Jul 01, 2024

Tutorial on PARADISe: PARAFAC2-based Deconvolution and Identification System for processing GC–MS data

This protocol is a draft, published without a DOI.

Beatriz Quintanilla-Casas[1], Rasmus Bro[1], Jesper Løve Hinrich[1], Cleo L. Davie-Martin[1]

[1]University of Copenhagen

Beatriz Quintanilla-Casas
University of Copenhagen

Protocol Citation: Beatriz Quintanilla-Casas, Rasmus Bro, Jesper Løve Hinrich, Cleo L. Davie-Martin 2024. Tutorial on PARADISe: PARAFAC2-based Deconvolution and Identification System for processing GC–MS data. **protocols.io** **https://protocols.io/view/tutorial-on-paradise-parafac2-based-deconvolution-dfda3i2e**

Protocol status: Working
We use this protocol and it's working

Created: June 10, 2024

Last Modified: July 01, 2024

Protocol Integer ID: 101506

Keywords: GC-MS, Deconvolution, PARAFAC2, Untargeted, Deep learning

Abstract

The present protocol provides general guidelines for users working with PARADISe, a deconvolution and identification system for processing GC-MS data. This tool allows users to perform untargeted analysis of large datasets efficiently and minimizes inter-user variability. The final output is a peak table, in excel format, containing the peak area and the purified spectrum for each resolved compound.

Attachments

Tutorial on PARADISe...

2.6MB

Guidelines

Introduction:

Traditionally, GC-MS data have been handled through targeted analyses, which involve the identification and quantification of predefined compounds. Such a strategy, besides being time consuming, is often subject to inter-person variability (Ballin & Laursen, 2019). Furthermore, it usually ignores undiscovered compounds, even if the raw analytical data provides comprehensive information.

In contrast, untargeted methods are gaining importance and provide approaches that overcome many of the problems associated with targeted approaches. Various tools have emerged over the past decades to achieve efficient, untargeted profiling analyses of GC-MS data. Some of these tools were based on the analysis of single chromatograms or a batch of sample chromatograms, and with both open accessible e.g., XCMS (Smith, et al., 2006) and proprietary toolboxes (i.e., MassHunter®, Agilent™). Yet, most methods are still time-consuming and require users to make decisions on settings that can lead to biases.

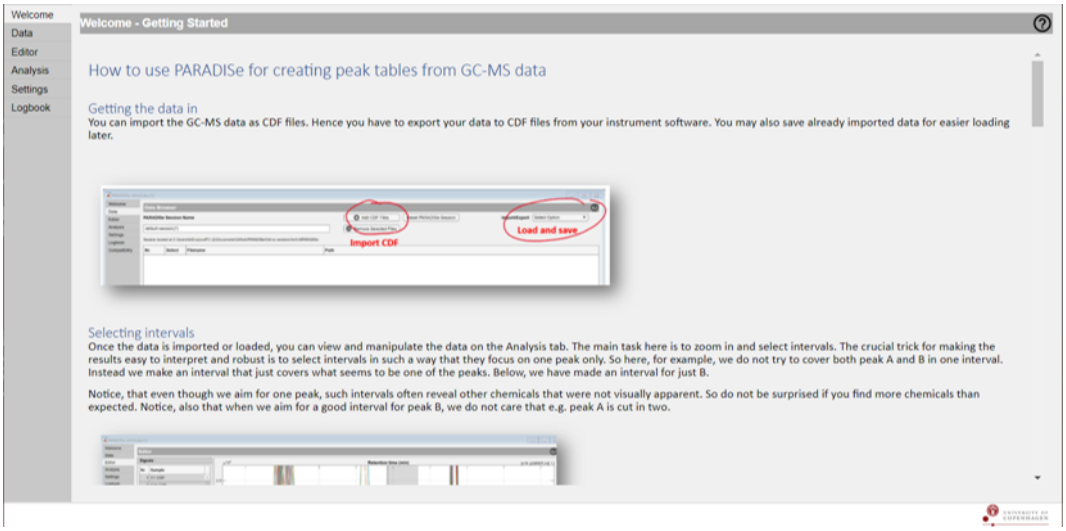
The present tool, PARADISe, is a PARAllel FACtor analysis 2 (PARAFAC2)-based deconvolution and identification system that applies a tensor decomposition model on the 3-way array considering all samples. It uses collinearity of the mass spectral mode in relation to the concentration of co-eluting peaks. One of the main advantages of the PARAFAC2 model is that it allows for retention time shifting between samples, which are commonly found in chromatographic data. Nonetheless, alignment algorithms can be automatically applied to GC-MS data in PARADISe to ease the confusion associated with interval selection when severe retention time shifts are present. The performance of PARAFAC2 models will be assured, as long as the user provides a large dataset (at least 20 samples) and the different chemical compounds present different individual mass spectra compared to each other, especially when they elute at very close retention times. The final output that PARADISe generates is a peak table of peak areas for each of the selected compounds. Moreover, if the user has access to a NIST MS database, the software will also identify the chemical compounds found.

> Note
>
> When samples are highly variable and/or the chemical profiles are very complex, a greater number of samples will be needed.

This protocol aims to provide an overall explanation of the PARADISe software (v6.0.1), from the basis of the methods to practical issues. The protocol will be updated when significant changes to the PARADISe workflow are effected. The PARADISe software adapts a semantic versioning scheme (per PARADISe v6.0.1), such that version number X.Y.Z denotes the major version X,minor version Y, and patch version Z (bug fixes and improvements). Thus, analysis perform in earlier versions of PARADISe will be loadable in newer versions, as long as the major version X is the same.

PARADISe tabs:

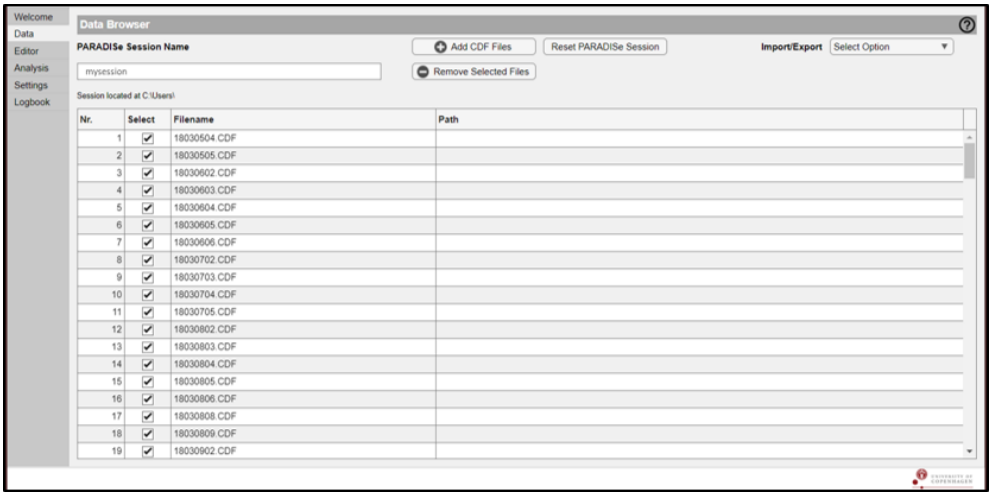
1. The **Welcome tab** (given in below figure) provides simple tips on getting started.



Welcome tab overview

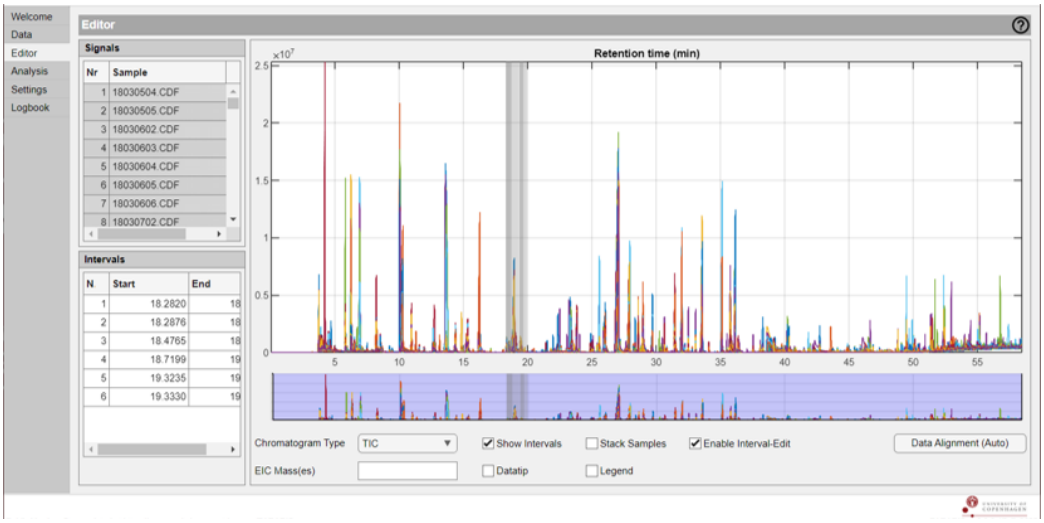
Furthermore, there is a series of short videos with extended information about the software and its utilities on our YouTube channel “Chemometrics and Machine Learning in Copenhagen” (https://www.youtube.com/@QualityAndTechnology), within the “PARADISe version 6” playlist.

2. The **Data tab** (given in below figure) is where the user reads in data, either by importing CDF files or previous PARADISe sessions. The Data tab also has the option to “Export” PARADISe intervals and chromatography data in the top right-hand corner, but this is not necessary as the current PARADISe session (and models) are automatically saved to the Session Name folder specified.



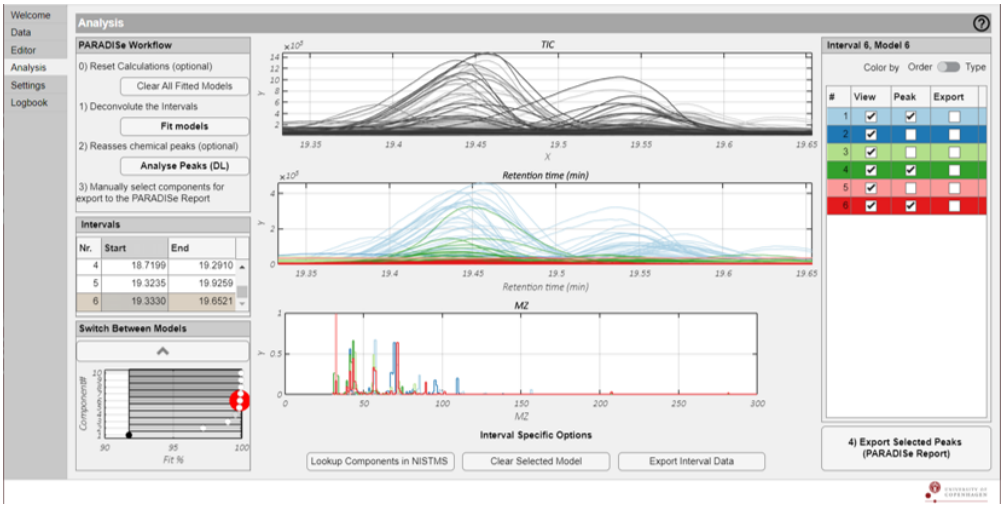
Data tab overview

3. The **Editor tab** (given in the below figure) is where the user can navigate through the overlaid chromatograms and select intervals. Different settings for data visualization are possible e.g., TIC/BPC or even specific m/z (EIC).



Editor tab overview

4. Models for the selected intervals are fitted and evaluated in the **Analysis tab** (given in the below figure) according to the features set in the **Settings tab** (see below). Once models are fitted, the user explores them manually in order to decide the optimum number of components and to select which components correspond to chemical peaks, which will end up in the final peak table.

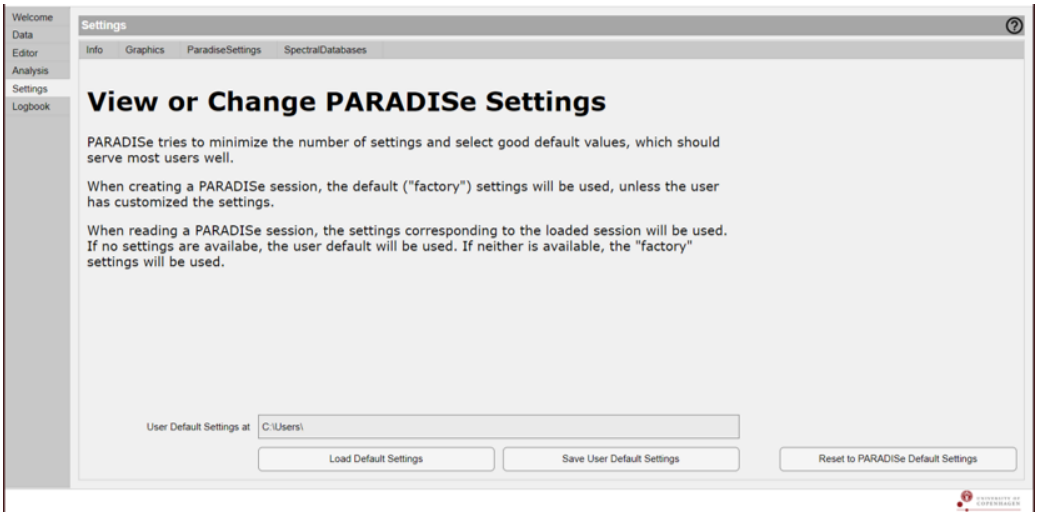


Analysis tab overview

5. In the **Settings tab** (below figure), overall PARADISe options are set. Usually, it is only the PARADISe session name, data save directory, and NIST directory where changes are required. However, modelling features can also be adjusted in this tab.

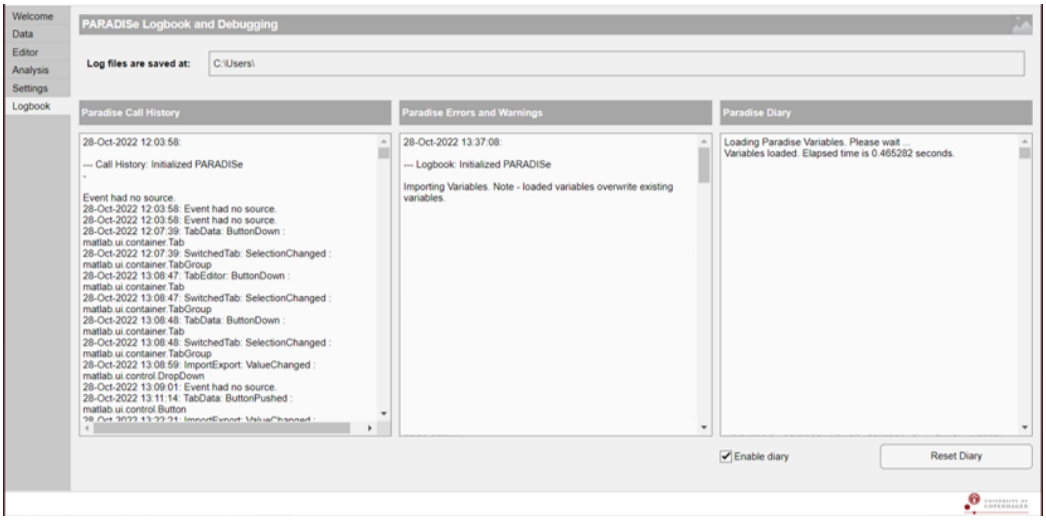
Note

Tip: The first time you run the program, it may be that no settings are shown. If that is the case, click the bottom “Reset to PARADISe Default Settings” under the Info tab.



Settings tab overview

6. The **Logbook tab** (Figure 6) contains a record of all “steps” taken in the PARADISe session and is saved automatically to the Session Name folder. It can be used for troubleshooting if the user encounters persistent problems.



Logbook tab overview

Note

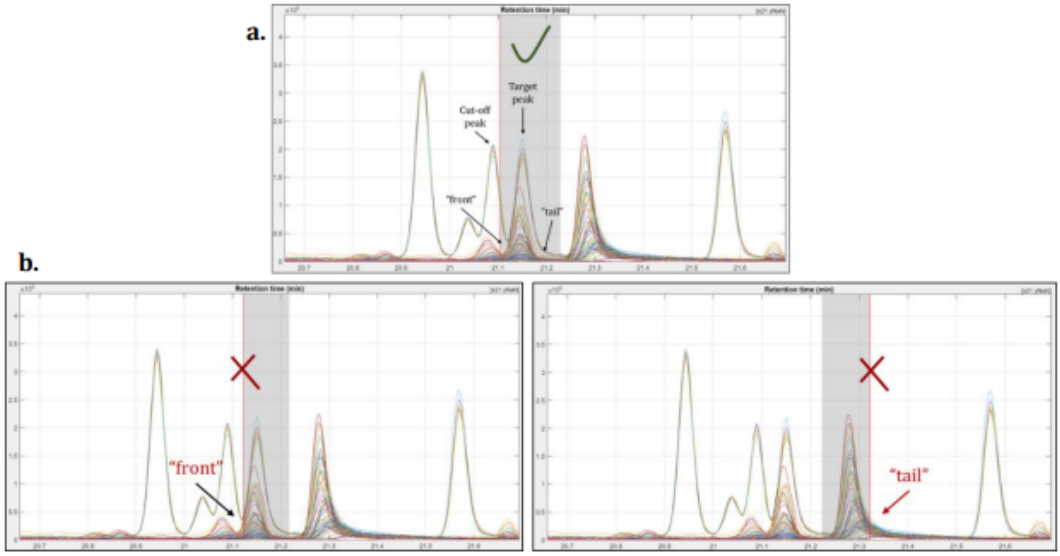
Tip: At the top-right corner of every tab, there is a help button (question mark icon) that provides some tooltips.

Troubleshooting:

PART A. Interval selection

Example 1. Cut-off peaks

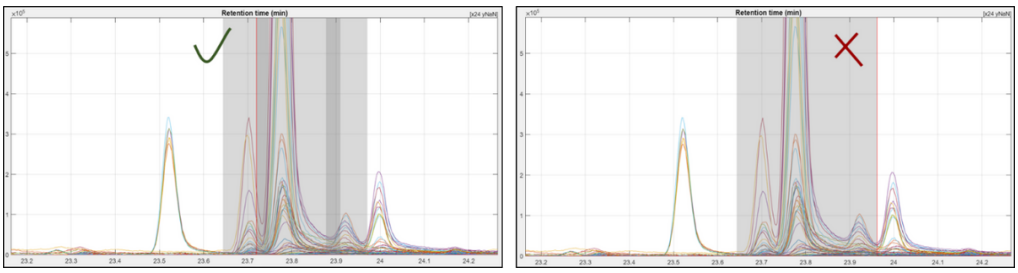
The interval should capture the entire target peak, including the curved fronting and tailing areas nearer the baseline, even if it means that interval includes part of another (cut-off) peak (Figure below a). The portion of cut-off peak will be excluded in later model evaluation steps. Figure below b show examples of a poor interval selection, where either the “front” (left) or “tail” (right) of the target peak are not correctly captured by the interval (grey shading).



Dealing with cut-off peaks: a. Correct interval selection for a given target peak; b. Examples of poor interval selection

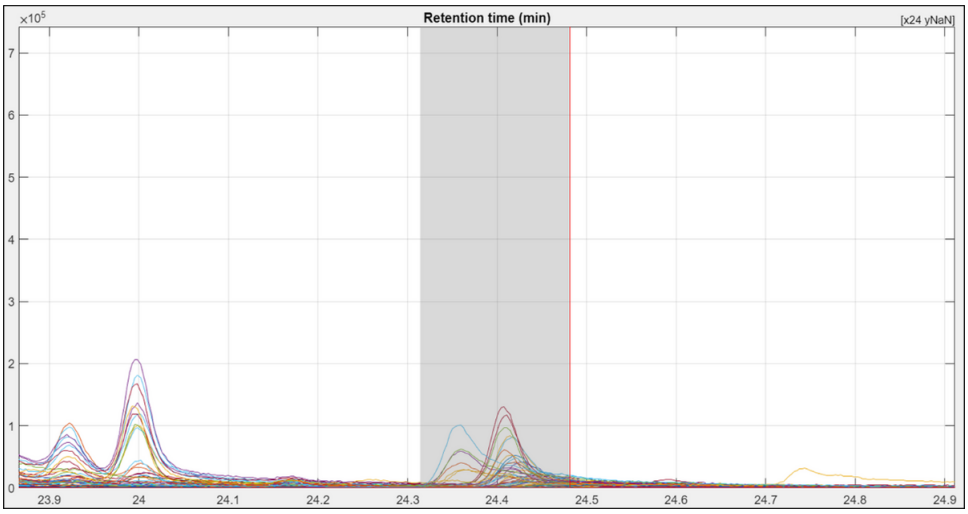
Example 2. Overlapped intervals

It can happen that boundaries of some intervals will end up overlapped, especially for compounds with very close retention indices. In Figure below, we have three peaks that do have some overlap, but are clearly distinct and their boundaries are clear. The example to the left shows the three peaks selected appropriately as three separate, slightly overlapping, intervals. It is advisable to avoid lazy interval selection in these cases, as shown in the right example, where a single interval covers all three peaks. The more peaks in an interval, the higher the risk of running into problems with recovering the pertinent information correctly.



Dealing with overlapped intervals: close but not co-eluted compounds

Nonetheless, when target peaks are actually co-eluting, we have no choice but to define a single interval covering both peaks. For instance, in Figure below, the two peaks seems to be independent, but co-eluting to the extent that the boundaries between them are not clear. Therefore, it is perhaps safer to define both peaks in a single interval than to risk incorrectly selecting individual intervals and cutting off part of the peak. In case they are actually two different compounds, they will be separated based on their mass spectra during the later model selection.

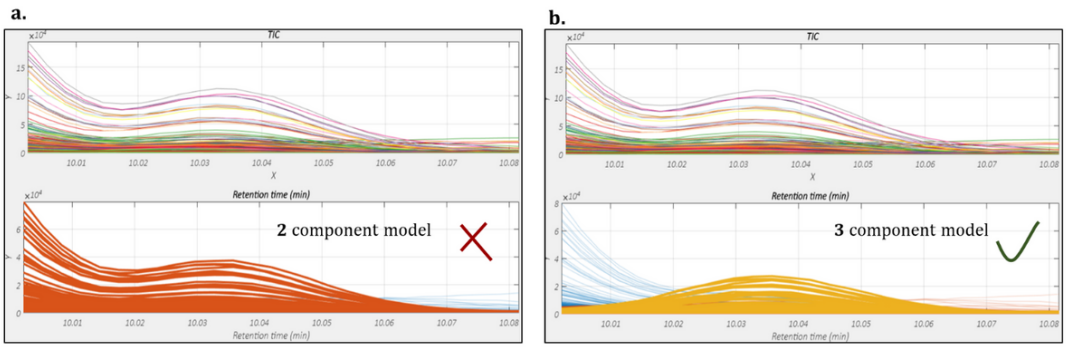


Dealing with overlapped intervals: co-eluting compounds

PART B. Model analysis

Example 1. A target peak and cut-off peak are described by the same model component

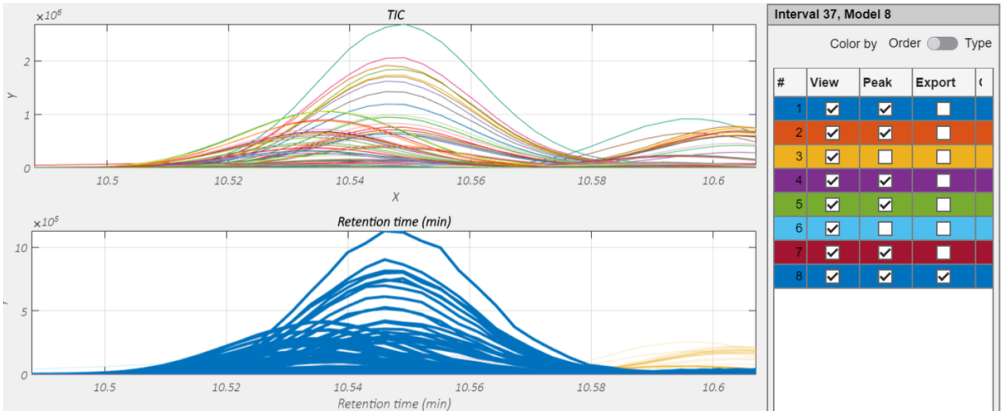
When a given interval contains a target and a cut-off peak, it can occur that both of them are described by the same component. Usually, this situation can be avoided by selecting a PARAFAC2 model with a greater number of components as seen in Figure below.



Increasing the number of components to avoid a single component describing a target and a cut-off peak:
a. 2-component model; b. 3-component model

Example 2. Two co-eluting target peaks are described by the same model component

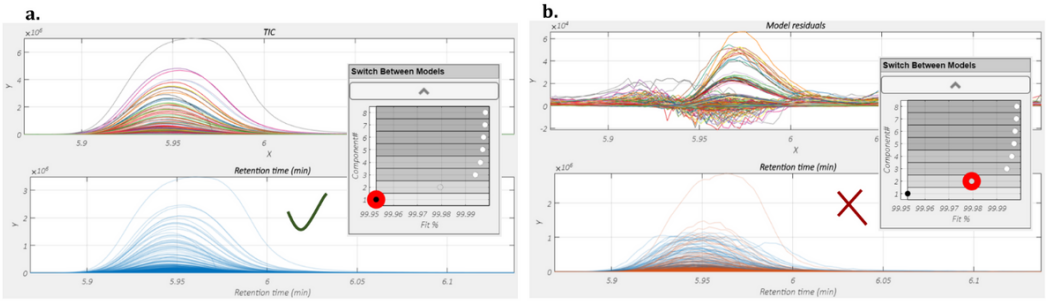
In this case, increasing the number of components does not solve the issue: two co-eluting peaks are still described by the same component (Figure below, blue component). This phenomenon might arise due to e.g., two co-eluting sesquiterpenes whose mass spectra cannot be distinguished, or (more likely), that the samples in the left-hand peak were run in a different analysis batch on the GC-MS to the right-hand peak and there has been a very subtle shift in retention time between the batches. Here, the overlap is too great and the peaks should be exported as a single model component.



Two co-eluting peaks described by the same component

Example 3. Dealing with saturated peaks

The example below shows a single peak in the TIC with a 99.95% fit for the 1-component model(Figure below a). It looks like the fit will drastically improve if choosing the 2-component model(Figure below b), but the reality is that you end up with much worse peak shape (1-component split into two wobbly peaks) and an improvement of only 0.03% in the fit. This phenomena is found in saturated (overloaded) peaks, where we it is recommended to select the 1-component model to avoid overfitting.



Saturated peak modelled with a. 1-component model and b. 2-component model

- If you run into other kinds of minor/major issues with PARADISe, we encourage users to check the Q&A forum available at https://ucphchemometrics.com/paradise/.

Materials

Equipment:

- **Software:**

The latest version PARADISe is publicly available at **https://ucphchemometrics.com/paradise/**
It is possible to download PARADISe as:

1. A stand-alone program in Windows (*.exe), which can be accessed directly by double-clicking the icon after installation.
2. A MATLAB app (*.mlapp) that should be accessed through the “Apps” tab on MATLAB . It requires MATLAB R2022a or later releases, available at www.mathworks.com.

- **Hardware:**

In general, the better the computer features, the better PARADISe will function. In particular, having several cores will allow parallel computations to take place, which can speed up the calculations. Also important, is that there is enough memory on the computer. Recommended requirements are as follows:

1. Windows 10/11 (for *.exe) and the MATLAB Runtime R2022a - freely available via Mathworks, https://se.mathworks.com/products/compiler/matlab-runtime.html
2. MATLAB R2022a or later (for *.mlapp)
3. Between 1 and 512 CPU cores.
4. 8 GB of RAM
5. Separately installed NIST MS Search Program (only if spectra matching is desired). PARADISe has been tested against the NIST11, NIST14 and NIST20 MS Search Program, but other versions should also work.

While it is possible to run the *.mlapp through MATLAB on Linux or iOS based systems, we have neither designed nor tested the software on these platforms.

Load GC-MS data (Data tab)

1 **First-time use**

Firstly, it is recommended you define a unique PARADISe session name.

> Note
>
> If the chosen session name already exists, all previous session details will be overwritten. The session name and PARADISe directory can also be defined in the Settings tab.

1.1 In PARADISe, GC-MS data can be imported as:

CDF (Computable Document Format) files: Data tab > Add CDF Files. It is advisable that all files are loaded from the same directory location (local) in order to avoid importing failures.nGenerally, it is possible to export raw data files as CDF files from the GC-MS software the user is working with e.g., exporting Data to AIA format from Chemstation.

- However, open source software for chromatographic data are also available to convert any type of raw data to CDF files, such as Openchrom from Lablicate.

1.2 Data array built in Matlab (*.mat): Data tab > Import/Export > Import chromatographic data. For the *.exe installation, the installation folder (default is C:\Program Files\University of Copenhagen\paradise\application\) contains a sub-folder named “ExampleFiles” which includes an example example-data-array.mat file. The format (v6.0.1) requires the following variables:

- **Data** which is an array with size: samples × retention time points (scans) × mz channels.
- **rt** which is a column vector specifying the time (in minutes) for each retention time point.
- **mz** which is a row vector specifying the mz values.
- **PathsAndFilenames** which is a cell array (samples × 2), where the first column contains the paths to the sample and the second column contains the filename of each sample.

- The *.mat file should be saved in –v7.3 or later – to allow reading the compressed file efficiently.
- The imported files can be removed from PARADISe by selecting them in the Data tab and then clicking the Remove Selected Files button. Furthermore, if the user presses the Reset PARADISe Session, all elements belonging to that specific session (data, intervals, and models) will be removed – this is equivalent to starting a new session.
- Note that if the imported dataset contains blanks (sample or column blanks), PARADISe will include them for modelling. Thus, it is the user’s choice to decide whether to keep or remove blanks from the data.

2 **Continued use**

The user is able to import previous PARADISe sessions (Data tab > Import/Export > Import Paradise session) or intervals that have been previously defined for any session (Data tab > Import/Export > Import Paradise session).

- Intervals can also be imported as an .xlsx file with the header row Start and End, which list the start and end times for each interval (in minutes), as shown in below figure.

	A	B
1	Start	End
2	6,020154602	6,10653631
3	14,79849297	14,92612
4	10,38031205	10,4699781
5		

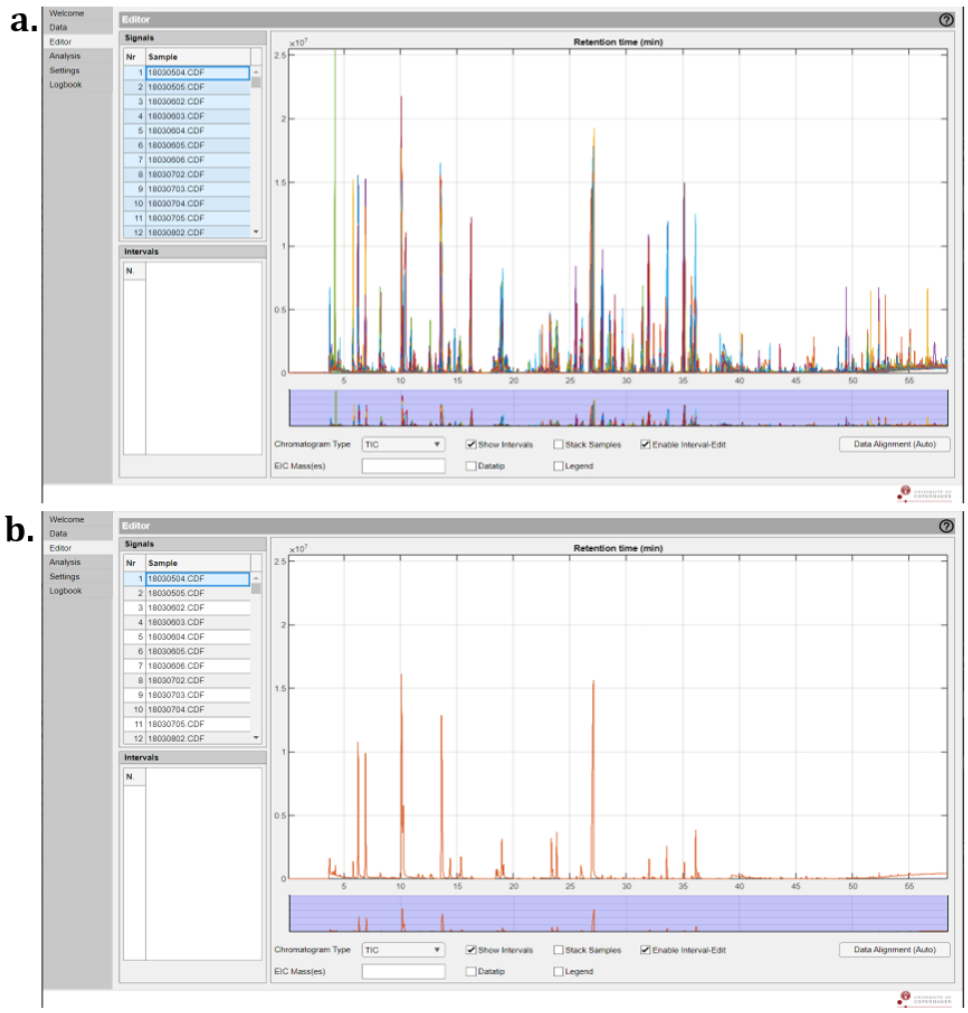
Intervals' .xlsx file

Visual assessment of GC-MS chromatograms

3 The Editor tab shows the overlaid chromatograms of all imported samples (Figure below a), but it is also possible to visualize a single or a few overlaid chromatograms by selecting specific samples in the left panel list called Signals (Figure below b).

Note

Be aware that the y-axis scale does not automatically adjust; it is fixed to the highest signal out of all of the imported chromatograms. This can hinder the quick assessment of individual files.

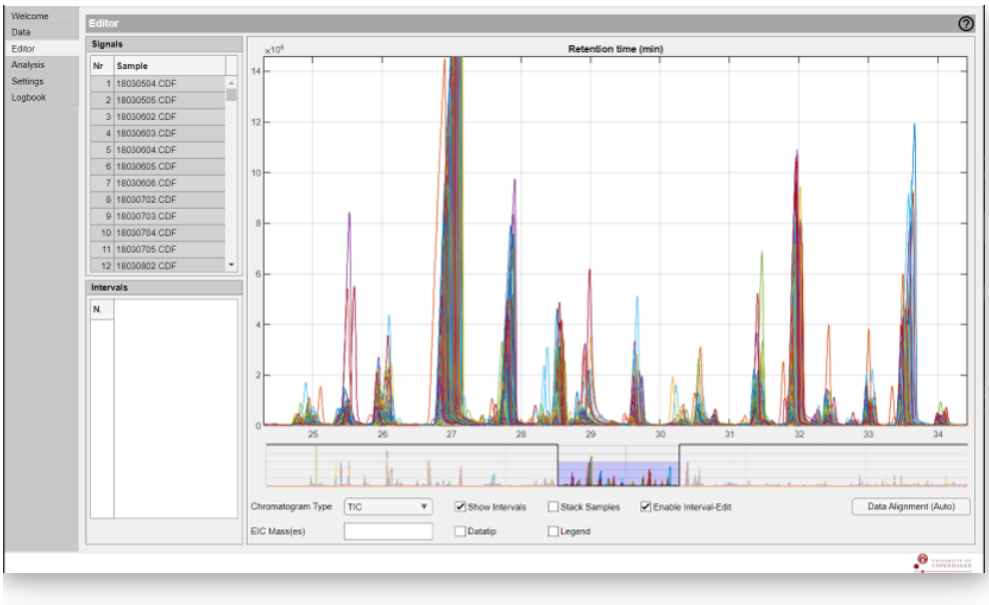


GC-MS chromatograms visualization in PARADISe: a. All samples; b. Selected samples

4 It is possible to navigate through the overlaid chromatograms by left-clicking and dragging the cursor over a selected area to zoom in and by right clicking and selecting Reset Zoom to zoom back out. The small lower panel shows that selected region highlighted in purple (Figure below).

Note

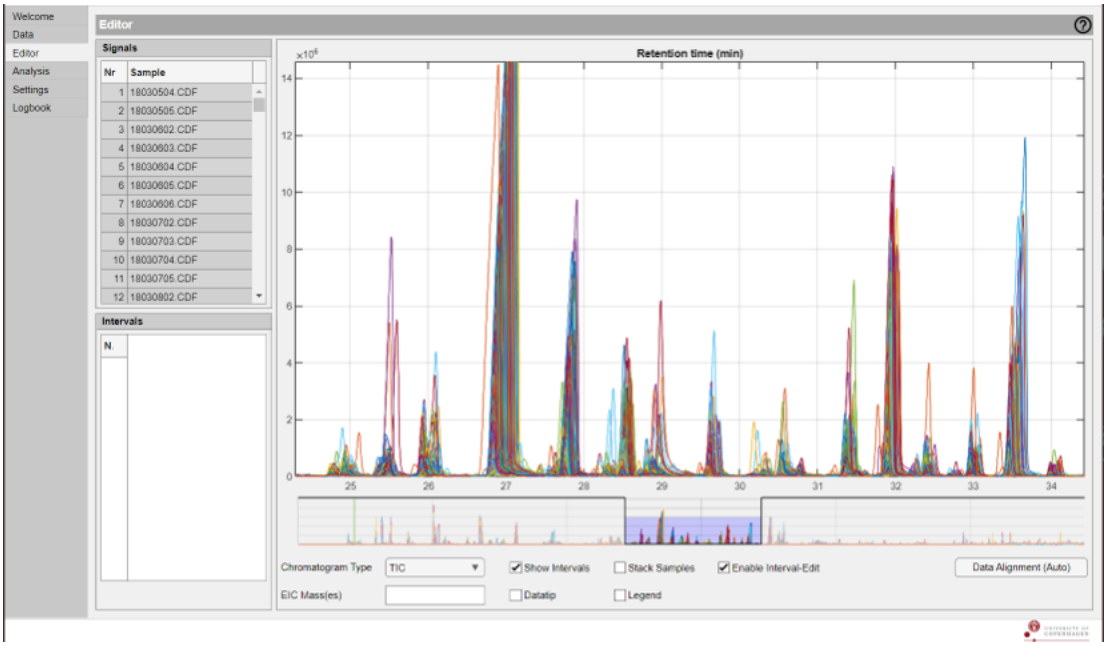
Tip: The user is also able to pan through the chromatogram while zoomed in by sliding the purple viewed section left and right or using the mouse scroll.



Zoomed in area of the overlaid GC-MS chromatograms

5 By default, the Total Ion Chromatogram (TIC) is shown, but the user can switch to the Base Ion Chromatogram (BPC).

- Furthermore, specific masses can be investigated by typing them into the EIC Mass(es) field, hover over the field to see a tooltip describing how to specify masses.
- These visualization tools are found at the bottom of the Editor tab (Figure below).



Zoomed in area of the overlaid GC-MS chromatograms

6 When having a large dataset, it is quite usual to observe retention time shifting across samples. Even though PARAFAC2 models handle shifted data, sometimes raw GC-MS data needs to be aligned in advance to ease the interval selection when shifts are too big.

7 The user is welcome to carry out this step automatically in PARADISe by pressing the Data Alignment (Auto) button at the bottom-right side of the Editor tab. Note that the alignment is a fully automated approach and it cannot currently be undone, so if the result is not satisfactory, the user will need to reload the data.

> Note
>
> The auto-alignment can take a very long time (>24 h) depending on the number of samples imported and the computer used.

8 The automated alignment consists of an initial coshifting (Larsen, Van den Berg & Engelsen, 2006) to handle possible single samples that are shifted dramatically differently from the bulk.

9 This is followed by a correlation optimized warping (Tomasi, Van den Berg & Andersson,2004), where the parameters are estimated based upon an optimization routine (Skov, Van den Berg, Tomasi & Bro, 2006). Usually, the results are simply assessed visually.

> Note
>
> **Tip:** If the alignment did not distort peaks and the intervals seem more easily identified, then the alignment is kept.

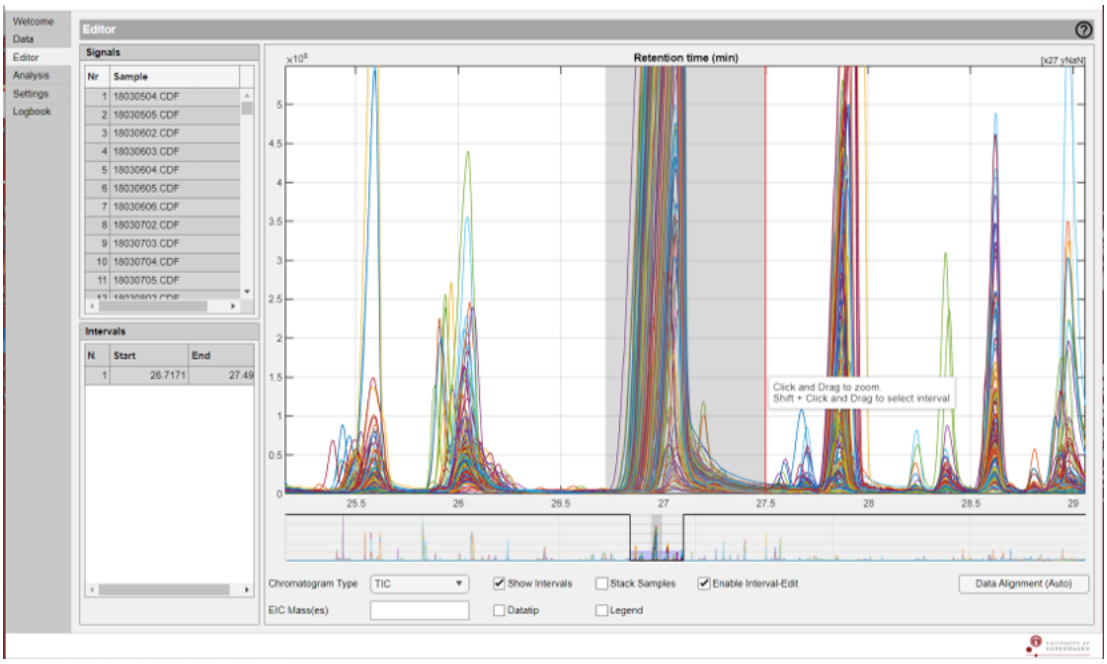
Interval selection

10 The user is in charge of performing the manual selection of intervals through the Editor tab as follows:

10.1 Hold down shift and then left click and drag from left to right to form the boundaries of the peak. The interval will appear shaded in grey and an interval entry for the peak will appear in the bottom left-hand “Intervals” panel.

10.2 The user is able to adjust interval boundaries by hovering the cursor against the boundary (a red line will appear at the boundary, along with a sideways arrow). Click and drag the boundary where desired.

- Alternatively, it is also possible to manually enter the numerical retention time for the boundary in the Intervals panel (bottom left-hand corner) (Figure below).



Editor tab where intervals are selected

10.3 Intervals can also be removed by right clicking either on the interval (grey area) or in the Intervals panel and selecting Delete/Remove Selected Intervals.

10.4 Repeat the interval selection across all peaks.

11 We are aware that GC-MS data is usually very complex and that could be a reason for the user to run into hesitations when setting intervals. Therefore, we provide some tips to handle some of the most common issues in this step in the **Troubleshooting section (part A)** of this protocol.

Fitting PARAFAC2 models

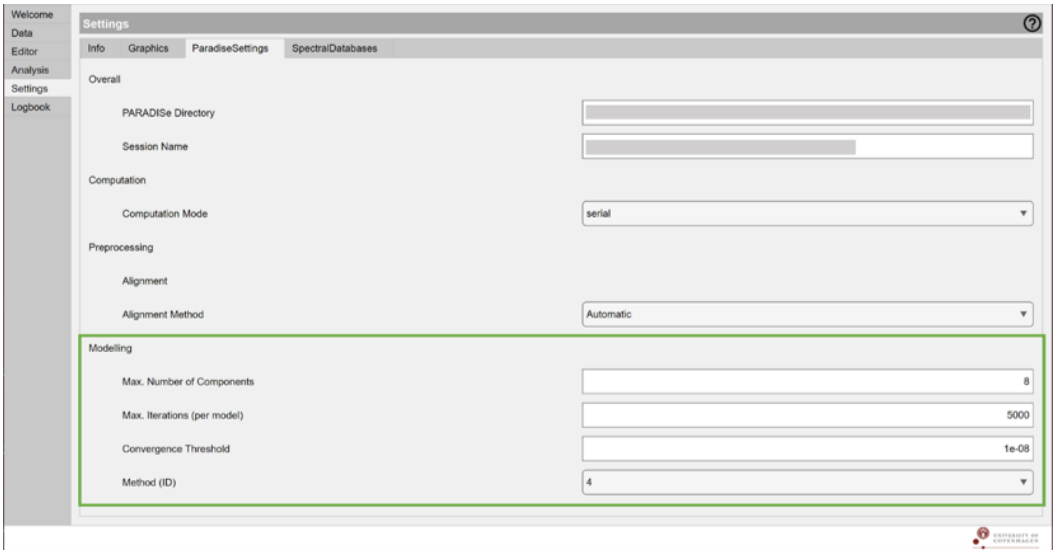
12 Once the intervals have been selected, one model for each interval will be fitted by pressing Fit models on the Paradise Workflow in the upper-left part of the Analysis tab. When the model calculations are complete, an informative pop-up window will appear.

Note

Depending on the number of samples and intervals set, this step can take a while. For this reason, we recommend the use of a (big) auxiliary dedicated computer to carry out the model fitting.

Note

Tip: The general recommendation is to not change the default settings. However, the user can go through the model settings (Settings tab > ParadiseSettings) to change the default modelling parameters (Figure below), if desired.



Modelling parameters (green square) in the Settings tab

12.1 **Max. Number of Components:** it is set to eight by default. You may wish to reduce or increase this number, depending on the complexity of the selected intervals e.g., co-eluted compounds within a single interval. It is also possible to add additional components to selected intervals and recalculate models later, if needed.

12.2 **Max. Iterations (per model) and Convergence Threshold:** The data is decomposed using PARAFAC2 based models, which require an iterative fitting approach.

- The fitting continues until the least squares error is below the convergence threshold or a maximum number of iterations are reached.
- The default convergence threshold is 10^{-8}, which is our recommendation. The default maximum number of iterations is 5000, which is enough for most datasets.
- If particularly difficult intervals are selected, such as many overlapping peaks, then more iterations can be beneficial.
- Decreasing the maximum number of iterations will speed up the modelling, but too few iterations can result in poor modelling.

12.3 **Method (ID):** while method 4 is selected by default, other methods are also available. It is beyond the scope of this tutorial to go into the details of these algorithms and we note that

there is little reason to change the default settings.

1. Nway PARAFAC2
2. Nway PARAFAC2 with non-negativity
3. Nway PARAFAC2 with fast non-negativity
4. Flexible coupling PARAFAC2 with non-negativity
5. Flexible coupling PARAFAC2 with fast non-negativity

13 The user is also able to switch computation mode to parallel when they have many intervals (>10) and many cores (>2), then PARADISe will use one core to fit one interval. Note that MATLAB has an upper limit of 512 cores.

Assessment of fitted models

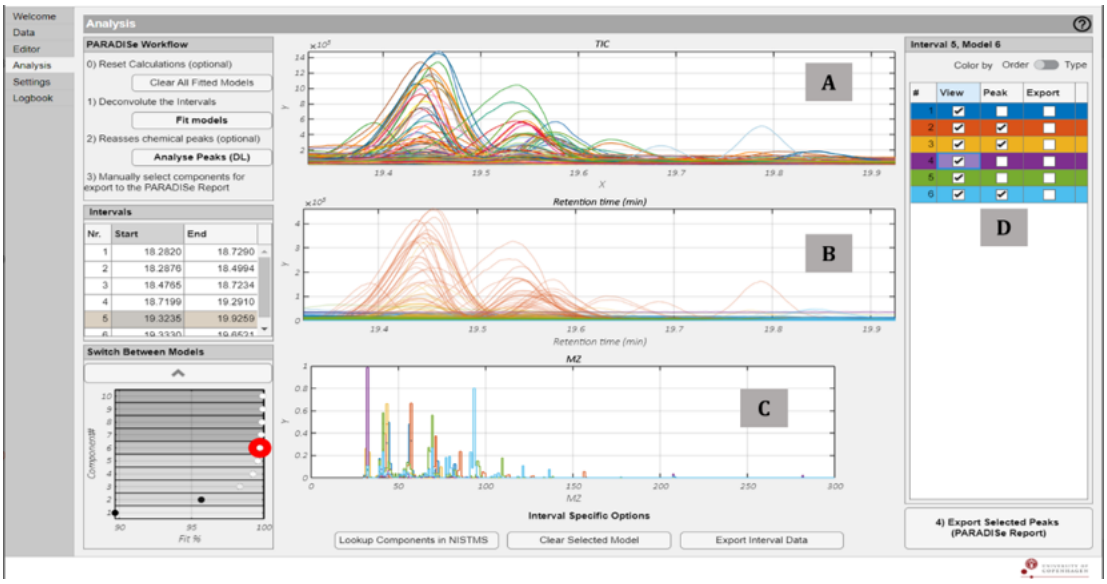
14 The user is able to inspect the fitted PARAFAC2 models in order to choose the optimal number of components.

Note

Tip: Note that PARAFAC2 components are not nested, as in PCA, so every model should be assessed independently according to the features described in this section.

15 PARADISe has a supporting tool available to identify actual chemical compounds, based on deep learning (Risum & Bro, 2019). This should automatically have been run after fitting the models, but if the model fitting was cancelled by the user, then just press the Analyse Peaks (DL) button on the Paradise Workflow in the upper-left panel of the Analysis tab.

16 This process takes much less time than the model-fitting step. As a result, the bottom-left panel in the Analysis tab (Figure below),where the models fitted with different number of components are displayed, will be painted according to the number of chemical compounds found.



Overview of the Analysis tab after fitting models and having run the deep learning tool

17 The darker the background, the greater the number of true peaks present in that model. Contrarily, the background remains white where the deep learning tool finds no peaks. That same panel also shows:

> **Note**
>
> PARADISe does not always guess correctly which components in the models are peaks, but it provides a good guide.

17.1 **Fit (%):** indicator of how well the model describes the information contained within the GC-MS chromatograms.

- A higher fit % value means that a better fit has been achieved for a model including a given number of components; therefore, we aim to maximize the fit and expect values of at least >95%.
- Generally, increasing the number of model components will improve the fit, but when it reaches a plateau, increasing the number of components may result in an over-fitted model.

17.2 **Core consistency:** the internal colour of the circle gives an indication of the “Core consistency” of the model, which is a measure of how adequate the model is.

- The higher the core consistency (i.e., darker shading), the better. As the number of components in the model increases, the core consistency will typically decrease towards zero (black ~100%, white ~0%).
- Ideally, we aim to maximize both core consistency and fit, but in practice, the fit (%) will take precedence.

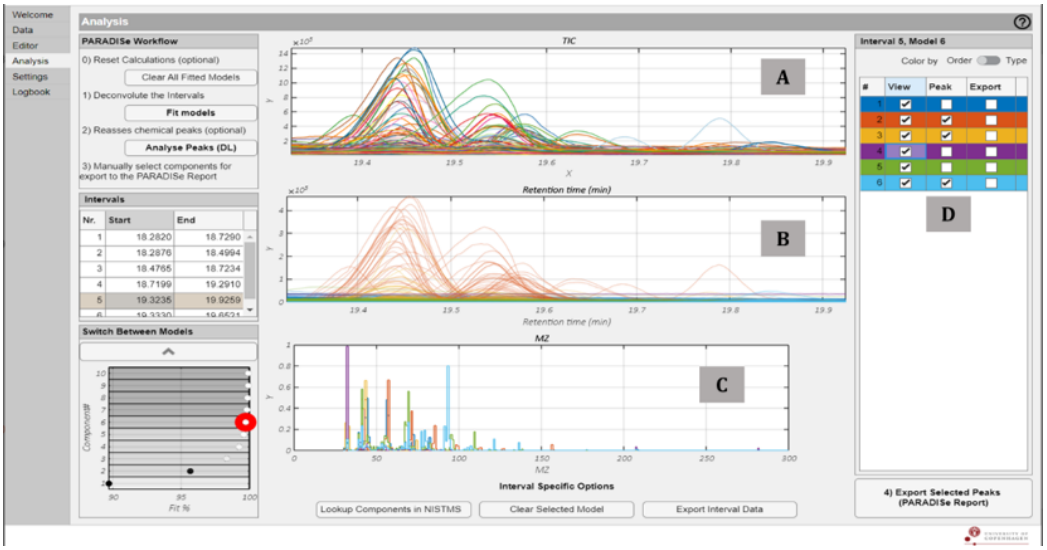
17.3 The **thick red circle** indicates which component model is currently being displayed.

17.4 The **arrow** at the top of the panel allows the user to increase the number of components in case the maximum number of components is not enough to describe the complexity of a giveninterval.

- After this, the new models have to be fitted by pressing again the Fit models button (upper-right panel) which will skip already calculated models.

18 Panel A in Figure below shows the Total Ion Chromatogram (TIC) at the given interval used for modelling (raw or aligned).

- Here, one can make a visual assessment of the number of peaks to expect within the current interval.
- By clicking on the TIC, it displays the Model residuals, which are essentially a measure of the TIC (i.e., the whole signal) minus the individual elution profiles for the model components (i.e., the structured signals (peaks) we are trying to model).

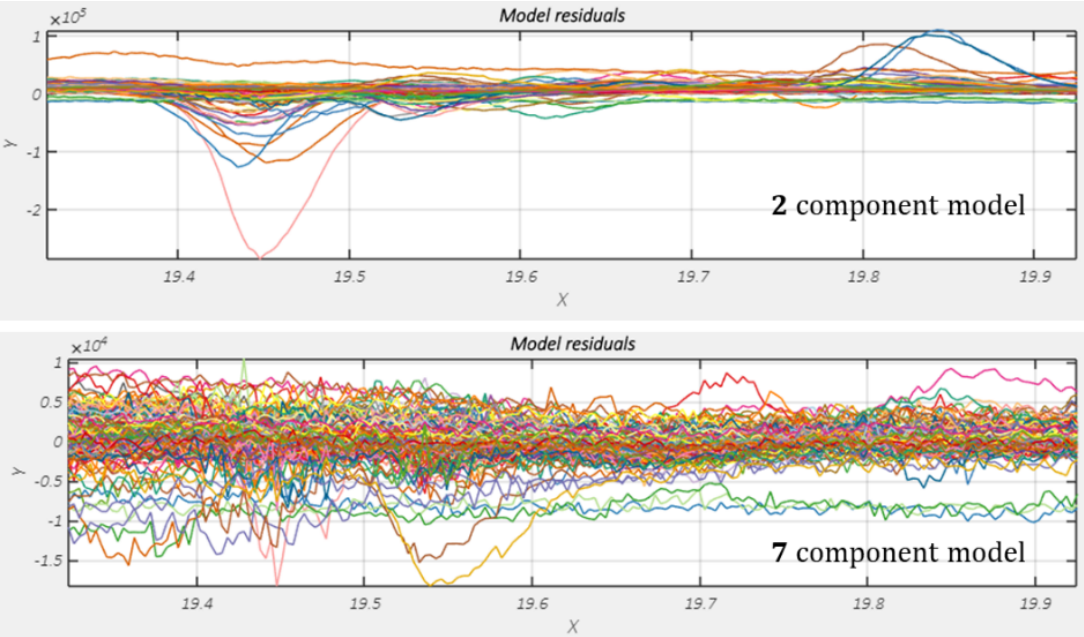


Overview of the Analysis tab after fitting models and having run the deep learning tool

19

Note

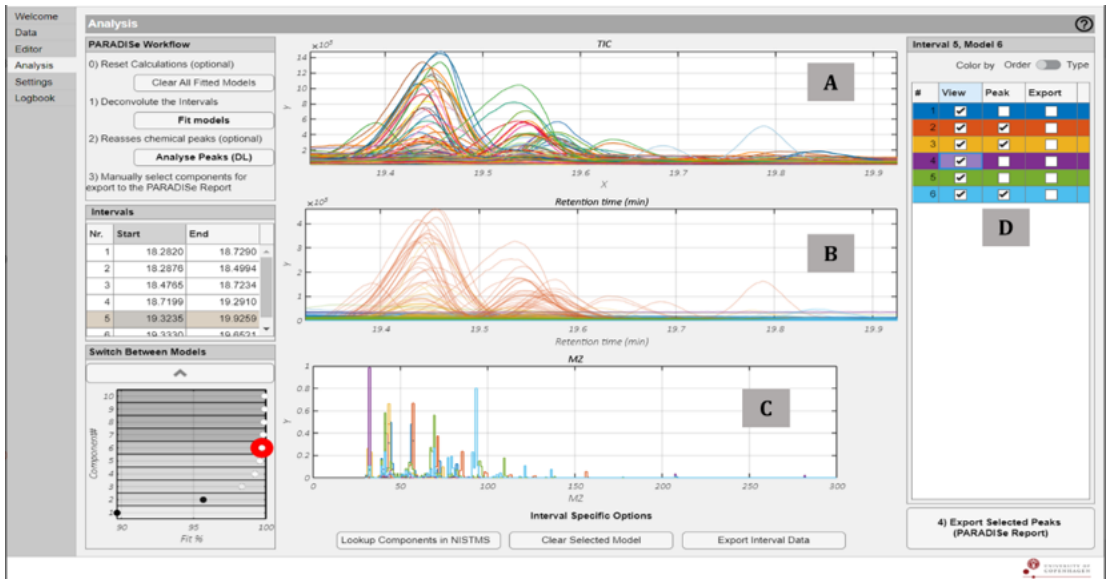
Tip: The residuals are the parts of the chromatograms that are not described by the model components. Residuals showing peak structures, as shown in Figure 13 (above - model with two components, fit ~96%), means that important information has been left out and that the model needs more components (Figure 13, below – model with 7 components, fit of ~100%).



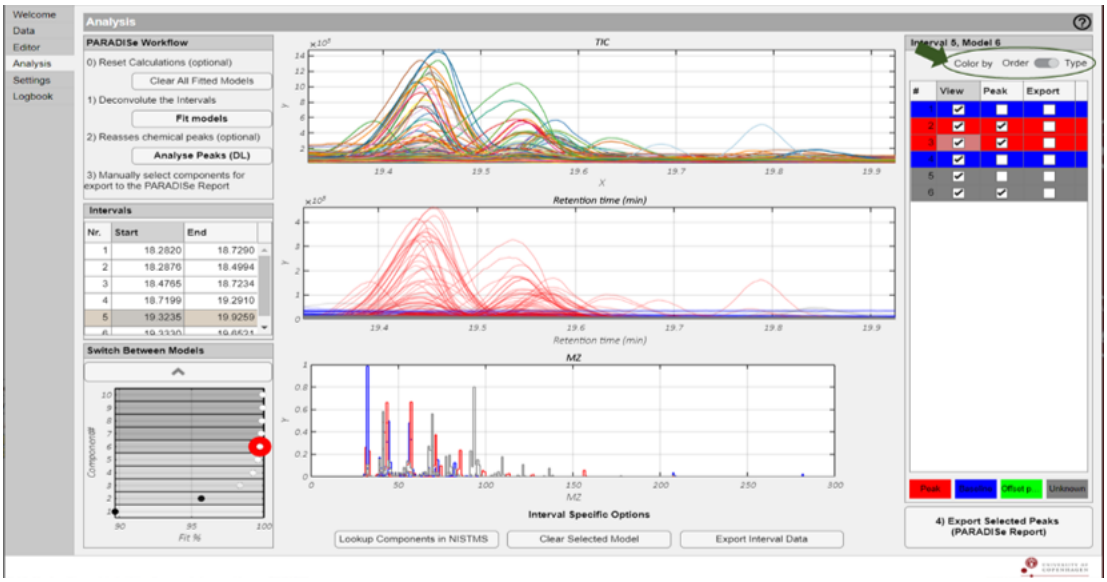
Model residuals of a two-component (above) and a 7-component model (below)

20 Panel B in Figure below a shows the individual elution profiles for the model components (i.e., the structured signals).

- Elution profiles can be coloured according to the number of components or the type of peak defined by the deep learning tool outcome (true peak=red, baseline=blue, offset peak=green, and unknown=grey).
- The user can switch the colouring type in the upper-right panel (Figure below b, green arrow).



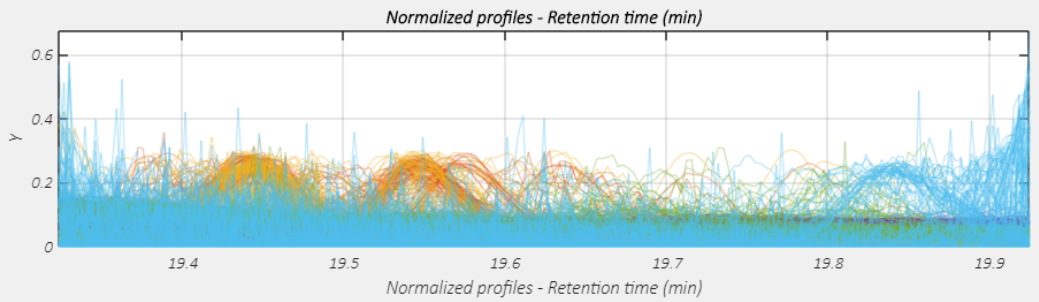
a. Overview of the Analysis tab after fitting models and having run the deep learning tool



b. Overview of the Analysis tab with components coloured by the type of peak

21 It is possible to change the view in panel B to see the normalized elution profiles by clicking directly on the panel. The normalization scales each component, such that its Euclidian norm is one across samples.

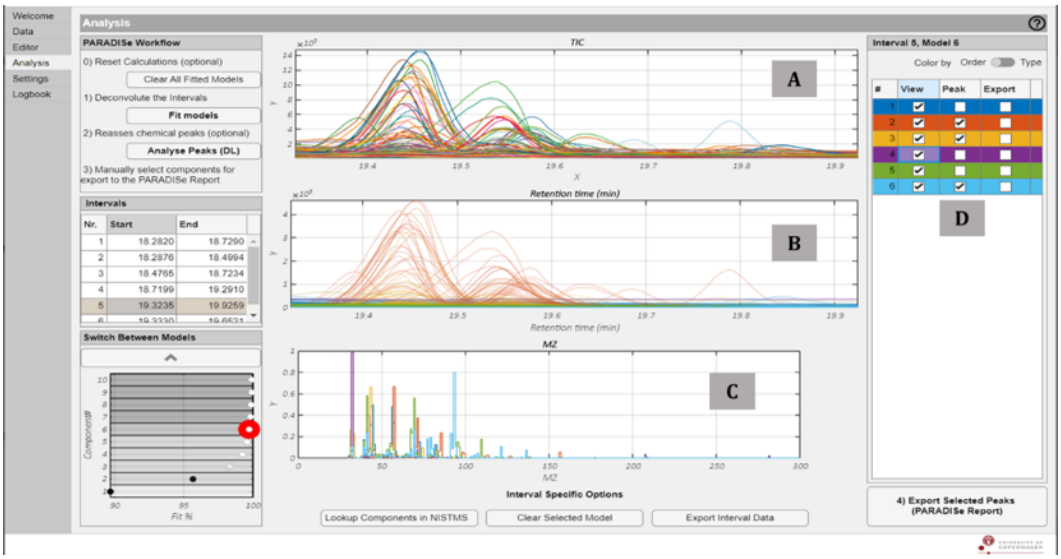
- This can sometimes be helpful to assess if components have a nice peak shape (i.e., approximately Gaussian).
- However, it can also scale up noise, which often gives rise to a saw-tack pattern (Figure below) – this happens especially for small peaks and peaks that are only present in a few samples.
- If this occurs, it is advisable to view the non-normalized profiles and iteratively remove the largest peak from view (untick the View box for that component).



Normalized elution profiles of a given model

22 In panel D (Figure below), the user is able to select/deselect the compounds to be displayed on panel B (View column), selected as true peaks (Peak column), or exported to the peak table (Export column).

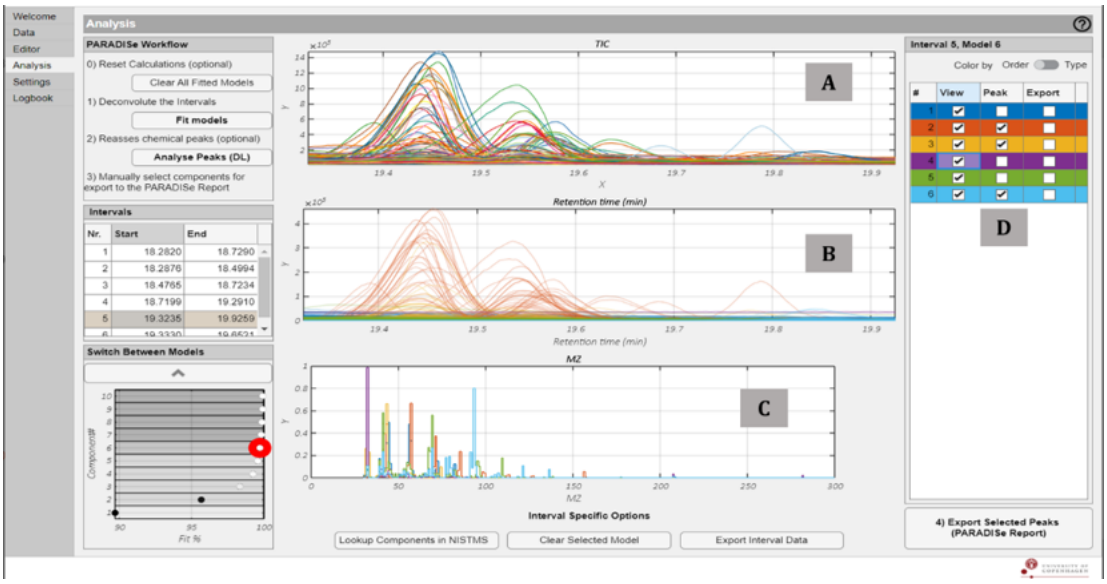
- In some cases, it may be necessary to deselect a large peak from view, in order to better visualize smaller peaks and baseline components.
- Selected components to be exported to the peak table are plotted as a bold line instead of a thin line.



Overview of the Analysis tab after fitting models and having run the deep learning tool

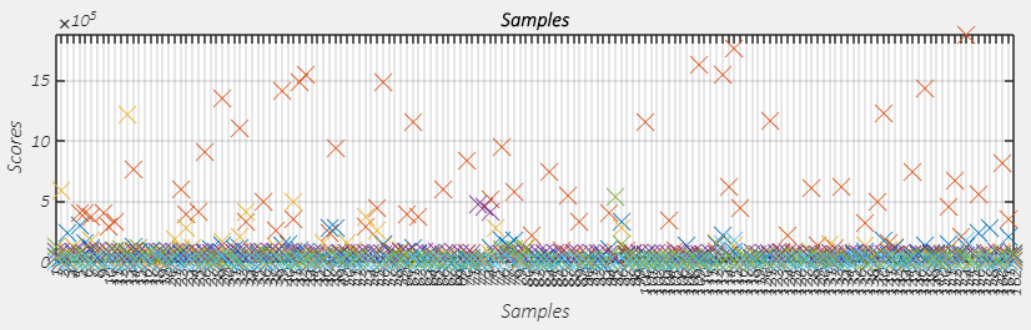
23 Finally, panel C in Figure below shows the resolved mass spectra for each model component (only if they have been selected in panel D to be visualized). This information is crucial to decide whether two peaks (model components) are indeed different chemical compounds, when they present different mass spectra.

- If, contrarily, these components have the same or very similar mass spectra, then it is very likely that the model is over-fitted and too many components have been selected.
- Additionally, the user can choose in which way the mass spectra is plotted (Bars, Lines, Stairs, and Stems) to the right of the plot.



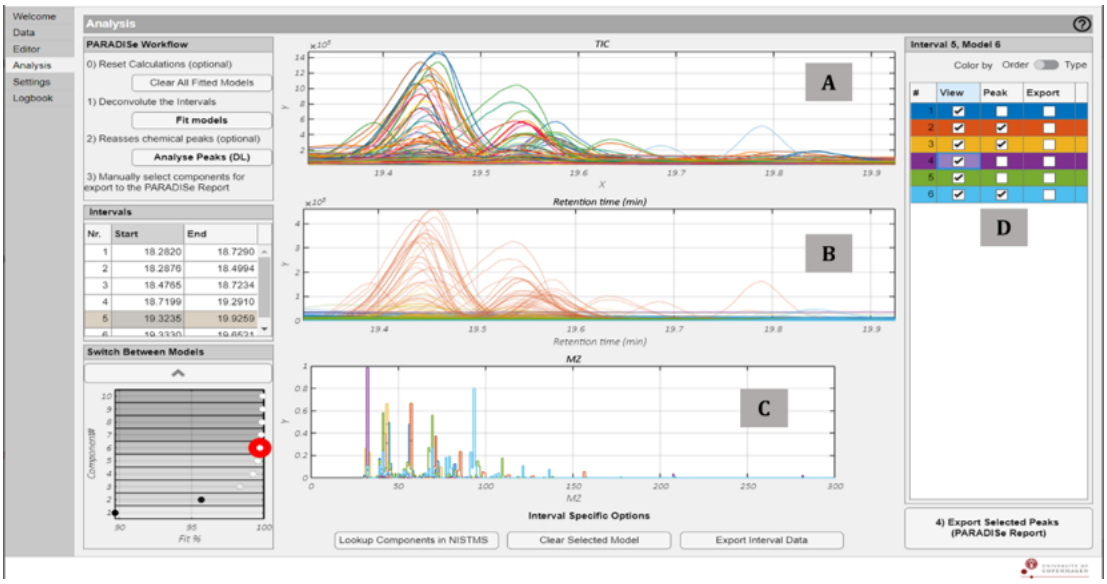
Overview of the Analysis tab after fitting models and having run the deep learning tool

24 Sometimes, it is also useful to check the distribution of relative concentrations (scores as estimated area) for each component between samples. This information can also be displayed in panel C by clicking on it once (Figure below).



Distribution of relative concentrations (estimated area) for each component

25 In case the user is interested on looking up if a given mass spectrum matches with a specific chemical compound, it is possible to do so by pressing the Lookup Components in NISTMS button at the middle-bottom part of the Analysis tab, where Interval Specific Options are shown (Figure below).



Overview of the Analysis tab after fitting models and having run the deep learning tool

26 This requires that the NIST MS Search Program has been separately obtained and installed.

> **Note**
>
> This is often included with vendor software – otherwise see **https://chemdata.nist.gov/**

27 The location of the MSSEARCH folder (containing the nistms.exe and nistms$.exe files) has to be specified at: Settings tab > SpectralDatabases > NIST > Location.

> **Note**
>
> An error warning that the NISTMS is not available can appear. If so, check that the NIST directory has been correctly specified in the Settings tab > SpectralDatabases. Try to check the NIST functionality by locating and running the “nistms.exe” application file independently

28 In that same tab, the user is also able to specify reporting parameters, such as how compounds are sorted in the peak list (by retention time within or across intervals), among others.

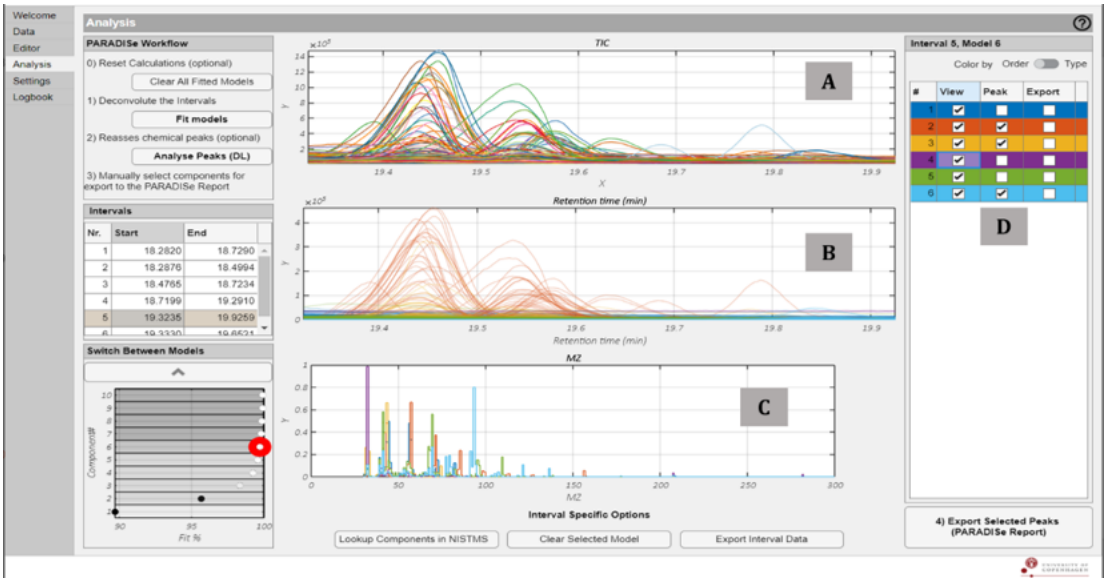
- The number of NISTMS hits and other search parameters needs to be specified through the nistms.exe program.

29 Back to the Interval Specific Options, it is also possible to Clear Selected Model (which can then be refitted by pressing Fit models) and Export Interval Data.

- Exporting an interval may prove useful when having problematic or illustrative intervals for further processing, the exported data is in .mat (MATLAB) format.

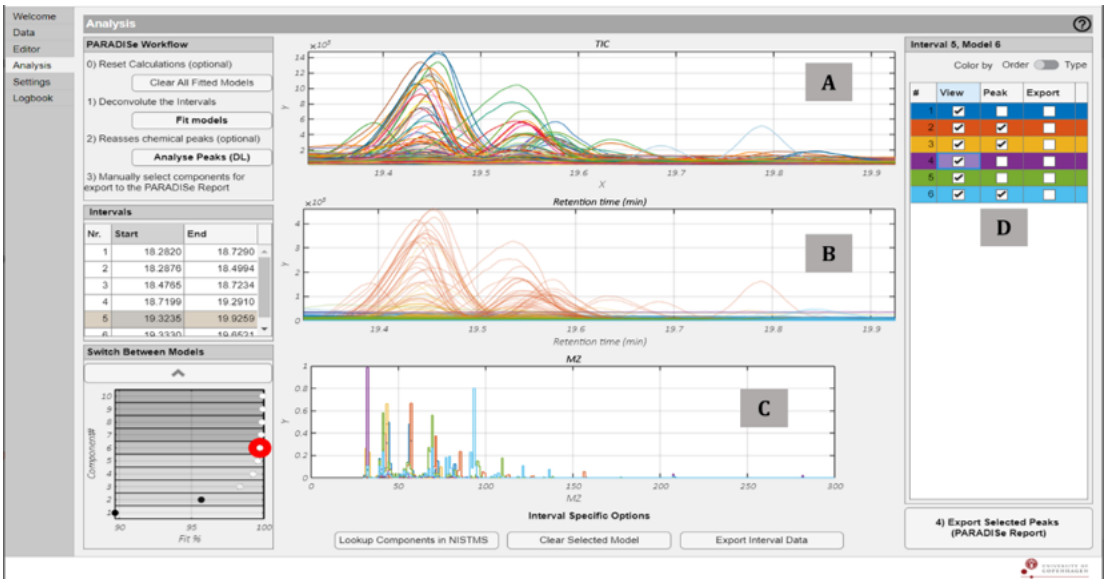
Create a PARADISe report

30 The user must go through all models fitted for the selected intervals (left-middle Intervals panel in the Analysis tab, Figure below) to decide the optimal number of components for each model and select which components should be exported to the peak table.



Overview of the Analysis tab after fitting models and having run the deep learning tool

31 After this, the final step is to create the peak table by pressing the Export Selected Peaks (PARADISe Report) in the bottom right-hand corner of the Analysis tab (Figure below). The user will be prompted to select the filename and where to save the report.



Overview of the Analysis tab after fitting models and having run the deep learning tool

31.1 **Overview** – Gives a few details on the data, the steps performed in PARADISe, and the software version used.

31.2 **Peak Area** – Gives the peak integration (peak areas) of all exported peaks for each sample, together with the interval and model information, and the most suitable chemical identification according to the NIST match factor (highest match factor).

31.3 **Resolved Mass Spectra** – Gives the estimated mz-spectra (as also shown in PARADISe) for each exported peak.

31.4 **Top NIST hits** for each component that has been exported, where the maximum number of hits is defined through nistms.exe.

31.5 **Interval Details** – Details on the PARAFAC2 model performance, which is nice to have along with the report in case of troubleshooting.

32 The peak table (Peak Area sheet) can now be used for further analysis outside of PARADISe.

Note

Note, if NISTMS search is not available, the report will still be generated, but the sheet Top NIST hits will be empty and the Compound Name and Match Quality will also not be available. The Resolved Mass Spectra can then be used as input to for spectral matching using other software.

Protocol references

References:

1. Ballin, N. Z., & Laursen, K. H. To target or not to target? Definitions and nomenclature for targeted versus non-targeted analytical food authentication. Trends in Food Science and Technology 2019, 86, 537–543. https://doi.org/10.1016/j.tifs.2018.09.025

2. Larsen, F.H.; van den Berg, F.; Engelsen, S.B. An Exploratory Chemometric Study of 1H NMR Spectra of Table Wines. Journal of Chemometrics 2006, 20, 198–208.https://doi.org/10.1002/cem.991

3. Skov, T.; Van den Berg, F.; Tomasi, G.; Bro, R. Automated alignment of chromatographic data. Journal of Chemometrics 2006, 20, 484–497. https://doi.org/10.1002/cem.1031

4. Smith, C.A.; Want, E.J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. Analytical Chemistry 2006, 78 (3), 779-787. https://doi.org/10.1021/ac051437y

5. Tomasi, G.; Van den Berg, F.; Andersson, C. Correlation Optimized Warping and Dynamic Time Warping as Preprocessing Methods for Chromatographic Data. Journal of Chemometrics 2004, 18, 231–241. https://doi.org/10.1002/cem.859

Associated Publications:

For further details, we encourage you to read the original publication:

Johnsen, L. G., Skou, P. B., Khakimov, B., & Bro, R. (2017). Gas chromatography–mass spectrometry data processing made easy. Journal of Chromatography A, 1503, 57-64. https://doi.org/10.1016/j.chroma.2017.04.052