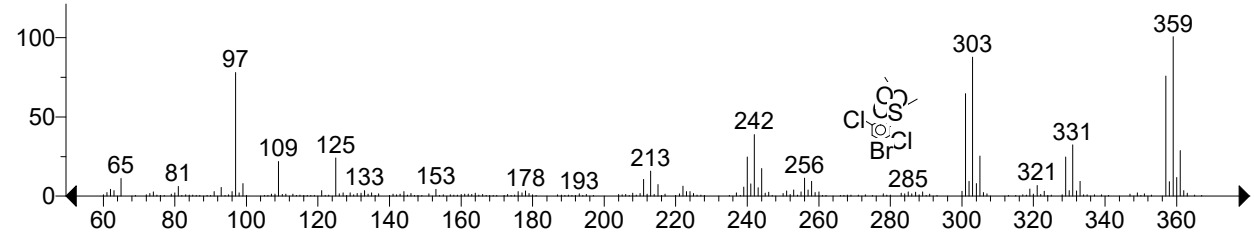

(replib) Bromophos-ethyl



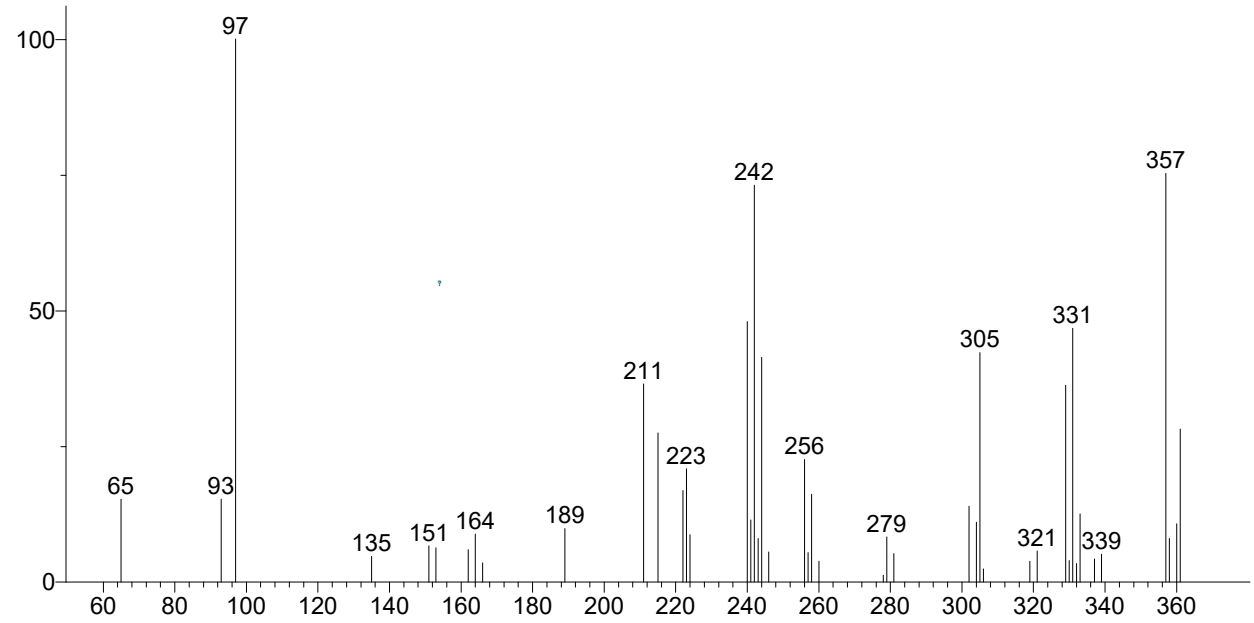

(Text File) Component at scan 2668 (16.364 min) [Model = +303u] in d:\codebank\phd4_deconvolution\160303_003l.cdf



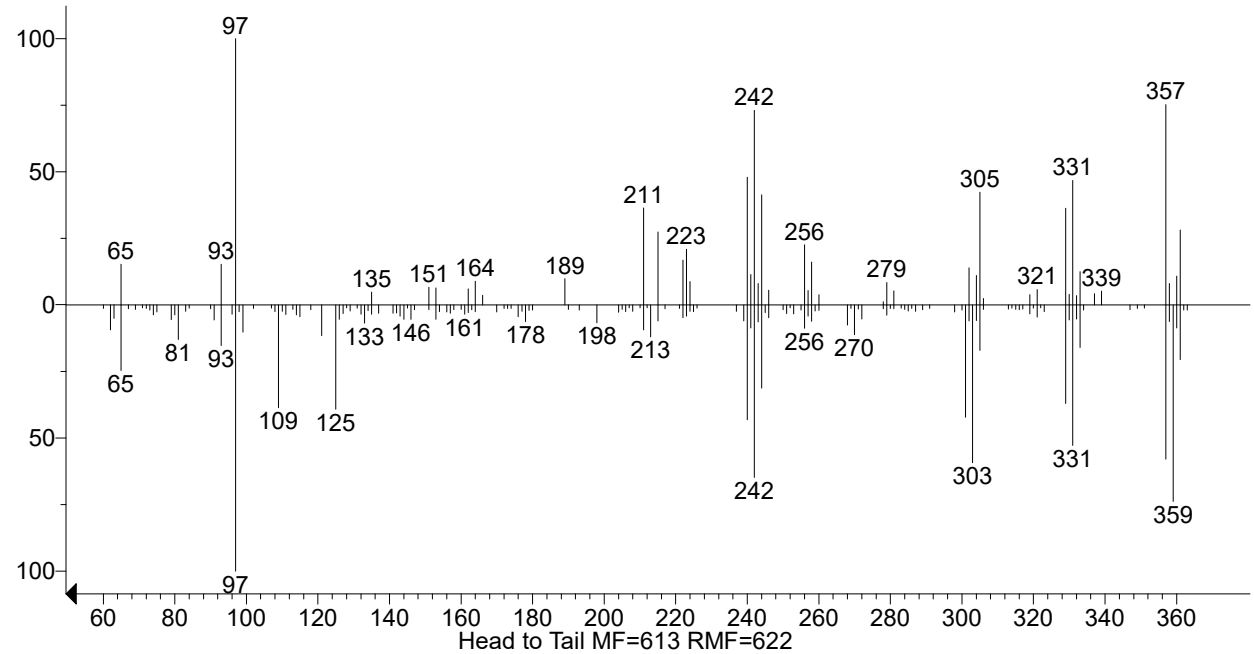

Head to Tail MF=839 RMF=839



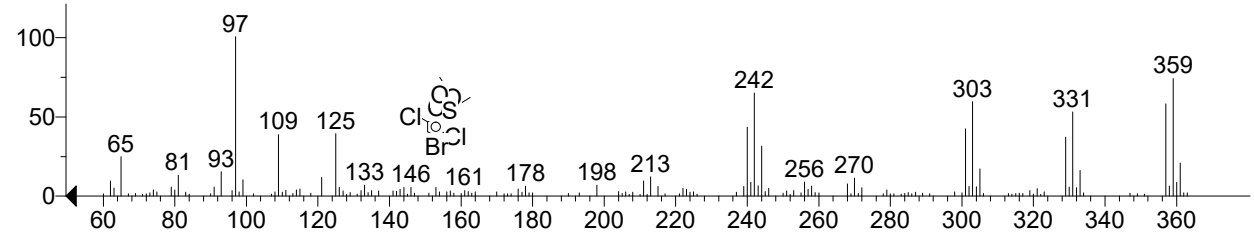

(replib) Bromophos-ethyl



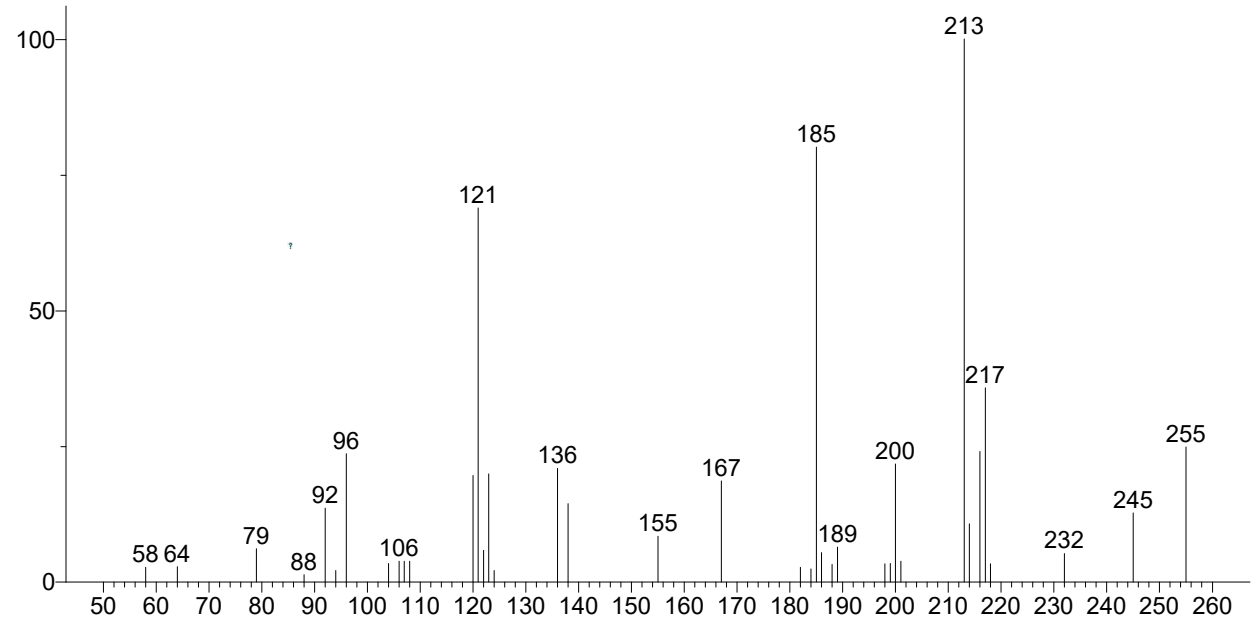

(Text File) Component at scan 2669 (16.368 min) [Model = TIC] in d:\codebank\phd4_deconvolution\160303_003I.cdf



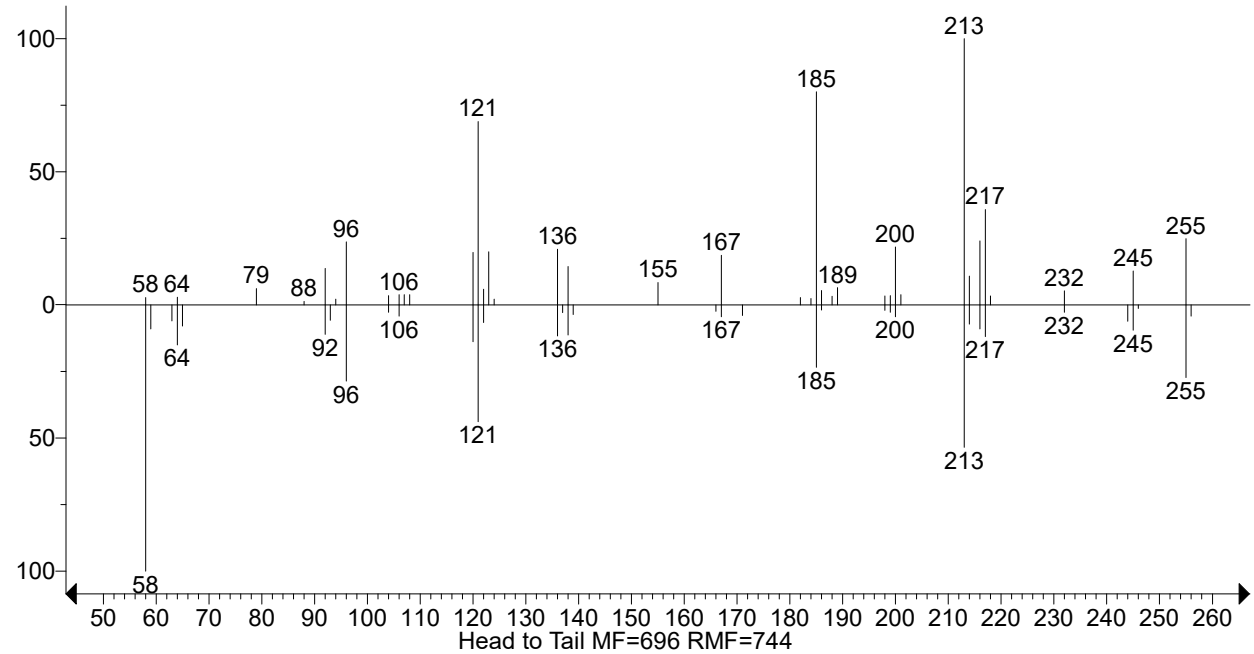

Head to Tail MF=613 RMF=622



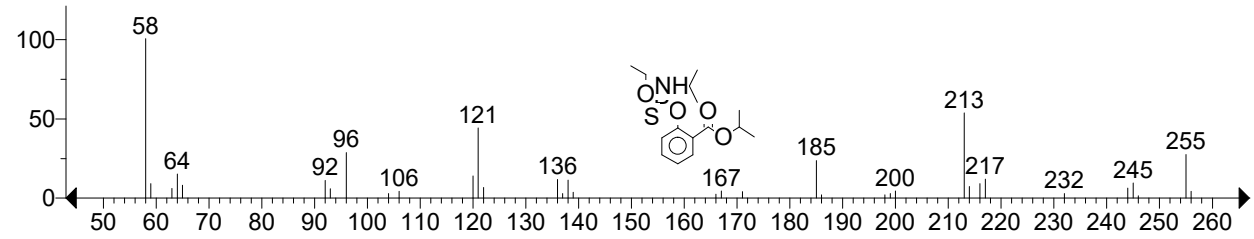

(mainlib) Bromophos-ethyl



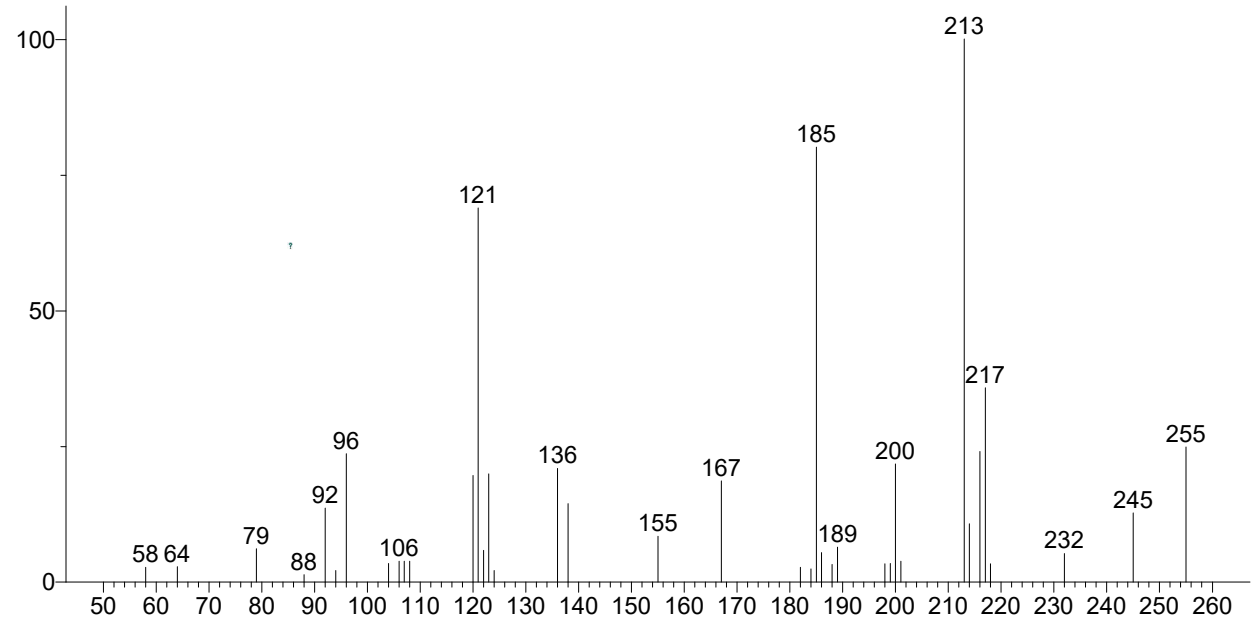

(Text File) Component at scan 2671 (16.375 min) [Model = +213u, -303u] in d:\codebank\phd4_deconvolution\160303_003l.cdf



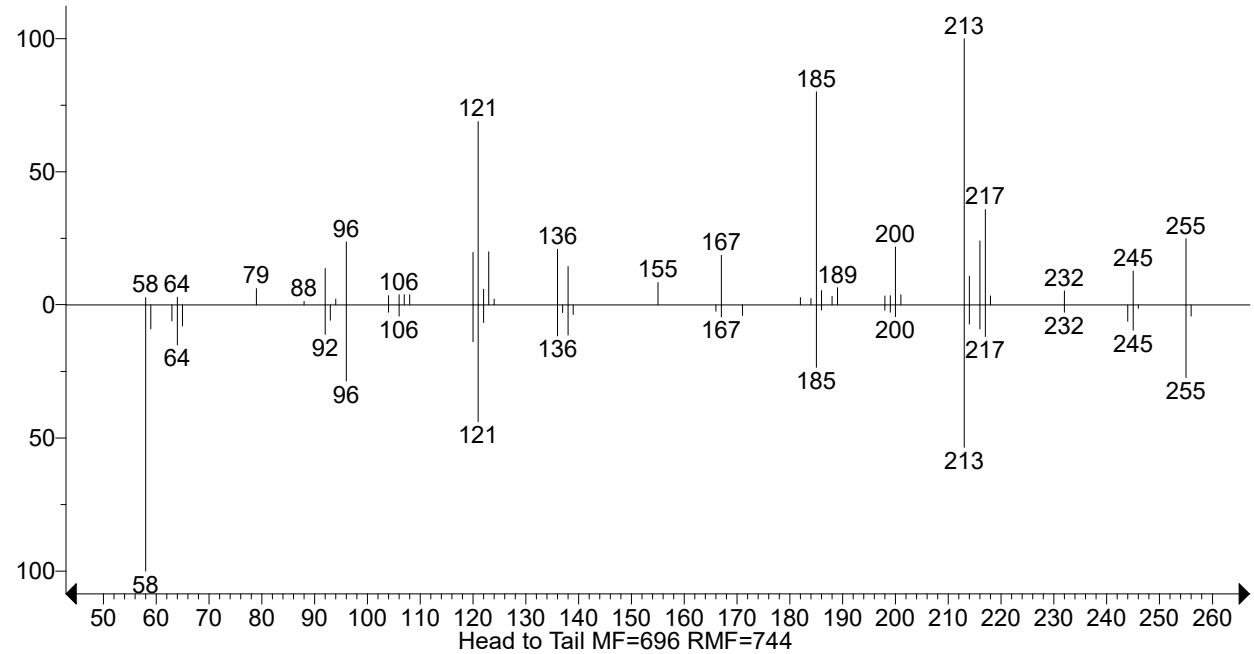

Head to Tail MF=696 RMF=744



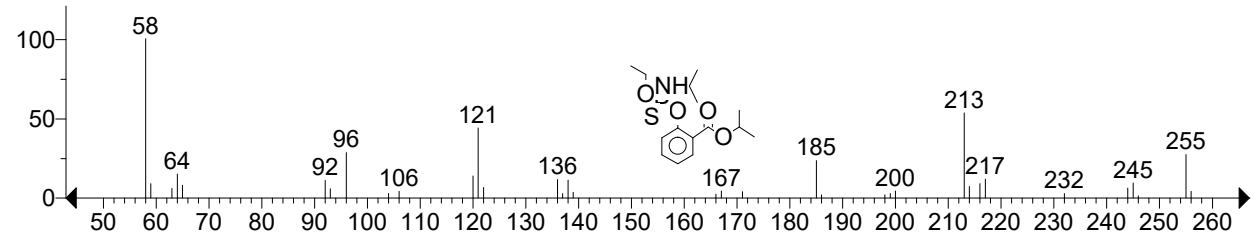

(replib) Benzoic acid, 2-[[ethoxy[(1-methylethyl)amino]phosphinothioyl]oxy]-, 1-methylethyl ester



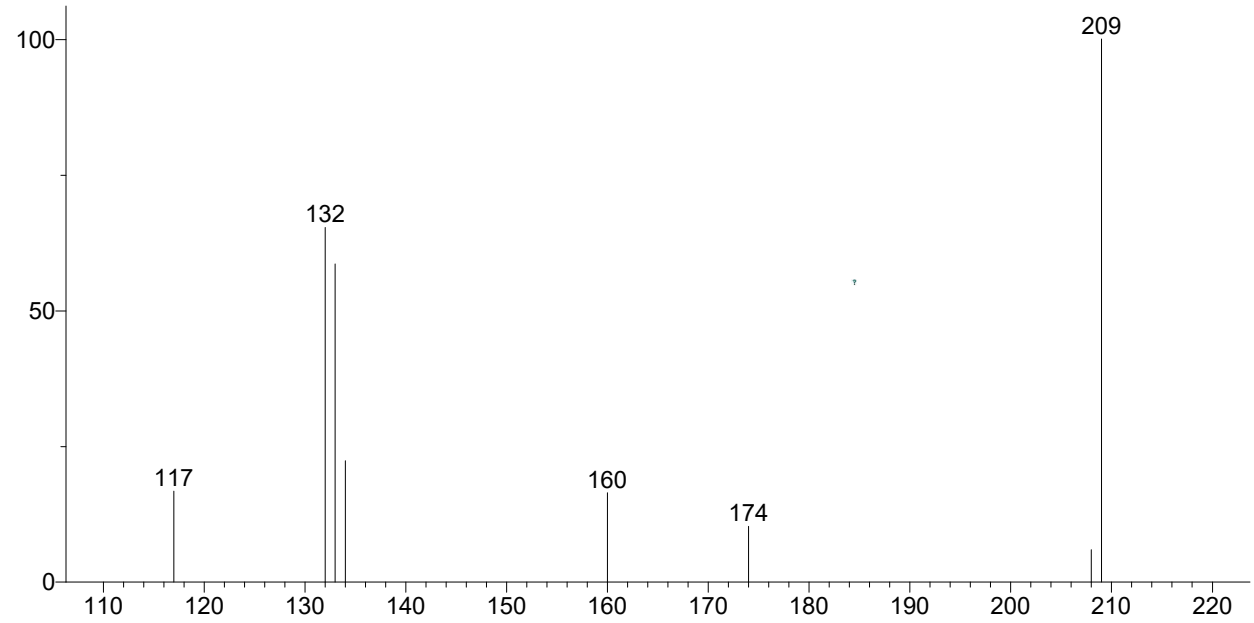

(Text File) Component at scan 2671 (16.375 min) [Model = +213u, -303u] in d:\codebank\phd4_deconvolution\160303_003l.cdf



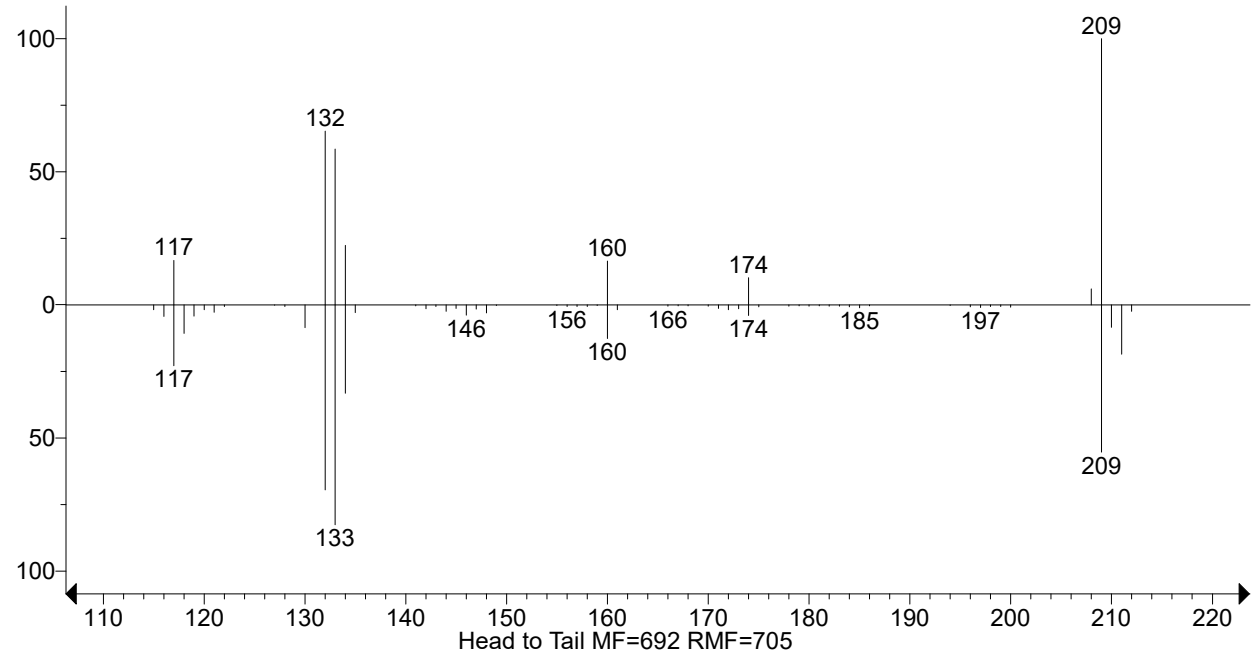

Head to Tail MF=696 RMF=744



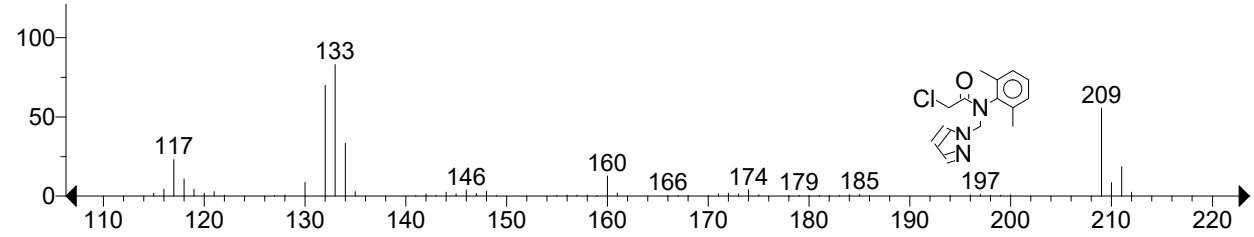

(replib) Benzoic acid, 2-[[ethoxy[(1-methylethyl)amino]phosphinothioyl]oxy]-, 1-methylethyl ester




(Text File) Component at scan 2677 (16.398 min) [Model = +209u] in d:\codebank\phd4_deconvolution\160303_003l.cdf




Head to Tail MF=692 RMF=705




(mainlib) Acetamide, 2-chloro-N-(2,6-dimethylphenyl)-N-(1H-pyrazol-1-ylmethyl)-