

1 Mass spectrometry data processing in MZmine 3: feature detection 2 and annotation

3
4 Steffen Heuckeroth^{1&}, Tito Damiani^{2&}, Aleksandr Smirnov³, Olena Mokshyna², Corinna Brungs², Ansgar Korf⁴,
5 Joshua David Smith^{2,5}, Paolo Stincone⁶, Nicola Dreolin⁷, Louis-Félix Nothias^{8,9}, Tuulia Hyötyläinen¹⁰, Matej
6 Orešič^{10,11}, Uwe Karst¹, Pieter C. Dorrestein¹², Daniel Petras^{6,13}, Xiuxia Du³, Justin J.J. van der Hooft^{14,15}, Robin
7 Schmid^{1,2,12}, Tomáš Pluskal^{2*}

8
9 ¹University of Münster, Münster, Germany, ²Institute of Organic Chemistry and Biochemistry of the Czech Academy of
10 Sciences, Prague, Czech Republic, ³University of North Carolina at Charlotte, Charlotte, NC, USA, ⁴Bruker Daltonics GmbH &
11 Co. KG, Bremen, Germany, ⁵Charles University, Prague, Czech Republic, ⁶University of Tuebingen, Tuebingen, Germany,
12 ⁷Waters Corporation, Wilmslow, United Kingdom, ⁸University of Geneva, Geneva, Switzerland, ⁹Université Côte d'Azur, CNRS,
13 ICN, Nice, France, ¹⁰Örebro University, Örebro, Sweden, ¹¹University of Turku and Åbo Akademi University, Turku, Finland,
14 ¹²University of California San Diego, La Jolla, CA, USA, ¹³University of California Riverside, Riverside, CA, USA, ¹⁴Wageningen
15 University & Research, Wageningen, the Netherlands, ¹⁵University of Johannesburg, Johannesburg, South Africa

16

17

18 & These authors contributed equally

19 * Corresponding authors

20 Tomáš Pluskal, tomas.pluskal@uochb.cas.cz

21 Robin Schmid, rschmid1789@gmail.com

22

23

24

25 **Abstract**

26 Untargeted MS experiments produce complex, multi-dimensional data that are practically impossible to
27 investigate manually. For this reason, computational pipelines are needed to extract relevant information from
28 raw spectral data and convert it into a more comprehensible format. Based on the sample type and/or goal of
29 the study, a variety of MS platforms can be used for such analysis. MZmine is open-source software for the
30 processing of raw spectral data generated by different MS platforms: liquid chromatography–MS (LC–MS), gas
31 chromatography–MS (GC–MS), and MS–imaging. Moreover, the third version of the software, described herein,
32 supports the processing of ion mobility spectrometry (IMS) data. The present protocol provides three distinct
33 procedures to perform feature detection and annotation of untargeted MS data produced by different
34 instrumental setups: LC–(IMS–)MS, GC–MS, and (IMS–)MS imaging. For training purposes, example datasets are
35 provided together with configuration batch files (i.e. list of processing steps and parameters) to allow new users
36 to easily replicate the described workflows. Depending on the number of data files and available computing
37 resources, we anticipate this to take between 2 and 24 hours for new MZmine users and non-experts. Within
38 each procedure, we provide a detailed description for all processing parameters together with
39 instructions/recommendations for their optimization. The main generated outputs are represented by aligned
40 feature tables and fragmentation spectra lists that can be used by other third-party tools for further downstream
41 analysis.

42

43

44

45

46

47

48

49

50

51 Introduction

52
53 Driven by rapid technological advances, the field of mass spectrometry (MS) has undergone substantial progress
54 since the early 2000s.¹ The unprecedented sensitivity and resolving power reached by modern MS instruments
55 enable the comprehensive characterization of both biological and non-biological samples. Liquid
56 chromatography (LC)–MS currently represents the most popular technique for the untargeted profiling of
57 complex mixtures, as hundreds to thousands of molecules can be detected in a single analysis. Moreover,
58 fragmentation (MS², also called tandem MS or MS/MS) data can be simultaneously collected for the
59 identification of these compounds.² Gas chromatography (GC)–MS offers another robust platform for global
60 metabolite profiling of microbial³, plant⁴, and human⁵ samples. Although especially suited for volatile and
61 nonpolar compounds, GC–MS can also be used for the analysis of more polar compounds through sample-
62 derivatization procedures.⁶ The predominant ionisation technique in GC–MS is electron ionisation (EI), because
63 of its universal applicability and the high reproducibility of the fragmentation spectra it produces.⁷
64 Complementary to chromatography-based techniques is MS imaging, which enables the spatial mapping of
65 molecules in tissue samples and is now an established tool in clinical practice.⁸ Compound annotation in MS
66 imaging often relies on the precursor *m/z* measurement. Nevertheless, MS imaging and LC–MS data can be
67 combined (aligned) to raise annotation confidence.⁹ Finally, ion mobility spectrometry (IMS) is being increasingly
68 adopted in disciplines like MS-based metabolomics and lipidomics, as it can provide an additional dimension for
69 metabolite separation and identification.^{10,11}

70 The growing versatility and throughput of MS platforms also pose challenges in terms of volume and
71 complexity of the produced multi-dimensional datasets. In fact, although manual data investigation still plays a
72 crucial role, computational pipelines are essential to streamline the processing of untargeted MS data. General-
73 purpose data processing tools are typically provided by MS instrument vendors. However, research applications
74 often go beyond the scope of vendor software and demand flexible processing solutions that rely on newly
75 published approaches.¹² MZmine is an open-source framework for processing MS data from different instrument
76 vendors and setups. Over the years, thanks to community efforts and collaboration with other open-source
77 projects, MZmine has become one of the most popular tools for visualising and analysing untargeted MS data.
78 The third version of the software, MZmine 3, has been released recently⁹ and includes several new
79 functionalities such as a re-designed graphical user interface (GUI), improved feature detection workflows, and
80 support for MS imaging and IMS data. In this protocol, we provide stepwise instructions for processing
81 untargeted MS data from several different platforms, using MZmine 3 (see [Overview of the method](#)).

82

83 Feature detection and annotation

84 The goal of MS data processing is to turn raw spectral data into a list of detected ions, to estimate their
85 abundance, and to assign chemical annotations based on multiple criteria.¹³ In MZmine, this is done in a three-
86 step approach. First, *raw* spectral data are centroided and intensity thresholds can be applied to exclude low-
87 intensity signals (e.g. electronic noise) from further processing. The second step is known as *feature detection*
88 (also ‘feature finding’ or ‘peak picking’) and represents the cornerstone of the processing. A feature can be seen
89 as an *m/z* signal (more often a group of signals) related to a single metabolite detected during MS analysis. Based
90 on the instrument setup, a *feature* can be characterised by additional identifiers such as retention time (RT) in
91 chromatography–MS experiments or spatial coordinates in MS imaging data. Untargeted MS experiments
92 typically yield hundreds to thousands of *features*, although a relatively small portion corresponds to meaningful
93 metabolites detected in the sample.¹⁴ For this reason, the goal of *feature detection* is to retain all relevant
94 *features* in the raw spectral data while discarding ‘noisy’ signals. Moreover, *features* detected in different
95 samples can be aligned to enable consistent sample-to-sample comparison (e.g. statistical analysis). The third
96 and last step of MS data processing in MZmine is *feature annotation*. Here, various chemical annotations can be
97 assigned to each *feature* based on additional information retrieved from raw spectral data (e.g. isotope pattern,

98 MS² spectra), using dedicated modules (e.g. lipid annotation¹⁵), or via leveraging the direct integration of
 99 MZmine with other popular annotation tools (e.g. SIRIUS¹⁶, GNPS¹⁷).
 100

101 **Table 1. Terminology for MZmine 3 data processing.**

Term	Explanation
Feature	In the field of mass spectrometry, the term <i>feature</i> is used to refer to a ‘meaningful’ signal produced by a chemical entity detected during MS analysis. Features are characterised by a mass-to-charge ratio (<i>m/z</i>), intensity, and, based on the type of MS experiment, additional identifiers: <ul style="list-style-type: none"> - In chromatography–MS experiments (e.g. LC–MS or GC–MS), <i>features</i> are associated with the RT of the chromatographic peak. - In MS imaging data, features are associated with spatial coordinates. - When IMS is used, features are also associated with their ion mobility value.
Feature list	List of features detected in a single raw data file.
Aligned feature table	List of features obtained by merging (i.e. aligning) feature lists of multiple samples.
Mass list	During spectral processing (see Fig. 1), each mass spectrum in the data files is processed individually and stored as a list of <i>m/z</i> and intensity pairs (called the <i>mass list</i>) that is readily usable by MZmine for the subsequent <i>feature detection</i> steps
Extracted ion chromatogram (EIC)	Signal intensity of a specific <i>m/z</i> displayed at any one retention time of the chromatographic run.
Extracted ion mobilogram (EIM)	Signal intensity of a specific <i>m/z</i> displayed along the mobility dimension.
Chromatogram and mobilogram resolving	Splitting EIC or EIM traces that contain multiple peaks into individual features. This includes the correct splitting of partially co-eluting peaks (e.g. shoulder peaks).
RT-resolved features	Features resulting from a chromatogram resolving process (see Procedure 1 – Step 6). See also ‘Chromatogram and mobilogram resolving’ section.
IMS-resolved features	Features resulting from a mobilogram resolving process (see Procedure 1 – Step 9). See also ‘chromatogram and mobilogram resolving’.
Mobility scan	Mobility scans are the individual MS spectra collected during each IMS separation cycle. Each mobility scan corresponds to a data point in a mobilogram. Frame spectra are obtained by merging these mobility scans.
Frame spectrum	<i>Frame spectrum</i> (also referred to as ‘summed frame spectrum’) is the mass spectrum resulting from the sum of all the mobility scans collected during an IMS separation cycle. Each <i>frame spectrum</i> corresponds to a data point in a chromatogram.
Isotopic pattern	Distribution of <i>m/z</i> signals in an MS ¹ spectrum that arises from isotopologues of the same molecule. Isotopologues are molecules that differ only in the isotopic composition of their atoms.
Isotopologue feature	Features generated by isotopologues of the same chemical entity.
Precursor ion	Ion selected and subjected to an MS/MS experiment to produce smaller fragment ions.
Component	In GC–EI–MS, components refers to a group of chromatographic signals with overlapping RTs and peak shape, which includes precursor and product ions generated in the EI source.
Data-dependent acquisition (DDA)	MS data acquisition mode where a certain number of precursor ions from an MS ¹ scan is selected for subsequent fragmentation experiments, one by one. For example, in topN-DDA, the most N-abundant signals in an MS ¹ survey scan are selected for fragmentation.

Data-independent acquisition (DIA)	MS data acquisition mode where all ions within a selected m/z range are selected from an MS^1 scan and subsequently fragmented. For example, in <i>all ion fragmentation</i> experiments (e.g. AIF, MS^E), the full m/z range undergoes subsequent fragmentation.
Spectral deconvolution	A procedure to computationally reconstruct MS spectra for co-eluting analytes in GC–EI–MS data. This is needed due to the extensive in-source fragmentation caused by EI.
Scan-to-scan tolerance	Tolerance between (usually) consecutive scans of the same instrument in a single acquisition. For example, the scan-to-scan m/z tolerance will depend on the mass accuracy, resolution, and precision of your instrument. To optimise this parameter, we recommend manually inspecting the raw data to determine the typical mass accuracy fluctuation over consecutive scans.
Feature-to-feature tolerance	Tolerance between multiple features of the same acquisition (sample). For example, a value of feature-to-feature m/z tolerance will refer to the actual vs. expected difference of multiple ion adducts of the same compound (e.g. $[M+H]^+$ and $[M+Na]^+$). An RT tolerance would refer to the retention time difference of the different adducts. The accuracies within a sample are usually greater than across samples.
Sample-to-sample tolerance	Tolerance for the same compound between multiple instrumental acquisitions, for example, replicates or other samples. Usually described by m/z , RT, and mobility tolerances. These tolerances are usually greater than the feature-to-feature tolerances within one sample.
MS level	Stage of fragmentation for a given scan. ‘MS level = 1’ means no fragmentation, ‘MS level = 2’ means tandem fragmentation (i.e. MS^2), ‘MS level = 3’ means MS^3 fragmentation, etc.
m/z peak	A peak within an MS spectrum corresponding to a detected ion.
Profile data	When MS data are acquired in profile mode, an m/z peak is represented by a collection of signals over several data points. Profile raw data preserves the original information in the data, which may be lost during centroiding. However, data files in profile mode can be significantly larger in memory size. Profile data can be converted to centroid data (i.e. <i>centroiding</i>).
Centroid data	MS data displayed as discrete m/z signals and corresponding intensities (i.e. m/z and intensity pairs). Centroid data files are significantly smaller in memory size than profile data files.
Centroiding	Converting raw mass spectra acquired in profile mode into discrete m/z values with associated intensities (centroided spectra).
Module	A module can be seen as a piece of software designed to carry out a specific task, independently of the rest of the system. A module takes some input data, performs a set of processing operations, and produces an output file, which can then be taken as input by another module. By doing so, different modules can be combined into a unique processing pipeline without affecting the entire system (modular architecture). This gives the user the flexibility to design custom workflows or even develop their own module.

102

103

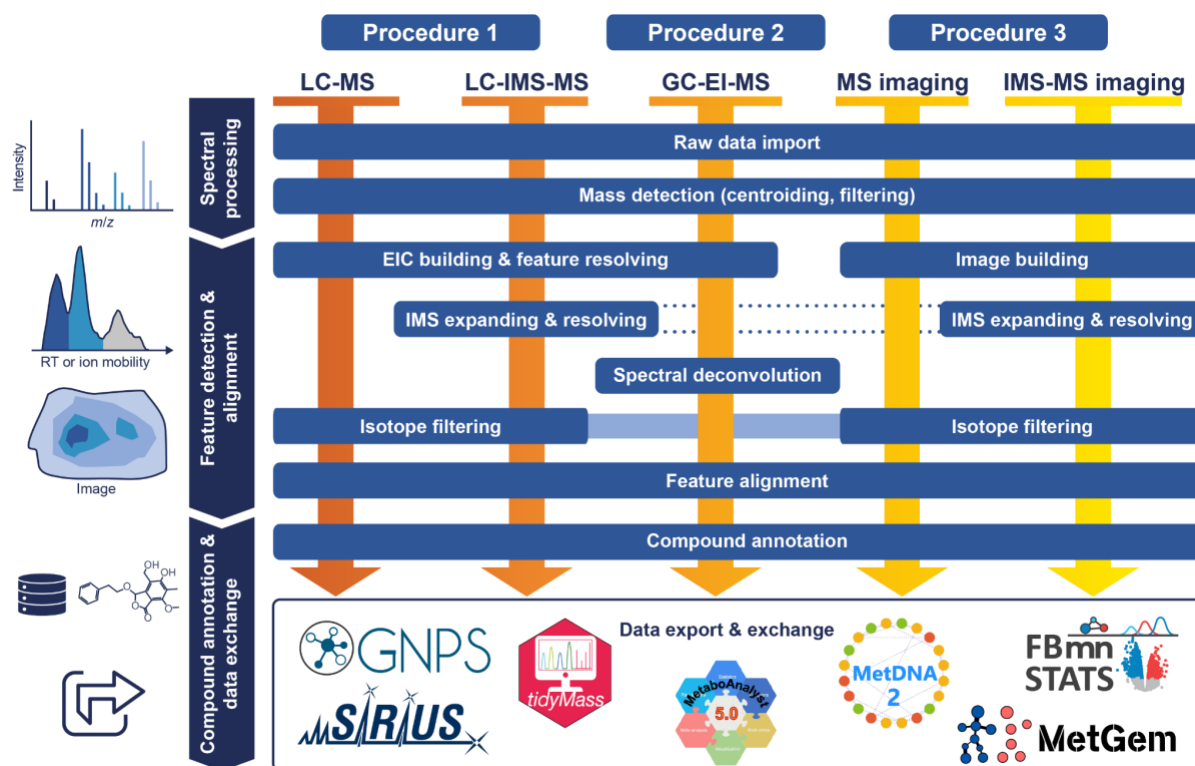
104 Overview of the method

105 Different data processing workflows are needed for different instrumental setups (e.g. chromatography-MS vs
106 MS imaging) and/or acquisition settings.¹⁸ The present protocol describes how to use MZmine 3 to perform
107 *feature detection* and *annotation* on MS data produced by three different platforms: LC–MS, GC–EI–MS, and MS
108 imaging. Moreover, LC–MS and MS imaging data acquired with IMS devices (i.e. LC–IMS–MS and IMS–MS
109 imaging) can also be processed.

110 Since many data processing parameters are dataset-specific and require insights into raw spectral data (e.g.
 111 chromatographic setup, MS analyser performance) and/or domain-specific knowledge, generally-applicable
 112 values cannot be given. Similarly, showcasing this protocol on example data might generate confusion among
 113 non-experts, as the provided parameter values cannot be readily used for processing different datasets. For this
 114 reason, we structured the present protocol as follows. First, we provide various example datasets and
 115 corresponding batch files (i.e. list of processing steps and related parameters, see **Table 1**) to help new users
 116 getting familiar with the software (see **Reproducing the procedures with the 'Batch mode'** in the **Equipment**
 117 **setup** section). Then, we describe three distinct procedures to perform *feature detection* and *annotation* on
 118 different data types (**Fig. 1**):

- 119 - [Procedure 1](#) for the analysis of LC-MS and LC-IMS-MS data.
- 120 - [Procedure 2](#) for the analysis of GC-EI-MS data.
- 121 - [Procedure 3](#) for the analysis of MS imaging and IMS-MS imaging data.

122 In each procedure, we give instructions for selecting the correct processing steps (i.e. MZmine modules)
 123 on the data type and, for each module, we provide parameters description as well as recommendations for their
 124 optimization. We encourage new MZmine users to first process the provided example datasets using the
 125 corresponding batch files as described in [Reproducing the procedures with the 'Batch mode'](#) section. Thereafter,
 126 the same batch files can be used as a starting point and adjusted for processing new datasets based on the
 127 instructions given in each procedure. If needed, more detailed explanations and tutorial videos can be found in
 128 the MZmine [online documentation](#).



129
 130 **Figure 1: Overview of the data processing workflows in MZmine.** The main data processing steps of the three procedures
 131 described in the present protocol are outlined.

132 Applications of the method

133 MZmine can process MS data for various applications including metabolomics, lipidomics, natural product
 134 research, or environmental studies. Although the presented protocols use example datasets acquired from
 135 biological samples, MZmine has been used to process (IMS-)MS data from virtually any sample types, including
 136 food¹⁹, dissolved organic matter²⁰, archaeological artefacts²¹, or tattoo pigments^{22,23}. [Procedure 1](#) and [3](#) describe

137 how to process LC–IMS–MS and IMS–MS imaging, respectively. As highlighted in **Fig. 1**, the same pipelines can
138 be used to process non-IMS data by skipping the ‘IMS expanding and resolving’ step. [Procedure 2](#) covers the
139 processing of untargeted GC–EI–MS data, which requires a dedicated spectral deconvolution step to handle the
140 extensive in-source fragmentation produced by EI. GC–MS data produced by ‘softer’ ionisation techniques (e.g.
141 chemical ionisation) can be processed using the workflow described in [Procedure 1](#).

142 At the time of writing, MZmine supports the following open data formats: .mzML²⁴, .mzXML²⁵, .imzML²⁶,
143 .netCDF²⁷, and .aird²⁸. Moreover, MZmine supports proprietary formats from Thermo Scientific (.raw) and
144 Bruker Daltonics (.d and .tdf/.tsf). Raw data files from other vendors can also be processed but must first be
145 converted into an open format using vendor-provided or third-party software. The MSConvert²⁹ tool from the
146 ProteoWizard package³⁰ supports the conversion of AB SCIEX, Agilent, Bruker, Shimadzu, Thermo Scientific, and
147 Waters raw data. A step-by-step guide for data conversion with MSConvert is provided in the [online](#)
148 [documentation](#). Both *profile* and *centroid* data can be imported in MZmine. Centroiding of *profile* data can be
149 performed during MZmine processing (see [Procedure 1 - Step 2](#)). However, we recommend using already
150 *centroided* data because of the smaller file size and memory consumption. More information and the latest
151 supported data formats are provided in the [online documentation](#).
152

153 **Comparison with other methods**

154 Over the years, several open software tools for MS data processing have been developed and widely adopted
155 by the scientific community. These include, among others, XCMS³¹, OpenMS³², and MS-DIAL³³. All of these
156 software packages are equipped with a user-friendly graphical user interface (GUI) that greatly assists
157 researchers lacking programming skills.¹² In this regard, MZmine places great emphasis on the development and
158 continuous improvement of highly-interactive GUIs that enable the user to make informed choices on key
159 processing parameters (see, for example, **Box 3**). Furthermore, MZmine can save results from each individual
160 processing step, which can be manually (re-)inspected by the user. This simplifies workflow optimization and
161 backtracking of potential errors during the setting of parameters.

162 At the time of writing, MZmine is among relatively few software packages that support the full processing of
163 IMS data (both LC–IMS–MS and IMS–MS–imaging).⁹ Moreover, one unique function of MZmine is the possibility
164 to combine MS data from various instrumental setups, for example, LC-MS and MS-imaging. To do so, users are
165 normally required to master different data processing software, for instance, specific to chromatography–MS or
166 MS–imaging, and to use a third, external tool to integrate the results. The alignment and annotation of LC–IMS–
167 MS and IMS–MS–imaging data is showcased in [Procedure 3 - Step 10](#).
168

169 **Box 1 - Contributing to MZmine 3**

Since its inception in 2004,³⁴ the MZmine project has evolved into a collaborative, community-driven effort, and nowadays constitutes one of the most popular tools for processing untargeted MS data. Thanks to the modular architecture of MZmine, new modules can be programmed and tested independently, without the need to modify other pieces of the software. Over the years, this has greatly facilitated contributions to the project by new developers and researcher teams from all over the world.⁹

To facilitate the use of the software by new users, the MZmine community creates extensive documentation materials and tutorial videos. The online documentation (<https://mzmine.org/documentation>) provides detailed description of each individual module and is constantly updated with the latest software releases and features. Moreover, it contains a step-by-step guide to program new processing modules and add them to the MZmine codebase.

Anyone can contribute to the MZmine community by:

1. writing documentation: <https://mzmine.org/documentation>,
2. writing code and developing modules: <https://github.com/mzmine/mzmine3>
3. answering questions or discussing developments on the GitHub issues page: <https://github.com/mzmine/mzmine3/issues>

170

171 **Limitations of this protocol and software**

172 Untargeted feature detection workflows in MZmine are primarily designed for data acquired in DDA mode, in
173 particular collision-induced dissociation (CID) and higher-energy collisional dissociation (HCD). Rule-based
174 annotation approaches (e.g., see lipid annotation modules, [Procedure 1 - Step 20](#)) may be limited when used on
175 data produced with different ionisation techniques, such as electron impact excitation of ions from organics
176 (EIEIO). Nonetheless, annotation of EIEIO spectra can still be performed via spectral library search against
177 reference libraries produced using EIEIO.

178 At the time of writing, MZmine 3 is not optimised for processing MS data acquired in *data-independent*
179 *acquisition* (DIA) mode. We recommend users to explore other more established tools for DIA data processing,
180 such as MS-DIAL³³ and Skyline³⁵.

181 The present protocol does not illustrate all MZmine 3 features that may be useful for untargeted MS data
182 processing (e.g. blank sample subtraction). Information and tutorials on how to use additional software
183 functions can be found in the [online documentation](#).

184 **Processing large datasets**

185 In MZmine 3, special attention was directed toward scalability due to the ever-increasing study sizes and
186 availability of public data. Nevertheless, when processing very large datasets (e.g. hundreds to thousands of data
187 files), we recommend applying the following measures to minimise the memory consumption and reduce the
188 chance of software crashes during processing. All points are covered in more detail in the online documentation.

- 189 ● Set the temporary file directory to a fast local drive (e.g. solid-state drive) with enough free space (see
190 **Materials** section).
- 191 ● Use the recommended file formats (e.g. native .tdf and .tsf formats for Bruker data, .mzML format for
192 all other vendors, .imzML format for imaging data).
- 193 ● Optimise the batch file on a subset of representative samples (e.g. pooled QCs, randomly selected
194 samples) before proceeding to the full processing.
- 195 ● Run MZmine through the command line interface. This will avoid memory usage by the GUI.
- 196 ● Use the advanced data import (see [Procedure 1 – Step 1](#)) to perform the mass detection during the
197 data import. By doing so, all signals below the specified noise level are not imported, thus saving
198 memory and processing time.
- 199 ● In the batch file, set the ‘Original feature list’ parameter to either ‘IN PLACE’ or ‘REMOVE’ (see
200 [Procedure 1 – Step 5](#)) in all steps.
- 201 ● Adjust processing parameters. In general, increasing the noise level and other feature constraints (e.g.
202 minimum feature height) will reduce the amount of data to be processed. Missed features will be
203 recovered during gap-filling.
- 204 ● Use the ‘Join aligner’ module for the feature alignment (Procedure 1 – Step 13) and the ‘Peak finder’
205 module for the gap-filling (Procedure 1 – Step 14), as they are optimised for large data volumes.
- 206 ● If the experimental design includes replicate samples and features are expected to be detected in a
207 minimum number of replicates, we recommend applying the ‘Min aligned features (samples)’ filter

208 (Procedure 1 – Step 22) before the gap-filling step. This will remove the ‘non-reproducible’ features and
209 reduce the volume of data being processed.

210 If software crashes still occur after implementing these measures, we recommend upgrading the RAM of your
211 processing PC, or consider using a computer cluster.

212 **Transparent and FAIR data processing**

213 Modern scientific research is required to adhere to the FAIR principles (Findable, Accessible, Interoperable, and
214 Reusable) to ensure transparency, reproducibility, and reusability of the produced results.³⁶ Every scientific
215 publication should provide clear instructions on how and where to access the experimental data and any digital
216 object used, for instance, software tools, algorithms, and workflows.³⁷ MZmine is often part of larger
217 computational workflows and even minor differences in the produced output files can impair the reproducibility
218 of downstream data analysis steps. Therefore, to guarantee full reproducibility of the processing output in
219 MZmine, the following elements are necessary:

- 220 - **Raw data files.** If raw files were converted to open formats (e.g. *.mzML*), ideally, also the original files
221 in the vendor format should be shared. The vendor formats might contain additional metadata that is
222 lost during conversion. Also, profile mode raw data is usually converted to centroid *.mzML* files, which
223 reduces the volume of data.
- 224 - **MZmine batch file.** A batch file contains the complete configuration of a list of processing steps,
225 specifically, modules and their related parameters. Virtually any processing pipeline can be saved as a
226 batch file and executed in the MZmine GUI or in the command-line mode. Loading batch files into the
227 GUI allows to visualise the structure and review settings. Batch files offer a convenient way to share a
228 precise description of the data processing and help others reproduce it.
- 229 - **MZmine version** used to perform the processing. MZmine uses semantic versioning –
230 major.minor.patch version. Stable releases are permanently stored and available in the MZmine GitHub
231 repository, different from the development release that is only meant for testing. For this reason, it is
232 strongly recommended to use stable MZmine releases to process data for scientific publications.
233 Starting with MZmine 3.4.0, information about the MZmine version is also included in the batch file.
234 Furthermore, batch files also contain versions for each step that are updated if any user parameter
235 changes.
- 236 - **Output files** produced by MZmine and exported for downstream data analysis, for example, feature
237 lists and MS spectra files stored in *.csv* and *.mgf*, respectively.
- 238 - **Metadata file** that contains the list of input raw data files and the corresponding sample information
239 based on the experimental design of the study.

240 A way to ensure open and long-term access to the above-listed files is to upload them into public MS data
241 repositories such as GNPS/MassIVE,¹⁷ MetaboLights³⁸ or MetabolomicsWorkbench³⁹. By doing so, the uploaded
242 files are assigned an accession number that can be easily referenced in scientific publications, databases and
243 other resources.

244

245

246

247 **Materials**

248

249 **Software**

- 250 ● MZmine 3 (latest release)
- 251 ● (Optional) MSConvert from the ProteoWizard package (latest release)^{29,30}
- 252 ● (Optional) FTP client (e.g. WinSCP)

253 **Equipment**

254 **Hardware**

- 255 ● Personal computer or other computing resources
 - 256 ○ Windows, Linux, or macOS platform
 - 257 ○ Small datasets (< 100 files): 4+ CPU cores, 8+ GB RAM
 - 258 ○ Medium datasets (100–5,000 files): 8+ CPU cores, 24+ GB RAM
 - 259 ○ Large datasets (> 5,000 files): 8+ CPU cores, 64+ GB RAM
 - 260 ○ MS imaging datasets: 8+ CPU cores, 64+ GB RAM
- 261 ● Internet connection

262 **Datasets**

263 ▲ **CRITICAL** All datasets used in this protocol are publicly available in open data repositories and listed under the
264 accession numbers provided below. The MassIVE datasets can be downloaded using an FTP client by following
265 this step-by-step guide. If you are processing your own dataset(s), make sure data are converted to
266 recommended file formats (see Processing large datasets section).

- 267 ● LC-IMS-MS data from plant extracts (Procedure 1): MSV000091634. This dataset was acquired from
268 the LC-IMS-MS analysis of hydroalcoholic extracts of Piperaceae plants (9 data files). MS data were
269 acquired with a quadrupole time-of-flight (QTOF) mass spectrometer equipped with a trapped-IMS
270 (TIMS) device. Fragmentation MS2 spectra were collected in PASEF mode (i.e. parallel accumulation-
271 serial fragmentation⁴⁰).
- 272 ● GC-EI-MS from clinical trial (Procedure 2): ST000981. This dataset was acquired from a study on healthy
273 research cats receiving clindamycin administered with a synbiotic or a placebo. More information about
274 the dataset can be found in the original publication.⁴¹
- 275 ● MS imaging data from sheep brain samples (Procedure 3): MSV000090328. This dataset was acquired
276 from the LC-IMS-MS and IMS-MS imaging analysis of sheep brain samples. Hydrophilic interaction
277 chromatography (HILIC) chromatography was used. MS data were acquired with a QTOF equipped with
278 a TIMS device. A matrix-assisted laser desorption ionisation (MALDI) source was used for the IMS-MS
279 imaging analysis.
- 280 ● (Optional) LC-IMS-MS data for lipid annotation (Procedure 1 - Step 20): MSV000091642. This dataset
281 was acquired from the LC-IMS-MS analysis of sheep brain samples extracted using methyl-tert-butyl
282 ether.⁴² HILIC chromatography was used. MS data were acquired with a QTOF equipped with a TIMS
283 device. Fragmentation MS2 spectra were collected in PASEF mode.
- 284 ● (Optional) LC-MS data for statistical analysis (Procedure 1 - Step 24): MTBLS265. LC-MS analysis of
285 blood samples from 30 patients. The dataset includes three replicates per sample required by

286 MetaboAnalyst for multivariate statistics. MS data were acquired with an Orbitrap MS instrument.
287 More information about the dataset can be found in the original publication.⁴³

288 Batch files

289 ▲ **CRITICAL** We provide configuration batch files for each example dataset to easily replicate all three
290 procedures described in this protocol. The provided batch files are optimised for the corresponding example
291 dataset. The same batch files should not be used to process different data without adaptation, as this would
292 likely produce unreliable results; rather, they represent a good reference and starting point for parameter
293 optimization.

- 294 ● Batch file for Procedure 1: batch_procedure-1.xml. Batch file for processing LC–IMS–MS data in the
295 native Bruker format. A spectral library search step for feature annotation is included. Export steps for
296 feature-based molecular networking and the SIRIUS software are included.
- 297 ● Batch file for Procedure 2: batch_procedure-2.xml. Batch file for processing GC–EI–MS data
298 (centroided). A spectral library search step for feature annotation is included. An export step for
299 feature-based molecular networking is included.
- 300 ● Batch file for Procedure 3: batch_procedure-3.xml. Batch file for processing IMS–MS data in the native
301 Bruker format.
- 302 ● (Optional) Batch file for Procedure 1 – Step 21: batch_lipid_annotation.xml. Batch file for processing
303 LC–IMS–MS lipidomics data in the native Bruker format. Feature annotation is done using the lipid
304 annotation module.
- 305 ● (Optional) Batch file for Procedure 1 – Step 25: batch_metaboanalyst.xml. Batch file for processing LC–
306 MS data (centroided). An export step for statistical analysis in MetaboAnalyst is included.

307 ⓘ TROUBLESHOOTING

308 Spectral libraries

309 ▲ **CRITICAL** The provided batch files include a step of spectral library search for feature annotation (see
310 [Procedure 1 – Step 19](#)). This requires a spectral library file to be imported into MZmine (see [Procedure 1 – Step](#)
311 [18](#)). The following public spectral libraries can be freely downloaded from the MassBank of North America
312 (MoNA, <https://mona.fiehnlab.ucdavis.edu/downloads>).

- 313 ● Spectral library for Procedure 1: ‘LC–MS/MS Positive Mode’ library from MoNA.
- 314 ● Spectral library for Procedure 2: ‘GC–MS Spectra’ library from MoNA.

315 Equipment setup

316 MZmine 3 installation

- 317 ● Download and install the latest stable release of MZmine from
318 <https://github.com/mzmine/mzmine3/releases/latest>.
- 319 ● Open MZmine and set a temporary file directory to a local drive with enough free space (preferably a
320 solid-state drive). To do so, navigate to ‘Project → Set preferences → General → Temporary file
321 directory’ and browse the desired directory. Changes in the ‘Temporary file directory’ require a restart
322 of the software to take effect.
- 323 ● (Optional) Additional memory-usage options can be set as described in the online documentation.

324 (Optional) Reproducing the procedures with the ‘Batch mode’

325 In MZmine, several parameters have to be set, but only a few are crucial to tune processing for specific datasets.
326 This generally requires insights in the spectral raw data and domain-specific knowledge depending on the

327 application. We encourage first-time users to download the example datasets and use the corresponding batch
328 files to run the processing pipeline as described below.

329 ▲ **CRITICAL** The batch files provided in the present protocol are optimised for each corresponding example
330 dataset. The same batch files should not be used to process different data without adaptation, as this would
331 likely produce unreliable results. Nonetheless, they can be used as a reference and as a starting point for
332 parameter optimization.

333

- 334 ● Open MZmine and navigate to 'Project → Batch mode'. This will open the dialogue shown in Extended
335 Data Figure 1, which can be used to load, inspect, edit, and run batch files.
- 336 ● Click the 'Load' button and import the batch file corresponding to the dataset and procedure (e.g.
337 batch_procedure-1.xml for Procedure 1).
- 338 ● Select 'Replace' and click 'OK' to load the batch file. All batch processing steps are now displayed in the
339 'Batch queue' panel. Values for the individual parameters are already set. Double-click on any step in
340 the Batch queue to open the corresponding dialogue box to review and/or modify the parameters.
341 Some of these dialogue boxes offer a 'Show preview' option for interactive parameter optimization. For
342 the preview to work, data must be already imported in MZmine.
- 343 ● Double-click on the 'Import MS data' step and select the data files to import:
 - 344 ○ Browse the MS data files to process (see [Procedure 1 – Step 1](#)) using the corresponding
345 buttons. Either Individual files (i.e. 'Select files' button) or all files in a directory (i.e. 'From
346 folder' button) can be imported. Alternatively, data files can be drag-and-dropped in the 'File
347 names' panel.
 - 348 ○ Browse the spectral libraries to use for feature annotation (see [Procedure 1 – Step 18](#)).
349 Alternatively, spectral libraries can be drag-and-dropped in the 'Spectral library files' panel. If
350 spectral matching is not used for feature annotation and no spectral library is imported, make
351 sure to also remove the 'Spectral library search' step from the Batch queue.
- 352 ● Based on the batch file being used, double click on the export steps and select a directory in your
353 filesystem for the results export. One or more of the following export steps can be present:
 - 354 ○ 'Export/Submit to GNPS-FBMN': Set the 'Filename' to a suitable file path (e.g.
355 'C:\Data\project_gnps' on Windows).
 - 356 ○ 'Export for SIRIUS': Set the 'Filename' to a suitable file path (e.g. 'C:\Data\project_sirius').
 - 357 ○ 'Export for statistics (MetaboAnalyst)': Set the 'Filename' to a suitable file path (e.g.
358 'C:\Data\project_metaboanalyst').
- 359 ● Click the 'OK' button in the dialogue window to start the batch processing.

360

361 **Box 2 - The *Processing wizard***

The *Processing wizard* is a tool for the quick and user-friendly generation of data processing workflows for different MS platforms (e.g. LC-MS, GC-MS). The goal is to make the generation of processing workflows more beginner-friendly by reducing the number of parameters to set. To open the *Processing wizard*, navigate to 'Processing wizard' in the MZmine menu. The wizard is organised in the following sections: sample introduction (e.g. HPLC, MALDI), IMS, MS analyser, data acquisition (e.g. DDA, DIA). After selecting the desired instrumental setup, each section can be configured in the tabs shown in the bottom panel. Directories for input data import and output files export can also be specified. Although default parameters are provided, adjustments might be needed based on the specific user's application or instrument performance. After setting all the required parameters, click the 'Create batch' button to open and review the so-created batch file in the dialogue window (see **Extended Data Figure 1**). More information about the *Processing wizard* can be found in the [online documentation](#).

Welcome to MZmine 3 × Processing Wizard × **5. Create batch configuration**

UHPLC - - Orbitrap - - DDA - = Create batch Save presets Load presets Local presets

1. Select instrument setup

2. Overview of selected workflow

3. Tabs: Configure each part of the workflow

4. Workflow part-specific parameters

6. Produced batch configuration:

Batch queue

Local CSV d... Show

Spectral libr...

Import MS data

Mass detection

Mass detection

ADAP Chromatogram Builder

Smoothing

Local minimum feature resolver

¹³C isotope filter (formerly: isotope grouper)

Isotopic peaks finder

Join aligner

Feature list rows filter

Peak finder (multithreaded)

Duplicate peak filter

Correlation grouping (metaCorrelate)

Ion identity networking

Export molecular networking files (e.g., GNPS, FBMN, IMN, MetGem)

Export for SIRIUS

Export all annotations to CSV file

Select files Clear

In sub folders

From folder

All *.json All *.mgf

All *.msp All *.jdx

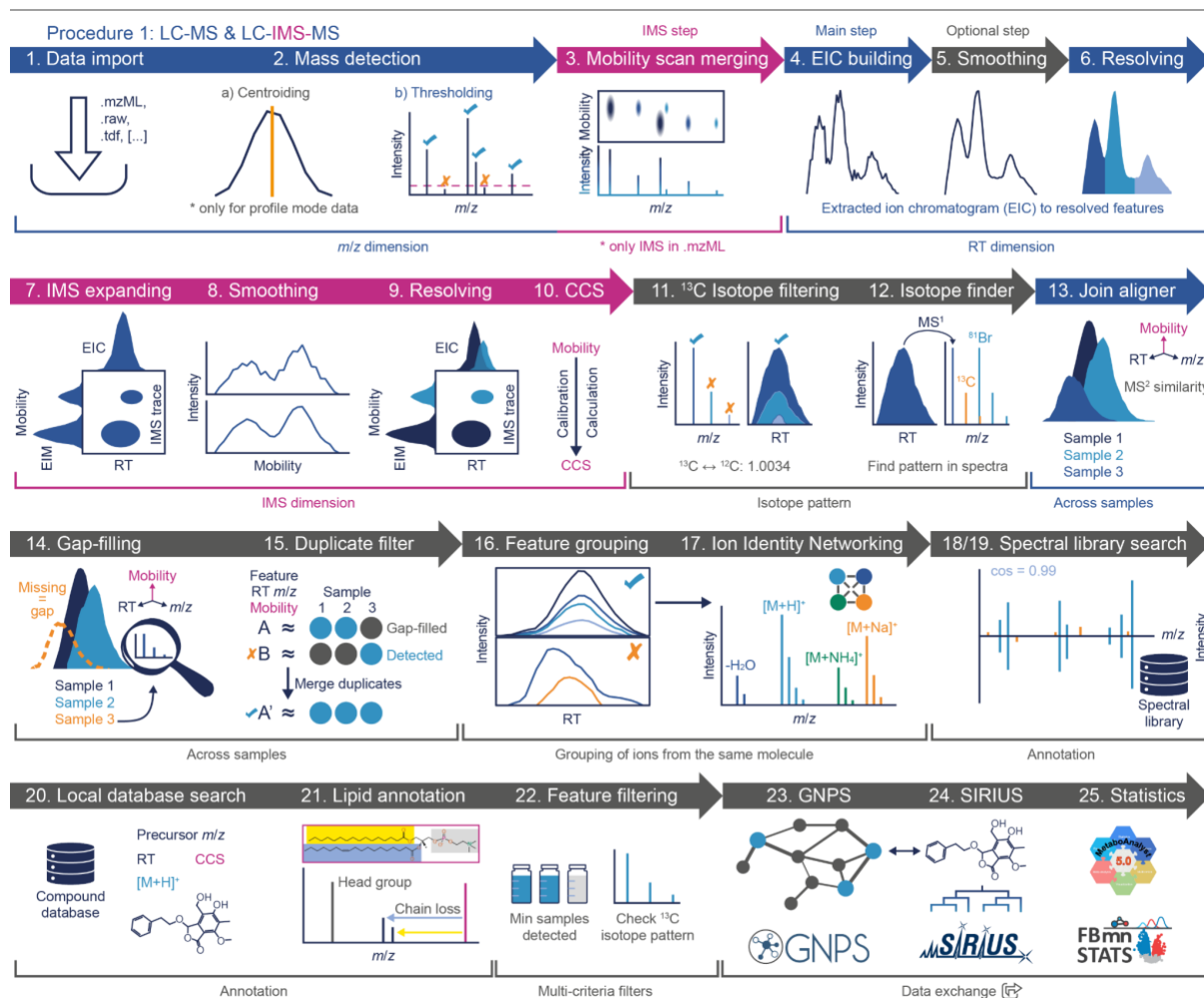
362

363

364

365 Procedure 1: LC-MS and LC-IMS-MS

366



367

368 **Figure 2: Schematic representation of the LC(-IMS)-MS workflow described in Procedure 1.** Steps strictly required for
 369 processing LC-MS data are coloured in blue. Additional steps required for LC-IMS-MS data are coloured in magenta.
 370 Optional steps for both workflows (in grey) can be applied to further refine or annotate feature lists. Steps are numbered
 371 according to the described procedure.

372 Data import and Mass detection

373 1. Import MS data

374 Raw data files can be imported into MZmine 3 by drag-and-dropping directly in the 'MS data files' tab
 375 in the main window. Another option is to use the 'Import MS data' module described step-by-step
 376 below.

- 377 ● Navigate to 'Raw data methods → Raw data import → MS data (advanced)'
- 378 ● Click on 'Select file' to directly browse the data files in your filesystem. Alternatively, all data
 379 files of a specific format can be imported from a directory using the corresponding button (e.g.
 380 'All *.mzML', 'All *.raw').
- 381 ● Disable the 'Advanced import' checkbox. This option is only needed for very large datasets
 382 (see **Processing large datasets** section).

- 383
- (Optional) ‘**Spectral library files**’ can be imported in a way similar to that described in step b. Spectral libraries are needed to perform spectral library search (see [Procedure 1 – Step 19](#)). If this is not the case, clear any text in the ‘Spectral library files’ panel.
- 384
- 385
- 386

387 2. Mass detection

388 Each mass spectrum in the data files is processed individually and stored as a list of m/z and intensity
389 pairs (called *mass list*), which is readily usable by the software for the subsequent *feature detection*
390 steps. An intensity threshold (i.e. ‘Noise level’) can also be set to exclude low-intensity signals (e.g.
391 electronic and/or chemical noise) from further processing.

392 ▲ **CRITICAL** The noise level in the spectra can vary greatly due to a number of reasons, such as the type
393 of the mass analyser and specific acquisition settings. For this reason, MZmine provides an interactive
394 visualisation panel to help the user optimise this step (see **Box 3**).

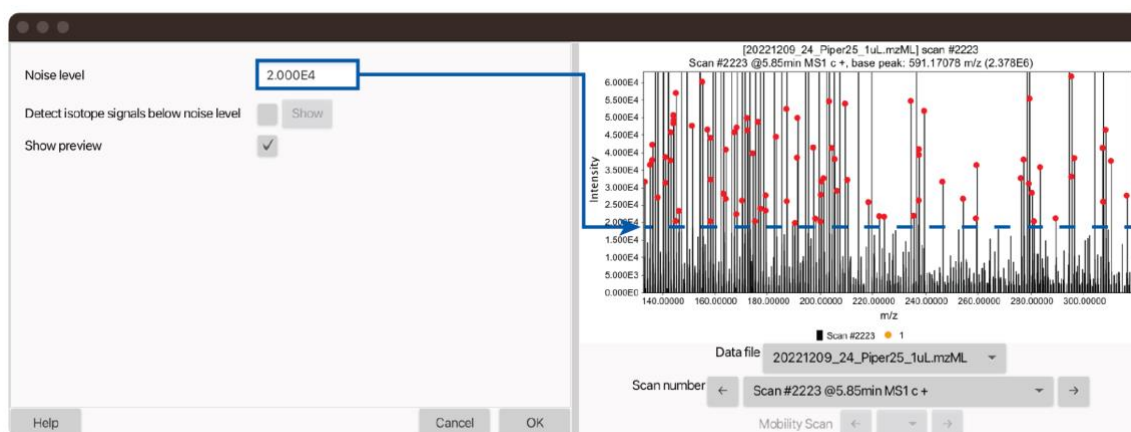
- Navigate to ‘**Raw data methods** → **Mass detection** → **Mass detection**’
 - Specify the ‘**Raw data files**’ to process (e.g. ‘All raw data files’ or ‘As selected in main window’).
 - Set the ‘**Scans**’ filters. Use the ‘Set filters’ button to run the mass detection on separate *MS levels* (see **Table 1**). When MS^2 spectra are collected (e.g. LC–MS/MS), the mass detection should be run twice: once for ‘MS level = 1’ and once for ‘MS level = 2’ (see for example batch file ‘*batch_procedure-1.xml*’).
 - Select the ‘**Mass detector**’ algorithm from the drop-down menu. Five *mass detection* algorithms are available. The choice of the algorithm depends on the raw data characteristics (e.g. profile vs centroid data, low vs high MS resolution). A detailed description of each mass detector option can be found in the [online documentation](#). We recommend the ‘Centroid’ mass detector for *centroid* data and the ‘Exact mass’ algorithm for *profile* high-resolution MS data.
 - Click the ‘Setup’ to set the ‘**Noise level**’. This will exclude the m/z signals below the specified intensity threshold from further processing. An interactive visualisation panel can be opened by ticking the ‘Show preview’ checkbox (see **Box 3**).
 - (Only IMS data) Specify the ‘**Scan types (IMS)**’ to be processed. This parameter only applies to IMS data and determines whether *mobility scans*, *frame spectra*, or both should be processed. Since *frame spectra* are obtained by merging multiple *mobility scans* (see **Table 1**), we recommend applying two different noise levels. This can be done by running the mass detection on ‘Mobility scans only’ and ‘Frames only’ separately (see e.g. batch file ‘*batch_procedure-1.xml*’).
- 395
- 396
- 397
- 398
- 399
- 400
- 401
- 402
- 403
- 404
- 405
- 406
- 407
- 408
- 409
- 410
- 411
- 412
- 413
- 414
- 415

416 **Box 3 – Setting the noise level**

MS background noise generally refers to non-specific m/z signals detected in the absence of a specific analyte. Such noise can arise for a number of reasons (e.g. electronic noise) and is characterised by several low-intensity signals with no clear pattern among them (often referred to as ‘grass’ in the mass spectra). Filtering out noise from raw spectral data prior to *feature detection* prevents a large number of irrelevant background signals from being retained as false *features*, which may complicate downstream processing steps. This also reduces memory consumption and computing time, especially when processing large datasets. Because the magnitude of background noise can vary greatly between different datasets, the best way to optimise this

parameter is by visually inspecting the raw spectra. To this end, an interactive visualisation panel can be opened directly from the 'Mass detection' dialogue window (see [Step 2](#)). Two drop-down menus can be used to select, respectively, the data file and spectrum to display (data needs to be imported first, see [Step 1](#)). Once a noise level is set, m/z signals above the threshold are automatically labelled with red dots. All unlabelled signals will be excluded from further processing. Ideally, the noise level should be set right above the 'grass' noise (see figure). Nevertheless, higher noise levels can be used to save computation time/cost when processing large datasets consisting, for example, of hundreds to thousands of data files.

In chromatography-based experiments, the noise level in the spectra can vary across the same chromatographic run. For instance, greater noise is often observed towards the end of GC-MS runs due to increased column bleeding. In these situations, different noise levels can be applied to different RT ranges using the 'Scans' filters. Note that the *mass detection* can be run on one RT segment at a time, therefore multiple 'Mass detection' calls are needed, one for each RT segment (see [Procedure 2 - Step 2](#)).



417

418

3. (Only IMS data in .mzML format) Mobility scan merging

419

In this step, *mobility scans* are merged to create *frame spectra* (see [Table 1](#)). This is required only when processing IMS data in .mzML format. When processing data files in the native Bruker format (.d and .tdf), *mobility scans* are merged in the background during the data import. If you are using the example dataset provided for this procedure, you can skip this step and go to [Step 4](#).

420

421

422

423

- Navigate to 'Raw data methods → File merging → Mobility scan merging';

424

- Specify the 'Raw data files' to process;

425

- Select the 'Show preview' checkbox and choose a frame to preview.

426

- Set the 'Noise level' to 0, to deactivate thresholding at this step. Similarly to [Step 2](#), a noise level can be set and will be applied to the merged *frame spectra*.

427

- Select the 'Merging type'. This determines how the signal intensities are calculated in the merged *frame spectra*. The 'Summed' option (recommended), sums the intensities for the same m/z value detected across the *mobility scans* being merged.

428

429

430

- Click the 'Clear filters' button for the 'Scans' filters, since the noise level was set for MS level 1 and MS level 2 separately.

431

432

- Choose the ' m/z weighting' method. This parameter determines how the m/z values are calculated in the merged *frame spectrum*. The 'Linear' method (recommended) attributes larger weight to more intense signals;

433

434

435

- 436
- 437
- 438
- Set the scan-to-scan '**m/z tolerance**'. This is the maximum allowed deviation for an *m/z* value detected across consecutive *mobility scans* to be considered the same. We recommend 0.005 *m/z* or 15 ppm as a starting point for most time-of-flight (TOF) instruments.

439 **Chromatogram building and resolving**

440 The following steps of chromatogram building and resolving require insight into the width, height, and number
441 of data points of chromatographic peaks, which depend on the instrument and the LC–MS method. Raw data
442 can be explored using the 'Raw data overview' module (see **Extended Data Figure 2**). See [online documentation](#)
443 for more details.

444

445 **4. EIC building with 'ADAP chromatogram builder'**

446 This step builds an EIC for each *m/z* value detected over a minimum number of consecutive MS¹ scans
447 in the LC–MS run. EICs matching a set of user-defined requirements (e.g. minimum number of data
448 points and intensity) are stored as features in a *feature list*. Although various EIC building algorithms
449 are available, we recommend using the *ADAP chromatogram builder module*. A detailed description of
450 this module is provided in the [online documentation](#).

- 451
- Navigate to '**Feature detection → LC-MS → ADAP chromatogram builder**';
 - Specify the '**Raw data files**' to process;
 - Set the '**Scan filters**'. Enable the checkbox, click the 'Show' button, and set the 'MS level filter' as 'MS1, level = 1'.
 - Set the '**Minimum consecutive scans**' as the minimum number of consecutive MS¹ scans where an *m/z* must be detected above a certain intensity (see the next parameter) for the corresponding EIC to be considered valid. This parameter largely depends on the MS acquisition settings used during the analysis. Usually, no less than 3–5 should be used, as lower values would produce false features.
 - Set the '**Minimum intensity for consecutive scans**' as the minimum intensity an *m/z* must exceed in consecutive MS¹ scans (see the previous parameter) for the corresponding EIC to be considered valid. A good starting point is 1–3 times the 'Noise level' used for the MS¹ level in the *Mass detection* ([Step 2](#)). If LC–IMS–MS data are being processed, consider the noise level applied to *frame scans*.
 - Set the '**Minimum absolute height**' as the minimum intensity the highest data point in the EIC must exceed for the corresponding EIC to be considered valid. A good starting point is 3–10 times the 'Noise level' used for the MS¹ level in the *Mass detection* ([Step 2](#)). If LC–IMS–MS data are being processed, consider the noise level applied to *frame scans*.
 - Set the '**m/z tolerance (scan-to-scan)**'. This is the maximum allowed *m/z* deviation between consecutive scans in the EIC. This parameter largely depends on the MS analyser type and performance. A good starting point is '0.003 *m/z* or 5 ppm' for Orbitrap instruments and '0.005 *m/z* or 15 ppm' for TOF devices.

473 ▲ **CRITICAL** The *m/z* tolerances must be specified as both an absolute value (in *m/z*) and
474 relative value (in ppm). The tolerance for each *m/z* value is calculated using the maximum of
475 the absolute and relative tolerances.

- 476
- Provide a '**Suffix**' (e.g. '_eic') to name the newly-created feature lists. This option is present in most of the modules described below. We recommend using a different suffix for each module to easily recognize the features lists produced by each processing step.
- 477
- 478

479

480 5. (Optional) Chromatogram smoothing

481 We recommend applying smoothing to EICs only if they exhibit a 'jagged' profile (i.e. large intensity
482 fluctuations of consecutive data points). Jagged EICs may cause inaccurate peak integration and
483 erroneous splitting of peaks into multiple features during the EIC resolving step (see [Step 6](#)). On the
484 other hand, excessive smoothing can lead to peak shape distortion and artefacts. For this reason, we
485 recommend using the 'Show preview' option to evaluate the effect of the chosen smoothing
486 parameters.

- 487 ● Navigate to '**Feature detection → Smoothing**';
- 488 ● Specify the '**Feature lists**' to process. When running modules individually, various options are
489 available (e.g. 'As selected in the main window', 'Feature list name pattern'). When using the
490 batch mode (see **Reproducing the procedures with the 'Batch mode'** section), the option
491 'Those created by previous batch step' must be selected.
- 492 ● Choose the '**Smoothing algorithm**'. We recommend using 'Savitzky Golay'.
- 493 ● Click the '**Setup**' button:
 - 494 ○ Tick the '**Retention time smoothing**' checkbox.
 - 495 ○ Set the number of data points to use for smoothing. We recommend using half the
496 number of data points of a chromatographic peak.
- 497 ● Tick the '**Show preview**' checkbox to open an interactive visualisation panel to help adjust the
498 smoothing parameters. Use the drop-down menus to select, respectively, the feature list and
499 feature to display. We recommend choosing a medium-intensity EIC trace that well represents
500 the 'jaggedness' in the data.
 - 501 ○ ▲ **CRITICAL** When changing the smoothing parameters, the preview does not
502 automatically update. It is necessary to select a new feature from the drop-down
503 menu to visualise the newly-set parameters.
- 504 ● Specify how to handle the '**Original feature list**'. This option determines whether to 'KEEP' in
505 memory or 'REMOVE' the input feature list(s) once the processing is completed. We
506 recommend using the 'KEEP' option during parameter optimization.
- 507 ● Provide a '**Suffix**' (e.g. '_RT-smooth') to name the newly-created feature lists.

508

509 6. EIC resolving with the 'Local minimum resolver'

510 The EIC traces built in the previous steps are stored in a *feature list* per sample. EICs might contain
511 multiple chromatographically separated peaks that need to be resolved into individual features.
512 Although various EIC resolving algorithms are available, we recommend using the *Local minimum*
513 *resolver* module when processing LC data. Refer to **Box 4** for a more detailed description of the
514 optimization of the Local minimum resolver. A detailed description of all feature resolvers is provided
515 in the [online documentation](#).

- 516 ● Navigate to '**Feature detection → Chromatogram resolving → Local minimum resolver**';
- 517 ● Specify the '**Feature lists**' to process.
- 518 ● Specify how to handle the '**Original feature list**'.

519 ● Enable the '**MS/MS scan pairing**' option. This will pair each resolved feature to the
520 corresponding MS² fragmentation spectrum (collected in DDA mode), based on the RT offset
521 between the chromatographic peak and the moment the MS² was triggered during the run.
522 Click the 'Show' button and set the following parameters:

523 ○ Set the '**MS1 to MS2 precursor tolerance (m/z)**' as the maximum allowed deviation
524 between the *m/z* associated with the feature, and the precursor *m/z* the MS² was
525 triggered for. As a starting point, the same *m/z* tolerance set in the chromatogram
526 building step ([Step 4](#)) can be used.

527 ○ Set the '**Retention time filter**' as 'Use feature edges'. This option pairs a feature with
528 the corresponding MS² spectrum only if the latter was triggered within the feature's
529 RT range. The 'Use tolerance' option uses a fixed tolerance between the feature's
530 peak apex and the RT of the MS² scan.

531 ○ (Optional) Enable and set a '**Minimum relative feature height**' to limit the pairing of
532 an MS² scan with multiple features. When an MS² scan can be paired with multiple
533 features within the specified tolerances, only those with intensity above X% of the
534 most intense feature will be considered. When enabled, the default value (25%)
535 should provide good results for most applications

536 ○ Disable the '**Minimum required signals**'. This parameter is designed to remove
537 empty MS² scans in spectral library building workflows.

538 ○ Ignore all the remaining parameters at this stage. These are related to IMS and have
539 no effect on the 'Retention time' dimension. They will be discussed in the **IMS**
540 **expanding and resolving** section.

541 ● Select '**Dimension → Retention time**';

542 ● Enable the '**Show preview**' option to open an interactive visualisation panel to help adjust the
543 resolving parameters (see **Box 4**)

544 ● Set the '**Chromatographic threshold**'. This parameter represents an important filter for
545 chromatographic noise (e.g. solvent background contaminants). Briefly, the X% least-intense
546 data points from the whole chromatogram are removed before the resolving. For LC-(IMS-
547)MS data, we recommend using a value between 50% and 90%.

548 ● Set the '**Minimum search range RT/Mobility**'. This is the RT window used for local minimum
549 search. A good starting point is the full-width at half maximum (FWHM, expressed in minutes)
550 of a typical chromatographic peak in the data.

551 ▲ **CRITICAL** The optimal RT search range mainly depends on the chromatographic system
552 setup and MS acquisition settings. An overly narrow RT search range can cause peak edges to
553 be truncated. Conversely, an overly wide search range might lead to the incomplete resolution
554 of closely-eluting peaks.

555 ● Set the '**Minimum relative height**' to 0 to ignore this parameter. We recommend not to use
556 this parameter, because modern MS analysers offer dynamic ranges spanning several orders
557 of magnitude.

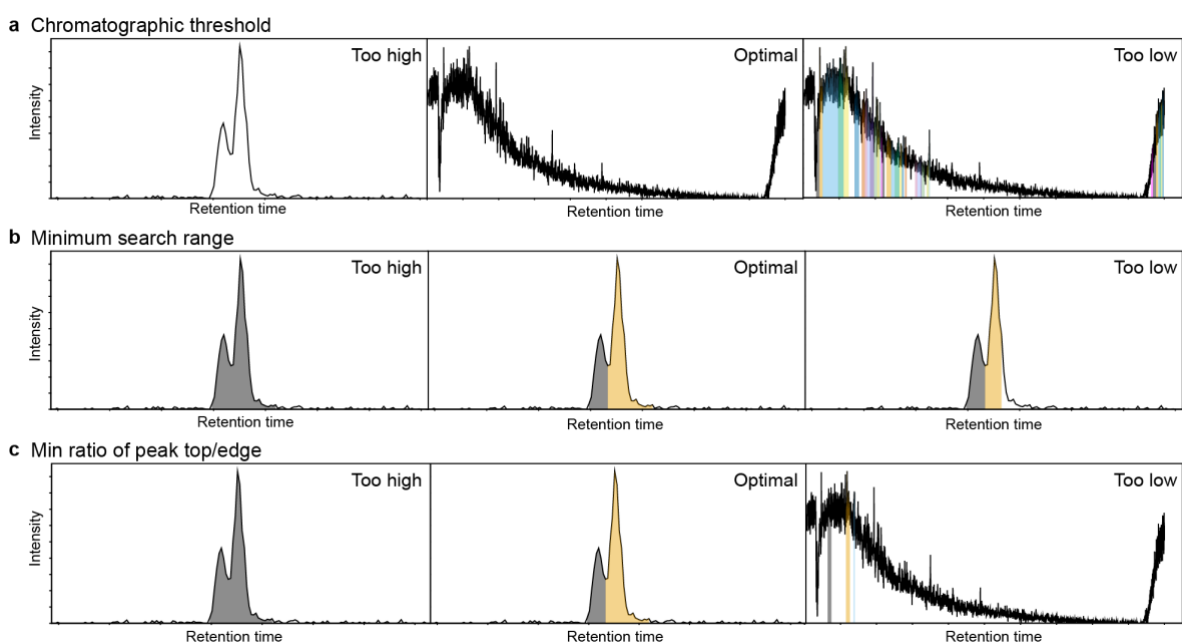
558 ● Set the '**Minimum absolute height**' a peak needs to reach to be retained as a feature after the
559 resolving procedure. We recommend using the same value as used in [Step 4](#).

560 ● Set the '**Min ratio of peak top/edge**'. This is the minimum ratio between the intensities of the
561 highest point (apex) and edges of a peak to be retained as a feature after the resolving step
562 (i.e. the peak apex must be X times more intense than the peak edges). We recommend setting
563 2 as a starting value.

- 564
- 565
- 566
- 567
- 568
- 569
- 570
- 571
- 572
- 573
- Set the **'Minimum scans (data points)'**. This parameter is very similar to the 'Minimum consecutive scans' settings in the *ADAP chromatogram builder* module ([Step 4](#)); therefore, the same recommendations can be followed.
 - Define the allowed **'Peak duration range'**. This is the acceptable duration of a chromatographic peak to be retained as a feature after the resolving. This parameter can be used to filter out noisy features based on their overly short, or long, duration. We recommend setting the minimum duration to 0 (the previous parameter already defines a minimum peak duration in terms of data points) and not be too strict when setting the maximum duration, because tailing chromatographic peaks can also be discarded.
 - Provide a **'Suffix'** (e.g. `'_RT-res'`) to name the newly-created feature lists.

574 **Box 4 – Optimise feature resolving**

The EIC resolving step plays a crucial role in the *feature detection* of chromatography–MS data. The goal is to split multiple peaks that are resolved or partially co-eluting in the EIC traces into individual *features* (i.e. *chromatographic resolving*). The *Local minimum resolver* module (recommended for LC–MS) assumes that a local minimum in an EIC trace corresponds to the valley between two adjacent peaks and uses it to split fully resolved or 'shoulder' peaks. Thorough optimization of the algorithm parameters is crucial to ensure reproducible detection of true *features* across all samples while minimising 'noisy' peaks to be retained as features. Particular attention should be given to the 'Chromatographic threshold', 'Minimum search range RT', and 'Min ratio of peak top/edge' parameters. Their effect on the EIC resolving results is illustrated in the figure below, and a more detailed description can be found in the [online documentation](#).



MZmine provides a preview panel (see [Step 6](#)) to interactively display the effect of the selected parameters on the detected features (**Extended Data Figure 3**). We recommend using this option for understanding and fine-tuning the algorithm, especially when working with a new dataset. Once opened, select the *feature list* and EIC traces to display from the corresponding drop-down menus. Two EIC traces are displayed in the two sub-panels. MZmine automatically tries to select a 'noisy' and 'good' EIC trace in the upper and lower panel,

respectively, based on the height-to-area ratio. The detected *features* are colour-filled. Resolved peaks are shown in different colours. During the optimization, all clear chromatographic peaks in the lower panel should be fully retained, while as few *features* as possible should be detected in the upper panel.

575 **IMS expanding and resolving (only for IMS data)**

576 At this stage, only the merged *frame spectra* have been examined for the detection of *RT-resolved features*,
577 while the IMS dimension has not yet been considered. In the next three steps (i.e. Steps 7–9), the individual
578 *mobility scans* are inspected to create *IMS-resolved features*. Similar to chromatogram building and resolving,
579 we encourage the user to explore the raw data and gain the necessary insights to choose the optimal processing
580 parameters. This can be done using the ‘Ion mobility raw data overview’ module (see **Extended Data Figure 4**).

581 ▲ **CRITICAL** Steps 7–9 are only required when processing LC–IMS–MS data. If you are processing LC–MS data,
582 skip these steps and go to [Step 11](#).

583

584 **7. (Only IMS data) Expanding features to the IMS dimension**

585 In this step, MZmine takes the *m/z* associated to every RT-resolved *feature* and searches the individual
586 *mobility scans* for signals to build the corresponding extracted ion mobilogram (EIM). A detailed
587 description of the ‘IMS expander module’ is provided in the [online documentation](#).

- 588 ● Navigate to ‘**Feature detection → LC–IMS–MS → IMS Expander**’.
- 589 ● Enable and set the ‘**m/z tolerance**’. This is the maximum allowed deviation between the *m/z*
590 of the *RT-resolved features* and the *m/z* signals in the individual *mobility scans*. In contrast to
591 the chromatogram building and resolving steps, a higher *m/z* tolerance might be needed; this
592 is because the mass accuracy in individual *mobility scans* tends to be lower compared to the
593 *frame spectra*. We recommend 0.005 *m/z* or 15 ppm as a starting point for most TOF–MS
594 instruments.
- 595 ● (Optional) ‘**Raw data instead of thresholded**’. When enabled, this option replaces the *noise*
596 *level* used in the *mass detection* with the provided intensity threshold.
- 597 ● Disable ‘**Override default mobility bin width (scans)**’ to use MZmine’s default binning of
598 mobility scans. This option is usually not required for general untargeted analysis.
- 599 ● Disable ‘**Maximum features per thread**’. When enabled, this option controls thread
600 parallelization, which affects RAM consumption and processing time. It is usually not needed
601 for processing LC–IMS–MS data.

602 **TROUBLESHOOTING**

603 **8. (Only IMS data, optional) Mobilogram smoothing**

604 Similar to [Step 5](#), we recommend applying smoothing to EIMs only if they exhibit a jagged profile, as
605 this may cause inaccurate resolving of mobility features (see [Step 9](#)). Since this is often the case for LC–
606 IMS–MS data, we generally recommend performing this step.

- 607 ● Navigate to ‘**Feature detection → Smoothing**’;
- 608 ● Specify the ‘**Feature lists**’ to process (see [Step 5](#)).
- 609 ● Choose the ‘**Smoothing algorithm**’. We recommend using ‘Savitzky Golay’.
- 610 ● Click the ‘**Setup**’ button:

- 611 i. Tick the **'Mobility smoothing'** checkbox.
- 612 ii. Set the number of data points to use for smoothing. We recommend using half the
- 613 number of data points of a mobility peak.
- 614 ● Tick the **'Show preview'** checkbox and select 'Mobility' as preview dimension to open an
- 615 interactive visualisation panel. Use the drop-down menus to select, respectively, the feature
- 616 list and feature to display. We recommend choosing a medium-intensity EIC trace that well
- 617 represents the 'jaggedness' in the data.
- 618 ▲ **CRITICAL** When changing the smoothing parameters, the preview does not update
- 619 automatically. It is necessary to select a new feature from the drop-down menu to visualise
- 620 the newly-set parameters.
- 621 ● Specify how to handle the **'Original feature list'** (see [Step 5](#)).
- 622 ● Provide a **'Suffix'** to name the newly created feature lists (e.g. '_IMS-smooth').
- 623

624 9. (Only IMS data) Mobilogram resolving

625 Similar to what was described in the **Chromatogram building and resolving** section, the EIM traces built

626 in the [Step 7](#) have to be split into individual mobility peaks using a resolving algorithm. We recommend

627 using the *Local minimum resolver* module. Since the parameters of this module are already described

628 in [Step 6](#), we focus on the differences between EIMs and EICs resolving here.

- 629 ● Navigate to **'Feature detection → Chromatogram resolving → Local minimum resolver'**.
- 630 ● Select **'Dimension → Mobility'**
- 631 ● Enable the **'MS/MS scan pairing'** option and proceed as described in **Step 6**, while considering
- 632 the following IMS-related parameters:
- 633 i. Enable the **'Limit by ion mobility edges'** option. This option pairs a mobility-resolved
- 634 feature with the corresponding MS² spectrum only if the latter was triggered within
- 635 the feature's mobility range (see the [online documentation](#) for more details).
- 636 ii. (Optional) **'Merge MS/MS spectra (TIMS)'**. This option only applies to fragmentation
- 637 MS² spectra acquired in PASEF mode. When enabled, multiple MS² spectra acquired
- 638 for the same precursor *m/z* and associated to the same feature are merged into a
- 639 single spectrum. We recommend enabling this option when low-abundant
- 640 compounds are of interest.
- 641 iii. (Optional) Enable and set the **'Minimum signal intensity (absolute and relative)'**.
- 642 When the **'Merge MS/MS spectra (TIMS)'** option is enabled, these two thresholds can
- 643 be used to remove low-intensity signals from the merged MS² spectra (see previous
- 644 parameter).
- 645 ● Set the **'Chromatographic threshold'**. A lower value compared to the EIC resolving step should
- 646 be used because mobility peaks are generally wider and less resolved than LC peaks. We
- 647 recommend using a value between 35% and 70%.
- 648 ● Adjust the **'Minimum search range RT/Mobility'** parameter. This is the mobility window used
- 649 for the local minimum search. We recommend starting the optimization at 0.005 for TIMS, 0.5
- 650 for travelling wave-IMS (TWIMS), and 1 for drift time-IMS (DTIMS) devices.
- 651 ▲ **CRITICAL** The optimal mobility search range mainly depends on the IMS unit and scale
- 652 employed by the instrument used for the analysis. For example, TIMS devices measure the ion

653 mobility in Vs/cm² (typically between 0.5 and 2.0), whereas time-dispersive instruments
654 (TWIMS and DTIMS) use milliseconds (typically between 0 and 90).

- 655 ● Provide a **'Suffix'** (e.g. **'_IM-res'**) to name the newly created feature lists.
- 656 ● All other parameters can be optimised as described in [Step 6](#).

657

658 10. (Only IMS data, optional) CCS calibration and calculation

659 MZmine supports two methods for CCS calibration. Either an external calibration file is imported, or a
660 calibration can be calculated using a list of reference compounds. The external calibration is described
661 below, whereas the calibration method based on reference compounds is covered in the
662 [documentation](#). After applying the vendor calibration software, the Agilent raw data folder contains a
663 'OverridelmsCal.xml', whilst Waters raw data contains a 'mob_cal.csv' file. Bruker raw data is
664 automatically calibrated during the import from *.tdf* raw files and does not require this step.

- 665 ● Navigate to **'Feature list methods → Processing → External CCS calibration'**.
- 666 ● Specify the **'Raw files'** to process
- 667 ● Select the external **'Calibration file'**.
- 668 ● Click the **'OK'** button. After applying the calibration, the CCS values are automatically
669 calculated in the [Isotope pattern finder](#) step.

670 Isotope filtering

671 During the EIC building, EIC traces are constructed for all the *m/z* signals detected during the *mass detection*. As
672 a consequence, signals generated by isotopologues of the same chemical entity produce multiple *features* in the
673 *feature list*, which constitutes redundant information for downstream data analysis. This is a common issue for
674 C-containing molecules, where the ¹³C isotope signal is easily detected. At the same time, the isotopic pattern
675 holds essential information for the purpose of feature annotation.

676 11. ¹³C isotope filter

677 This module removes ¹³C-related features from the processed *feature lists* and assigns the retrieved ¹³C
678 isotopic pattern to the monoisotopic peak.

679 ▲ **CRITICAL** This module removes *features* matching the filtering criteria from the *feature lists* being
680 processed. This also means that false ¹³C-related features can be erroneously discarded. Therefore, we
681 recommend using fairly strict tolerances, based on the instrument performance, to reduce such a risk.

- 682 ● Navigate to **'Feature list methods → Isotopes → 13C isotope filter'**.
- 683 ● Specify the **'Feature lists'** to process (see [Step 5](#)).
- 684 ● Specify how to handle the **'Original feature list'** (see [Step 5](#)).
- 685 ● Set the **'m/z tolerance (intra-sample)'**. This is the maximum allowed *m/z* difference between
686 the examined feature and its potential ¹³C-isotopologues in the feature list. We recommend
687 using a fairly strict tolerance, based on the MS analyser performance.
- 688 ● Set the **'Retention time tolerance'**. This is the maximum allowed RT deviation between
689 potential ¹³C-related features. Because isotopologues should produce fully overlapping
690 chromatographic peaks, a strict tolerance can be used.
- 691 ● (Only IMS data) Enable and set the **'Mobility tolerance'**. This is the maximum allowed IMS
692 deviation between potential ¹³C-related features. Here too, a strict mobility tolerance can be

- 693 used since isotopologues should undergo identical IMS separation. We recommend 0.008 for
694 TIMS and 0.5 for TWIMS and DTIMS.
- 695 ● Tick the **'Monotonic shape'** checkbox to filter ¹³C-related features only when the retrieved
696 isotopic pattern has a monotonically decreasing trend (typical for small molecules). For small
697 molecule applications, we recommend enabling this option.
 - 698 ● Set the **'Maximum charge'** state to be considered when calculating the *m/z* of ¹³C isotopes.
699 For small molecules applications, we recommend to use 1 or 2;
 - 700 ● Set the **'Representative isotope'** as 'most intense';
 - 701 ● Tick the **'Never remove features with MS²'** checkbox to avoid filtering ¹³C-related features for
702 which an MS² scan has been acquired (even though they match the filtering criteria).
 - 703 ● Provide a **'Suffix'** (e.g. '_deiso') to name the newly-created feature lists.

704 11. Isotope pattern finder

705 This module searches and annotates potential isotope patterns for each feature based on its *m/z* and a
706 list of chemical elements specified by the users. According to the retrieved isotope pattern, a charge
707 state is also assigned. Unlike the 'Isotope filter', this module does not remove isotopic features, but
708 annotates them as part of an isotopic pattern.

- 709 ● Navigate to **'Feature list methods → Isotopes → Isotope pattern finder'**.
- 710 ● Click the 'Setup' button and select the **'Chemical elements'** to consider for the isotope search;
- 711 ● Set the **'*m/z* tolerance (feature-to-scan)'**. This is the maximum allowed *m/z* difference
712 between the examined feature and its potential isotopologues. Since this module uses the raw
713 data to find potential isotope signals, a slightly wider *m/z* tolerance may be appropriate
714 compared to the [¹³C isotope filter](#).
- 715 ● Set the **'Maximum charge of isotope *m/z*'** to be considered when calculating the isotopes'
716 *m/z*. For small molecules applications, we recommend using 1 or 2.
- 717 ● Set the **'Search in scans'** as 'Single most intense'.

718

719 Alignment and gap-filling across samples

720 Any untargeted MS experiment performed on multiple samples aims at comparing, qualitatively or
721 quantitatively, the analytes detected across the set of analysed samples. However, chromatography-MS
722 experiments are subject to instrumental drift that produces fluctuations in RT, ion mobility, and mass accuracy
723 over the course of the analysis. As a consequence, the same analyte is almost never detected with the same RT,
724 *m/z* and mobility over consecutive LC-(IMS)-MS runs. The goal of *feature alignment* is to account for such
725 variations and align the *features* corresponding to the same molecular entity across different instrument runs.⁴⁴
726 By doing so, *feature lists* from multiple samples can be merged into a single, *aligned feature table*.

727 12. Join aligner

728 In MZmine, the *feature alignment* is based on alignment scores calculated using a combination of user-
729 defined tolerances and weights for each available analysis dimensions (i.e. *m/z*, RT, and ion mobility;
730 see **Box 5** for more details). Although various alignment algorithms are available, we recommend using
731 the 'Join aligner' module for LC-(IMS)-MS data. A more detailed description of this module is provided
732 in the [online documentation](#).

- 733 ● Navigate to **'Feature list methods → Alignment → Join aligner'**

- 734
- Specify the ‘**Feature lists**’ to process (see [Step 5](#)).
- 735
- Set an ‘**m/z tolerance (sample-to-sample)**’. This is the maximum allowed *m/z* deviation
- 736
- 737
- 738
- 739
- Set a ‘**Retention time tolerance**’. This is the maximum allowed RT deviation between the
- 740
- 741
- 742
- (Only IMS data) Enable and set a ‘**mobility tolerance**’. This is the maximum allowed mobility
- 743
- 744
- Set the ‘**Weight for m/z**’, ‘**Weight for RT**’ and ‘**Mobility weight**’. These weights define the
- 745
- 746
- 747
- 748
- Disable all other remaining options. They can be used in particular applications that require
- 749
- 750
- Provide a ‘**Feature list name**’ to name the newly-created aligned feature lists.

751 **Box 5 – Feature alignment**

During the alignment, multiple *features* can fall within one, or more, tolerance windows set for each analysis dimension (i.e. RT, *m/z* and mobility). The best alignment match is chosen using a weighted scoring system that considers all the available analysis dimensions to assign a global alignment score. Alignment scores are calculated using the following equation:

$$score_p = \sum_{dim} \left(1 - \frac{\Delta_{dim}}{tolerance_{dim}} \right) \cdot weight_{dim}$$

where:

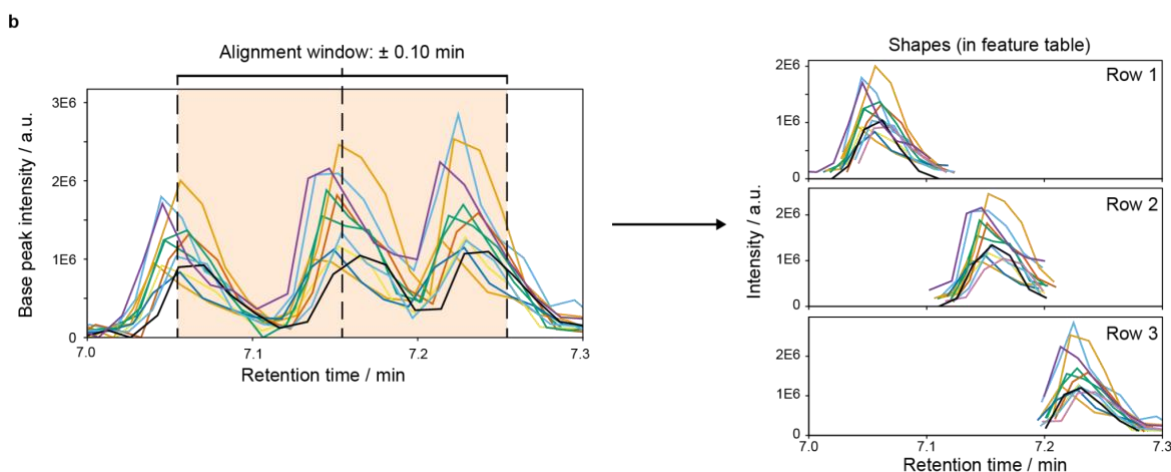
- Δ_{dim} is the difference in the value of the considered dimension (e.g. $dim = RT$) between the features being aligned. Perfectly matching RTs will produce a $\Delta_{RT} = 0$;
- $tolerance_{dim}$ is the maximum sample-to-sample allowed deviation for the considered dimension;
- $weight_{dim}$ is the importance given to each dimension in the calculation of the global alignment score;
- $score_p$ is the global alignment score for the features being aligned. It is obtained by summing the weighted score of each dimension.

Unlike other steps (e.g. [Chromatogram building and resolving](#)), no ‘Show preview’ option is available to interactively assess the alignment quality. Therefore, the alignment results can only be evaluated *a posteriori*. To do so, a set of metrics can be displayed for each feature in the *aligned feature list* (see orange columns in the figure). These include the ‘Aligned features’ (i.e. number of aligned samples), ‘Rate’ (i.e. ratio between the number of aligned and total samples), ‘Extra features’ (i.e. number of other possible alignment matches within the defined tolerances), ‘weighted distance score’ (i.e. reflects the alignment

$score_p$) and the average difference between the value of each dimension pre- and post-alignment (i.e. ' $\Delta m/z$ ', ' ΔRT ', and ' $\Delta Mobility$ ' columns).

a

ID	RT	m/z	Height	Area	Alignment (13 samples)						
					Rate	Aligned features	Σ Extra features	Weighted distance score	$\Delta m/z$ ppm	$\Delta m/z$	ΔRT
1	7.05	293.1860	2.0E6	9.1E4	0.85	11	9	0.96	0.23	0.0001	0.008
2	7.15	293.1860	2.5E6	1.3E5	0.85	11	17	0.97	0.26	0.0001	0.005
3	7.23	293.1860	2.8E6	1.5E5	0.85	11	11	0.96	0.34	0.0001	0.007



752

753 13. (Optional) Gap-filling

754 One of the main challenges in untargeted feature detection is reproducible detection of low-intensity
 755 *features*. In fact, these can be erroneously filtered out at various stages of the workflow, producing
 756 false missing values (i.e. gaps) in the *aligned feature table*. MZmine offers the possibility to re-inspect
 757 such gaps by checking for the presence of omitted signals in the original raw data. If a relevant signal is
 758 found, it is integrated and re-included in the *feature table*, thus filling the gap. In MZmine, this process
 759 is called 'gap filling' and can be performed using the *Peak finder* module. During the gap-filling, artefacts
 760 (duplicate features) can be introduced in the feature table in case of misaligned features (see the [online](#)
 761 [documentation](#) for more information). Such artefacts can be removed later using the 'Duplicate filter'
 762 module ([Step 15](#)).

- 763
- 764 • Navigate to 'Feature list methods → Gap filling → Peak finder'.
 - 765 • Specify the 'Feature lists' to process (see [Step 5](#)).
 - 766 • Set the 'Intensity tolerance'. Maximum allowed intensity deviation between consecutive
 767 scans when building the EIC for gap-filled features. A higher tolerance will retain more 'jagged'
 EICs during gap-filling. We recommend 20% as a starting point.
 - 768 • Set the 'm/z tolerance sample-to-sample'. Maximum allowed m/z deviation between gap-
 769 filled signal in the raw data and the feature's m/z in the *feature table*. It is a sample-to-sample
 770 tolerance and the same recommendation provided in [Step 13](#) can be followed.
 - 771 • Set the 'Retention time tolerance (sample-to-sample)'. This is the RT window (around the
 772 feature's RT) used to examine the raw data for gap filling. It is a sample-to-sample tolerance,
 773 therefore, the same recommendation provided in [Step 13](#) step can be followed.

- 774
- 775
- 776
- 777
- Set the **'Minimum scans (data points)'**. This parameter is very similar to the 'Minimum consecutive scans' settings in the *ADAP chromatogram builder* module ([Step 4](#)); therefore, the same recommendations can be followed.
 - Provide a **'Suffix'** (e.g. `'_gap-filled'`) to name the newly-created feature lists.

778

779 **14. (Optional) Duplicate filter**

780 This module is intended for removing artefacts (duplicate features) that can originate from the gap
781 filling of misaligned features. In fact, when a misaligned feature undergoes gap filling, all the correctly-
782 aligned signals are retrieved, thus creating a 'duplicate feature' (see the [online documentation](#) for more
783 information). The 'Duplicate filter' module removes such duplicates by merging them into one
784 consensus feature.

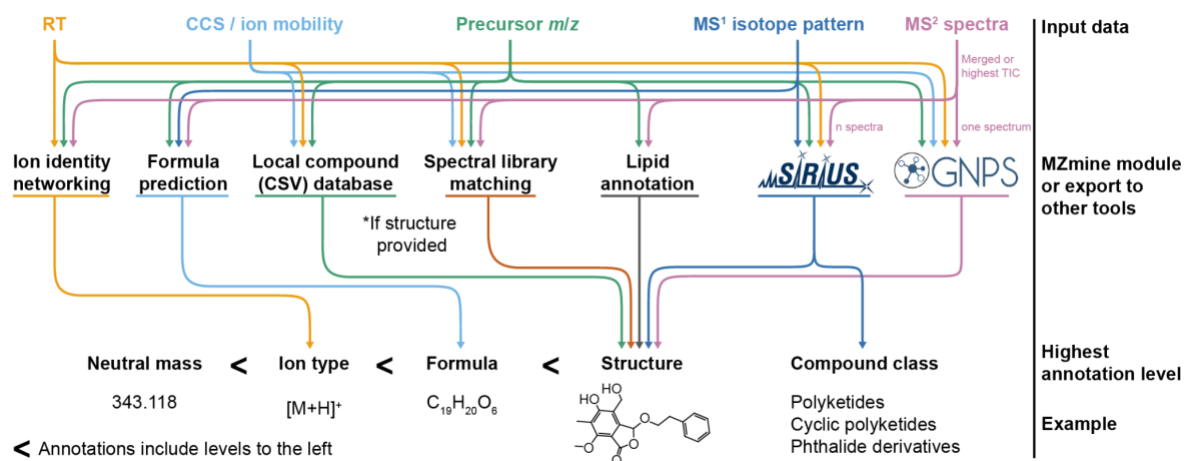
785 ▲ **CRITICAL** Similar to the ¹³C isotope filter' ([Step 11](#)), this module removes *features* matching the
786 filtering criteria. Therefore, we recommend using strict tolerances to avoid removing false duplicates.
787 As a rule of thumb, stricter tolerances than those used in the alignment step ([Step 13](#)) must be used.

- 788
- Navigate to **'Feature list methods → Feature list filtering → Duplicate feature filter'**
 - Select the **'Filter mode'**. This parameter determines how the RT and *m/z* of the consensus feature are calculated after the merging. The 'NEW AVERAGE' option (recommended) recalculates the features RT and *m/z* as average between the duplicate features.
 - Set the **'m/z tolerance'**, **'RT tolerance'** and, for mobility data, enable and set a **'Mobility tolerance'**. Features falling within these tolerances will be considered duplicates and thus removed. Therefore, we recommend using strict tolerances to avoid removing false duplicates.
 - Disable the **'Require same identification'** checkbox.
 - Provide a **'Suffix'** (e.g. `'_dup-filt'`) to name the newly created feature lists.

797

798 **Feature annotation**

799 MZmine offers various *feature annotation* modules to assign ion adducts, molecular formulas, and chemical
800 structures to the detected features. Furthermore, harmonised data exchange formats enable direct interface of
801 MZmine with other annotation tools. **Figure 3** provides an overview of the most popular modules and third-
802 party tools for feature annotation integrated with MZmine. The full list of available feature annotation tools is
803 provided in the [online documentation](#).



804

805 **Figure 3: Overview of popular modules and third-party tools for feature annotation integrated with MZmine.**

806 The various modules and third-party tools use different information retrieved during the MZmine preprocessing

807 (e.g. precursor m/z , isotope pattern, MS^2 spectra) to assign annotation.

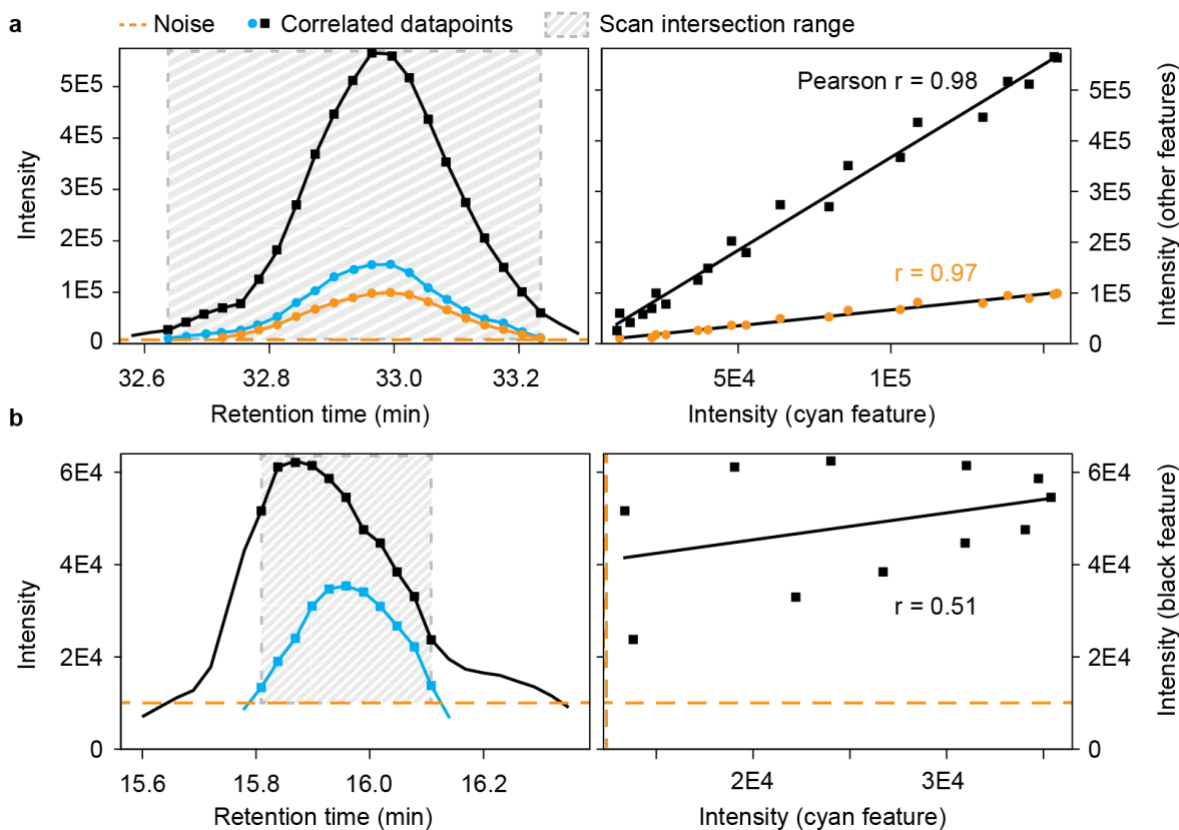
808

809

810

811 15. (Optional) Feature grouping - connecting ions of the same molecule

812 MZmine annotates features originating from the same chemical entity (e.g. multiple adducts) in two
 813 subsequent steps (Step 16 - Feature grouping and Step 17 - Ion Identity Networking). The metaCorrelate
 814 module searches for features with overlapping RT and chromatographic profiles to annotate them as
 815 'correlated features'. Besides checking if features fall within the same RT window, the chromatographic
 816 profile can be considered to distinguish between ions actually originating from the same molecule and
 817 features that are just co-eluting (Fig. 4).



818

819

Figure 4: Feature grouping by feature shape correlation. a, Features that originate from the same molecule exhibit highly correlated feature shapes at the same RT. b, The Pearson correlation drops significantly for features with different shapes or slight RT shifts.

821

822

823 • Navigate to 'Feature list methods → Feature grouping → Correlation grouping
824 (metaCorrelate)'.

825 • Specify the 'Feature lists' to process (see [Step 5](#)).

826 • Set the 'RT tolerance'. This is the maximum allowed RT deviation between features to be
827 grouped together. We recommend using a strict tolerance (e.g. \sim FWHM / 3) when the 'Feature
828 shape correlation' option is disabled. Otherwise, a wider tolerance (\sim FWHM \times 2) can be used
829 because the 'Feature shape correlation' option will provide a stringent filter for grouping.

830 • Set the 'Minimum feature height' for a feature to be considered for the grouping. Features
831 with intensity below this threshold will be disregarded. Set it to 0 to ignore this parameter.

832 • Set an 'Intensity threshold for correlation'. This threshold is used by the 'Feature shape
833 correlation' option, when enabled (see step g). Data points with intensity below this threshold
834 will be disregarded. Set it to 0 to ignore this parameter.

835 • (Optional) Enable and set the 'Min samples filter' by clicking the 'Show' button and set the
836 following parameters:

837 i. Set the 'Min samples in all' as the minimum number of samples (absolute or relative)
838 in which two features must be detected together to be grouped.

839 ii. Set 'Min samples in group' to 0 to ignore this parameter. This can be used when
840 sample groups are included in the experimental design and the information is

841 provided using the 'Sample metadata' module (see the [online documentation](#) for
842 more information).

843 iii. Set the '**Min %-intensity overlap**'. This is the minimum intensity overlap between the
844 smaller feature and the rest of the features being grouped. The intensity overlap is
845 calculated considering the sum of all data point intensities within the RT range of the
846 features being grouped. The default value (60%) should provide good results for most
847 applications.

848 iv. Enable the '**Exclude estimated features (gap-filled)**' option to ignore gap-filled
849 features during the grouping. We recommend using this option when smoothing was
850 applied during the processing ([Step 5](#) and [8](#)).

851 ● Enable the '**Feature shape correlation**' option and set the corresponding parameters by
852 clicking the 'Show' button. When enabled, the features' chromatographic profile is taken into
853 account for the grouping. We recommend using this option if most features have at least five
854 MS¹ data points (i.e. points-per-peak), two on each side of the apex. The following default
855 parameters should provide good results for most applications:

856 i. '**Min data points = 5**' and '**Min data points on edge = 2**'. These are the minimum
857 numbers of total data points and data points per peak side a feature must exhibit to
858 be considered for grouping. According to the typical number of points-per-peak in the
859 data, these values can be increased to make the feature grouping more strict.

860 ii. '**Measure = PEARSON**'. Although other correlation measures are available, we
861 recommend using the Pearson correlation as a starting point;

862 iii. Set the '**Min feature shape correlation**'. This is the minimum level of correlation
863 between the chromatographic profiles of the feature being grouped. A 85% Pearson
864 correlation threshold (default) should provide good results for most applications.

865 iv. Disable the '**Min total correlation**' option. This option represents an additional
866 constraint that considers all data points from all the features being grouped (see the
867 [online documentation](#) for more information).

868 ● (Optional) Enable the '**Feature height correlation**' option and set the corresponding
869 parameters by clicking the 'Show' button. This represents an additional constraint for the
870 grouping based on the heights of feature pairs across samples (see the [online documentation](#)
871 for more information). When enabled, the default parameters should provide good results for
872 most applications.

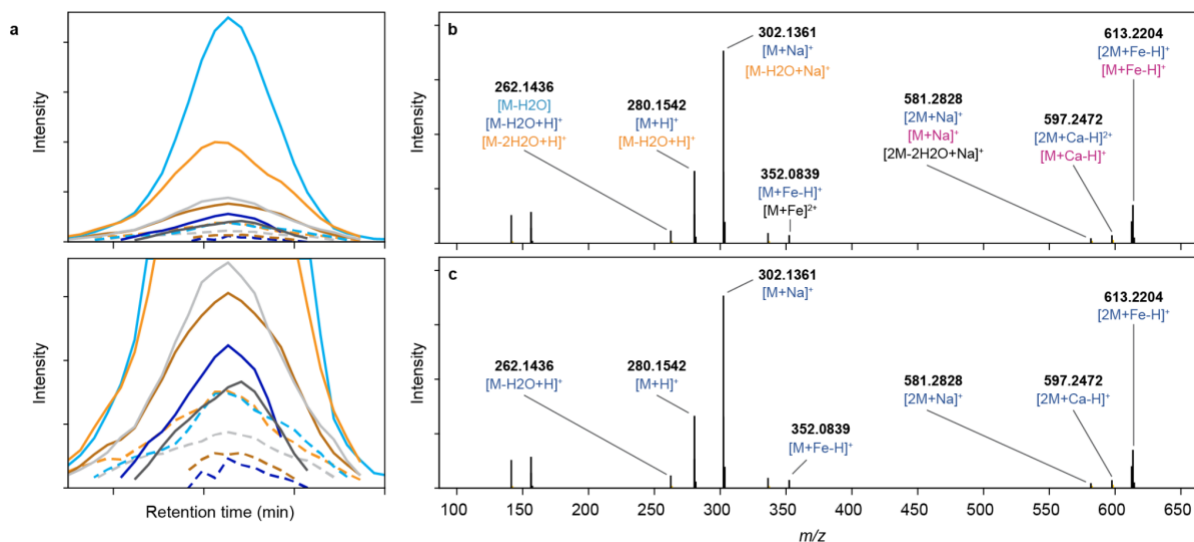
873 ● Provide a '**Suffix**' (e.g. '_corr') to name the newly created feature lists.

874

875 16. (Optional) Ion Identity Networking - Identification of ions

876 This step examines correlated features (i.e. features with same RT and chromatographic profile) to
877 annotate those generated by multiple adducts (e.g. [M+H]⁺ and [M+Na]⁺) or in-source modifications
878 (e.g. [M-H₂O]) of the same chemical entity. To do so, grouped features (see [Step 16](#)) are compared
879 pairwise against a user-defined list of adduct ions (e.g. [M+H]⁺), modifications (e.g. [M-H₂O]), and
880 multimers. When the *m/z* difference between two features corresponds to a possible adduct or in-
881 source modification, they are annotated (Fig. 5a). Multiple ion features that describe the same neutral
882 molecule (neutral mass) are then represented by an ion identity network, which is refined in a final step

883 to only retain the most confident networks (i.e. the largest number of ions pointing to the same neutral
 884 mass, see Fig. 5c). To perform this step, the *Feature grouping* module has to be run first (see [Step 16](#)).



885
 886 **Figure 5: Ion identity networking annotation refinement.** **a**, Grouped features are searched pairwise against a
 887 user library of ion adducts, in-source fragments, and multimers. **b**, Annotated IIN results for an example molecule
 888 in the MS1 spectrum. Each signal might be explained by different ions where annotations are linked in ion identity
 889 networks (coloured labels). **c**, After IIN refinement, only the best annotation that is supported by the largest
 890 network is retained. This means that $[M-H_2O]$ that was only defined by a single difference of m/z -18 can be defined
 891 as $[M-H_2O+H]^+$ through connections to other ion adducts, such as $[M+Na]^+$.

- 892 ● Navigate to **'Feature list methods → Feature grouping → Ion identity networking'**.
- 893 ● Set the **'m/z tolerance (intra-sample)'**. This is the maximum allowed m/z deviation when
 894 annotating two features as adducts pair or modification. This tolerance is very similar to the
 895 one used in the ^{13}C isotope filter (see [Step 11](#)); therefore, the same recommendations can be
 896 followed.
- 897 ● Select **'Check = ONE FEATURE'** to annotate two features if their m/z difference matches a
 898 possible adduct in at least one sample where the features were detected. The **'ALL FEATURES'**
 899 option is more stringent and requires the m/z difference to match in all samples the features
 900 were detected
- 901 ● Set the **'Min height'**. This is the minimum height for a feature to be considered for the adduct
 902 annotation. Set it to 0 to ignore this parameter and consider all features regardless of their
 903 intensity.
- 904 ● Define the **'Ion identity library'** to use for the annotation by clicking the **'Setup'** button and
 905 setting the following parameters:
 - 906 ○ Set **'MS mode'** as the ionisation polarity of the data.
 - 907 ○ Select the **'Adducts'** and **'Modifications'** to consider for the annotation from the
 908 corresponding lists. Adducts and modifications are combined together (e.g. $[M-H_2O+Na]^+$) to create the final list of ions to search for. Use the **'Reset positive'** and
 909 **'Reset negative'** buttons to show/restore the default adduct lists. User-defined
 910 adducts and modifications can be added manually by using the **'Add'** button.
 - 911 ○ Set the **'Maximum charge'** and the **'Maximum molecules/cluster'** of the adducts to
 912 be considered. We recommend setting both these parameters to 2 as a starting point
 913 for applications involving small molecules.
- 914 ● (Optional) Enable and set the **'Annotation refinement'** by clicking the **'Show'** button. These
 915 are additional constraints to consider for the adduct annotation and retain only the most
 916

917 confident annotation (see the [online documentation](#) for more information). When enabled,
918 the default parameters should provide good results for most applications.

919

920 17. (Optional) Import spectral libraries

921 To perform feature annotation based on spectral matching (see [Step 19](#)), spectral library files first have
922 to be imported into MZmine. The following file formats are supported: .json (e.g. the MoNA and GNPS
923 libraries), .mgf, .msp (e.g. the NIST library), and .jdx. Library files can be imported in MZmine by drag-
924 and-drop directly in the 'Libraries' tab in the main window. Alternative ways of importing files are
925 described in the [online documentation](#). Some of the most popular public spectral libraries can be freely
926 downloaded using the links provided below. We recommend using the .json format when available:

- 927 - MassBank of North America (MoNA): <https://mona.fiehnlab.ucdavis.edu/downloads>. Several
928 different libraries are available. If you are using the example dataset provided for this
929 procedure, we recommend downloading the 'LC-MS/MS Positive Mode' library.
- 930 - Global Natural Products Social Molecular Networking (GNPS): [https://gnps-
931 external.ucsd.edu/gnpslibrary](https://gnps-external.ucsd.edu/gnpslibrary). Several different libraries are available. If you are using the
932 example dataset provided for this procedure, we recommend downloading the 'GNPS-
933 LIBRARY' library.
- 934 - MassBank: <https://github.com/MassBank/MassBank-data/releases/tag/2022.12.1>. Download
935 the 'MassBank_NIST.msp' file.

936

937 18. (Optional) Spectral library search

938 Spectral library matching is the main approach for metabolite annotation in MS-based experiments.
939 The collected fragmentation spectra are matched against reference spectral libraries to assign putative
940 chemical structures to features matching a set of user-defined criteria, for example, *m/z* tolerance and
941 similarity score. In addition, RT- and CCS-based constraints can be used for chromatography–MS and
942 ion mobility data. In order to perform this step, spectral library files first have to be imported into
943 MZmine (see [Step 18](#)). *In-house* created, commercial (e.g. NIST) and open spectral libraries (e.g. MoNA,
944 GNPS) can be used.

945 ▲ **CRITICAL** Public spectral libraries contain mass spectra acquired under a wide range of instrumental
946 conditions and using a wide range of sample preparation and data curation protocols.⁵⁰ As a
947 consequence, spectra can vary greatly in terms of observed mass fragments, intensity ratios, and
948 spectral quality.⁵¹ For this reason, we recommend the users to treat annotation results with caution
949 when public data repositories are used for automated spectral matching.

- 950 ● The spectral library search can be performed on entire feature lists or individual features
951 selected from a feature list:
 - 952 ○ For the entire feature list(s). Navigate to '**Feature list methods** → **Annotation** →
953 **Search spectra** → **Spectral library search**'. In the dialogue box, specify the '**Feature**
954 **lists**' to process.
 - 955 ○ For individual features. Open the feature list → Select one or multiple features →
956 Right click → '**Search** → **Spectral library search**'
- 957 ● Specify the '**Spectral libraries**' to use for the spectral matching. Libraries must be already
958 imported (see [Step 18](#)).
- 959 ● Set the '**Scans for matching**'. For LC–MS/MS experiments, use the '**MS2 ≥ (merged)**' option.

960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006

TROUBLESHOOTING

- Set the '**Precursor m/z tolerance**'. This tolerance applies to the precursor m/z (not to the fragment m/z) and serves as a filter to reduce the number of library entries to match. A higher m/z tolerance can be set when public spectral libraries are used, as they are generally acquired with different instrument types, resulting in varying mass accuracy levels. We recommend 0.01 m/z or 20 ppm as a starting point.
- Set the '**Spectral m/z tolerance**' to pair the fragment m/z signals of the experimental and library spectra. As for the 'Precursor m/z tolerance' (step d), strict tolerances should be avoided. We recommend setting this tolerance higher than the 'Precursor m/z tolerance'. A good starting point is 0.01 m/z and 25 ppm.
- Enable '**Remove precursor**' to exclude the precursor m/z signal (± 4 Da) from the matching. This option can be useful when comparing experimental and library spectra acquired under different fragmentation methods (e.g. fragmentation mode or collision energies). This can result in varying intensities of the precursor ion, which decreases the overall spectral similarity. On the other hand, this option can reduce false library matches due to high abundant precursor ions in both the experimental and library spectra that match and account for most of the similarity.
- Set the number of '**Minimum matched signals**' that needs to be paired between the experimental and library spectra. A higher number (e.g. 6) results in increased confidence in the match. However, overly high values can impair the spectral matching for molecules with poor fragmentation patterns. We recommend 4 as a starting point for small molecules. Lower values will increase the probability of false library matches.
- Select the '**Similarity**' calculation algorithm to use for the spectral matching. This is the algorithm used to calculate the similarity between the experimental and library spectra. Choose the 'Weighted cosine similarity' for MS^2 data and the 'Composite cosine identity' for MS^1 and GC-EI-MS data, as this algorithm considers the relative intensity of neighbouring signals (more detailed information about the cosine similarity calculation can be found in the online documentation). With the 'Setup' button further parameters are available:
 - Choose the '**Weights**' for calculating the cosine similarity between the experimental and library spectra based on the m/z and signal intensity. The MassBank option gives more importance to matching fragments with higher m/z that might be more compound characteristic and can be used as a starting point.
 - Set the '**Minimum cos similarity**'. Only the library matches with cosine similarity above this threshold will be considered. A minimum similarity of 0.7, although the threshold to use largely depends on the next parameter.
 - Choose the '**Handle unmatched signals**' to determine how non-matching signals (i.e. m/z signals that occur only in the experimental or library spectrum) affect the cosine similarity. We recommend the setting 'KEEP ALL AND MATCH TO ZERO' for GC-EI-MS and MS^2 spectra. When this option is used, all unmatched signals weigh negatively on the overall score. For chimeric experimental spectra, for example, MS imaging data, the option 'KEEP LIBRARY SIGNALS' can remove additional signals in the experimental scans for the scoring. Information about the other options can be found in the [online documentation](#).
- Disable all the '**Advanced**' options by unticking the corresponding checkbox. These options can be used to add further constraints to consider in the library search. A detailed description of these options can be found in the [online documentation](#).

1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050

19. (Optional) Local compound database search

The local compound database search requires a compound table (i.e. a text file) with at least one of the following pieces of information: precursor m/z , neutral mass, chemical formula or chemical structure (i.e. SMILES). Additional details, such as RT, mobility, and CCS, can also be used as further annotation constraints (step h). When the neutral mass, chemical formula or SMILES is provided, the m/z of the corresponding adducts can be automatically calculated and used for the annotation.

- Navigate to **'Feature list methods → Annotation → Search precursor mass → Local compound database (CSV) search'**
- Specify the **'Feature lists'** to process (see [Step 5](#)).
- Click the **'Select'** button and browse the database file in your filesystem.
- Specify the **'Field separator'** of your database file (e.g. ';' for CSV files).
- In the **'Columns'** table, select the columns of the database file to import by ticking the corresponding checkbox. The names of columns to import are specified under 'Column name (csv)'. To edit them, double-click on the name, type the new column header and press enter. The attribute each database column corresponds to is specified under 'Data type (MZmine)'.
▲ CRITICAL For the import to be successful, the column headers in the database file must match exactly the names specified under 'Column name (csv)'.
- (Optional) Enable the **'Use adducts'** option to calculate and use the m/z of the adduct for the annotation, instead of the precursor m/z . When this option is enabled, MZmine automatically calculates the m/z of the specified adducts and/or in-source modifications based on the compound neutral mass. Therefore, to use this option, one among neutral mass, chemical formula or SMILES information must be provided in the database file. The list of adducts to search can be specified by clicking the 'Setup' button:
 - Set **'MS mode'** as the ionisation polarity of the data.
 - Select the **'Adducts'** and **'Modifications'** to consider for the annotation from the corresponding lists. Adducts and modifications are combined together (e.g. $[M-H_2O+Na]^+$) to create the final list of ions to search for. Use the 'Reset positive' and 'Reset negative' buttons to show/restore the default adduct lists. User-defined adducts and modifications can be added manually by using the 'Add' button.
 - Set the **'Maximum charge'** and the **'Maximum molecules/cluster'** of the adducts to be considered. We recommend setting both these parameters to 2 as a starting point for small molecules applications.
- Set the **' m/z tolerance'** as the maximum allowed deviation between the experimental mass and the exact mass provided in the database file. This tolerance mainly depends on the mass accuracy offered by the mass spectrometer used for the measurement.
- (Optional) Disable the **'Retention time tolerance'**, **'Mobility time tolerance'**, and **'CCS tolerance (%)'** options. When enabled, they are included as further annotation constraints. To do so, RT, mobility time and CCS value must be provided in the database file, respectively (zero and/or empty entries are ignored).
- (Optional) Disable the **'Filter filename header'** option. This option is intended for library building workflows and restricts the matching to a specific sample. If this is not the case, ignore this option.
- (Optional) Leave the **'Append comment fields'** field empty.

1051 20. (Optional) Lipid annotation

1052 MZmine offers a dedicated module for lipid annotation, which comes integrated with a set of pre-
1053 defined fragmentation rules for several glycerolipid and glycerophospholipid classes and subclasses.
1054 Furthermore, custom rulesets can be defined by the user and used for the annotation of derivatization
1055 products, oxidised forms, etc. The module first generates a database of lipid species based on a list of
1056 selected lipid classes/subclasses, the number of possible carbon atoms and double bond equivalents
1057 (DBE). From this database, theoretical precursor m/z are calculated and searched within the feature
1058 list. Moreover, *in silico* fragmentation spectra can be predicted (using both predefined and custom
1059 fragmentation rules) and matched against the experimental MS^2 data.

1060 ▲ **CRITICAL** Selecting many lipid classes with wide ranges of number of carbon atoms and DBE increases
1061 the size of the database exponentially and, thus, the computation time. For this reason, we recommend
1062 running the lipid annotation module multiple times and using specific ranges of carbon atoms and DBE
1063 for selected lipid classes (see the provided *batch_lipid_annotation.xml*).

- 1064 ● Navigate to **'Feature list methods → Annotation → Search spectra → Lipid annotation'**.
- 1065 ● Specify the **'Feature lists'** to process (see [Step 5](#)).
- 1066 ● Select the **'Lipid classes'** to consider for the database generation.
- 1067 ● Specify the ranges of **'Number of carbon atoms in chains'** and **'Number of double bonds in**
1068 **chains'** in all side chains combined. The selected ranges should be set in accordance with the
1069 selected lipid classes, as well as sample preparation and analysis methods, to minimise false
1070 positive annotations. For example, we recommend 14–26 carbons and 0–6 DBE for a lipid class
1071 with a single side chain, and 56–86 carbons and 0–18 DBE for cardiolipins.
- 1072 ● Set the **' m/z tolerance MS^1 level'**. This is the maximum allowed difference between the
1073 experimental and theoretical m/z values. This parameter mainly depends on the accuracy of
1074 the MS measurement.
- 1075 ● (Optional) The **'Show database'** button opens a separate window to visualise the lipid species
1076 database generated with the current parameters. Various info are displayed for each lipid
1077 species (e.g. exact mass, implemented fragmentation rule. Moreover, the 'Info' column
1078 highlights whether multiple annotations might occur if only the MS^1 information is considered
1079 (due to isomeric/isobaric overlap).
- 1080 ● (Optional) Activate the **'Search for lipid class specific fragments in MS/MS spectra'** option if
1081 MS^2 data were acquired. When deactivated, annotations are assigned based on MS^1 data only.
1082 Click the 'Show' button and set the following parameters:
 - 1083 i. Set the **' m/z tolerance MS^2 level'**. This is the maximum m/z allowed difference
1084 between the experimental and theoretical fragment signals.
 - 1085 ii. Set the **'Minimum MS/MS score'** between 0 and 100. This is the portion of intensity
1086 of the theoretical MS^2 spectrum explained by the experimental MS^2 spectrum.
 - 1087 iii. (Optional) Enable the **'Keep unconfirmed annotations'** checkbox to annotations
1088 based on MS^1 data only (these will be labelled in the feature list).
- 1089 ● (Optional) Enable the **'Search for custom lipid class'** option to use custom lipid classes for the
1090 annotation. To add a custom lipid class, click the 'Add' button and set the following
1091 parameters:
 - 1092 i. Define the **'Custom lipid class name'** (e.g. 'oxidised PC'), a **'Custom lipid class**
1093 **abbreviation'** (e.g. 'PC+O'), and a 'Lipid Backbone Molecular Formula' (e.g.
1094 ' $C_8H_{20}O_6PN$ ' for oxidised phosphatidylcholine). Multiple customised lipid classes can
1095 be defined and stored as a .json file.
 - 1096 ii. Click the 'Add' button to **'Add Lipid Chains'** to the backbone. At the time of writing,
1097 acyl and alkyl side chains are supported.

- 1098 iii. Enable the **'Add fragmentation rules'** option to add fragmentation rules for the
1099 custom lipid class being created. Multiple fragmentation rules can be set and stored
1100 as a .json file. Click the 'Add' button and set the following parameters:
1101 i. Select the **'Polarity'**
1102 ii. Select the **'Ionization method'** as the expected adduct type.
1103 iii. Set the **'Lipid fragmentation rule type'** (e.g. HEADGROUP_FRAGMENT).
1104 iv. Set the **'Lipid fragment information level'** as the level of structural
1105 information the fragment can provide for the annotation (e.g.
1106 'MOLECULAR_SPECIES_LEVEL').
1107 v. If a certain formula is needed for a fragmentation rule (e.g. headgroup
1108 fragment or headgroup neutral loss), specify the fragment's **'Molecular
1109 formula'** (e.g. 'C₅H₁₅NO₄P⁺' for the typical PC headgroup fragment). This is
1110 not needed for side chain fragments or side chain neutral losses.
1111 vi. A **'Molecular formula'** can be specified in the corresponding field. For
1112 example, this is needed for headgroup-related fragmentation rules (i.e.
1113 headgroup fragment or headgroup neutral loss). This is not needed for side
1114 chain-related fragmentation rules, as all the possible chain combinations are
1115 considered when generating the database.

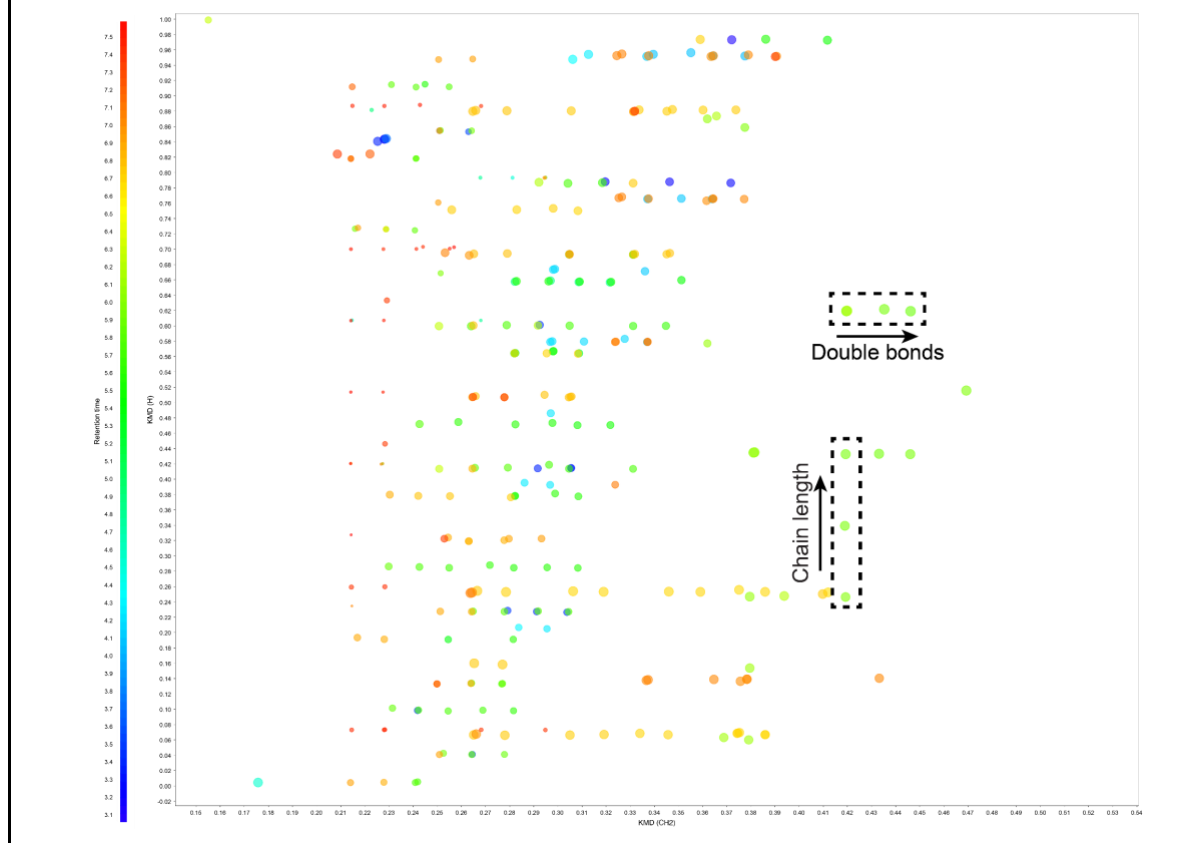
1116 **Box 6 – Visualisation tools for lipidomics**

The Kendrick mass defect (KMD) plot is a graphical tool to assist the identification of compounds that include repeating units in their chemical structures.^{52,53} Although originally developed for petroleum analysis, the KMD plot can be used to highlight differences in the acyl chain length and saturation of homologous lipid species.^{15,52,54,55} MZmine provides a module for the generation of KMD plots based on the repeating unit of interest (e.g. H₂, CH₂).⁵⁶ Moreover, the chromatographic information (i.e. RT) can be included in the plot in the form of a colour-coded scale,⁵⁷ thus enabling further visualisation possibilities (see figure). The KMD plot shown below can be generated as follows:

- a. Navigate to **'Visualization → Kendrick mass plot'**
- b. Select the **'Feature list'** to process;
- c. Select the **'Peaks'** (i.e. features) to be used in the generation of the plot. Click the 'Set to all' button to use all the features in the feature list. Specific *m/z* and RT ranges can be set using the 'Add' button.
- d. Select the **'Kendrick mass base for y-Axis'**. This is the structural unit of interest (e.g. CH₂, H) used to calculate the KMD displayed on the Y axis.
- e. Choose the variable displayed on the **'X-axis'** from the drop-down menu. The '*m/z*' and 'KM' options are available. However, a second Kendrick mass base can be displayed on the X axis by enabling the **'Kendrick mass base for x-Axis'** checkbox.
- f. Choose the variable displayed on the **'Z-axis'** (i.e. in the form of a colour-coded scale). Multiple options are available and, here too, another Kendrick mass base can be displayed on the Z axis by enabling the **'Kendrick mass base for z-Axis'** checkbox.
- g. Choose the **'Bubble size'** (i.e. data point size). The size of data points displayed in the KMD plot can be scaled to a variable chosen from the drop-down menu.
- h. Select the **'Z-axis scale'** and specify a **'Range for z-Axis scale'** to display. This can be used, for example, to display specific RT ranges on the Z axis.
- i. Select the **'Heatmap style'**. Various colour palettes are available

The plot displays the features annotated by the Lipid annotation module (Step 20) using two KMDs against each other (i.e. H on the Y-axis and CH₂ on the X-axis and). By doing so, homologous lipid species form easily

recognisable series based on the varying chain length (vertical) and level of unsaturation (horizontal). The size of data points and the colour scale represent the m/z and RT of each feature, respectively.



1117

1118 Feature filtering

1119 21. (Optional) Feature list rows filter

1120 This module allows the user to remove unwanted entries in a feature list using different filters. All
1121 features matching the filtering criteria will be either kept or removed from the feature list. Because
1122 several different filters are available and can be useful in specific applications, a few examples are
1123 explained below, and a detailed description for each filter is provided in the [online documentation](#).

- 1124 ● Navigate to **'Feature list methods → Feature list filtering → Feature list rows filter'**.
- 1125 ● Enable the **'Min aligned features (samples)'** filter and set the minimum number of samples
1126 (absolute or relative) in which a feature needs to be detected. This filter is commonly used to
1127 keep only features that were 'reproducibly' detected in analysis replicates or pooled quality
1128 control samples.
- 1129 ● Enable the **'Minimum features in an isotope pattern'** filter and set the minimum number of
1130 isotope signals to be detected in a feature. This filter is commonly used to remove all those
1131 features for which an isotopic pattern was not detected.
- 1132 ● Enable the **'Never remove feature with MS²'** option to always retain features associated with
1133 an MS² spectrum, regardless of the filters used. This option is commonly used when processing
1134 data for applications where the MS² data is the focus (e.g. molecular networking).

- 1135
- 1136
- Set **'Keep or remove rows'** to 'Keep rows that match all criteria'. The alternative option removes all features matching the selected criteria.

1137

1138 **Data export**

1139 MZmine enables the export of both quantitative (feature table) and qualitative (fragmentation spectra list)
1140 summaries of the feature detection and annotation workflow. Such outputs constitute the basis for a wide range
1141 of downstream data analysis such as feature-based molecular networking⁴⁶ (see [Step 23](#)), software packages for
1142 compound structure prediction (e.g. SIRIUS, see [Step 24](#)), and statistical and pathway analysis (e.g.
1143 MetaboAnalyst, see [Step 25](#)). Over the years, a number of other tools have integrated the output from MZmine
1144 into their pipelines; a list of such tools is available in the [online documentation](#). Besides export modules designed
1145 for specific third-party tools, export of feature lists is also possible via more general export modules available in
1146 MZmine and covered in the online documentation.

1147

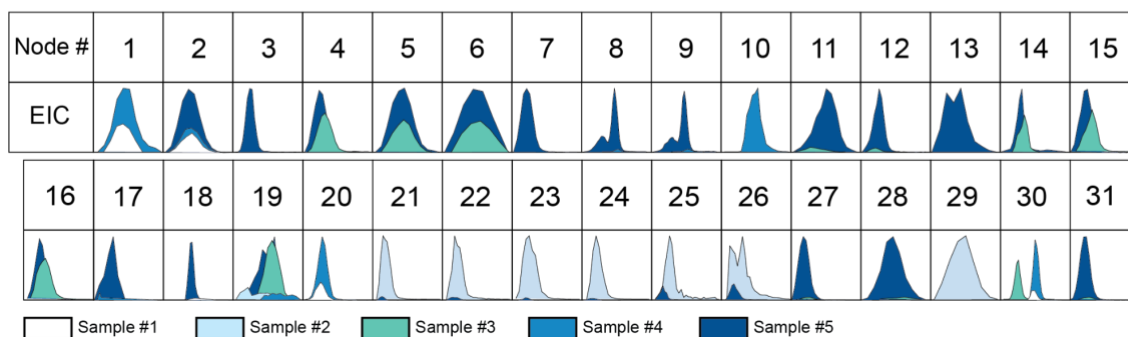
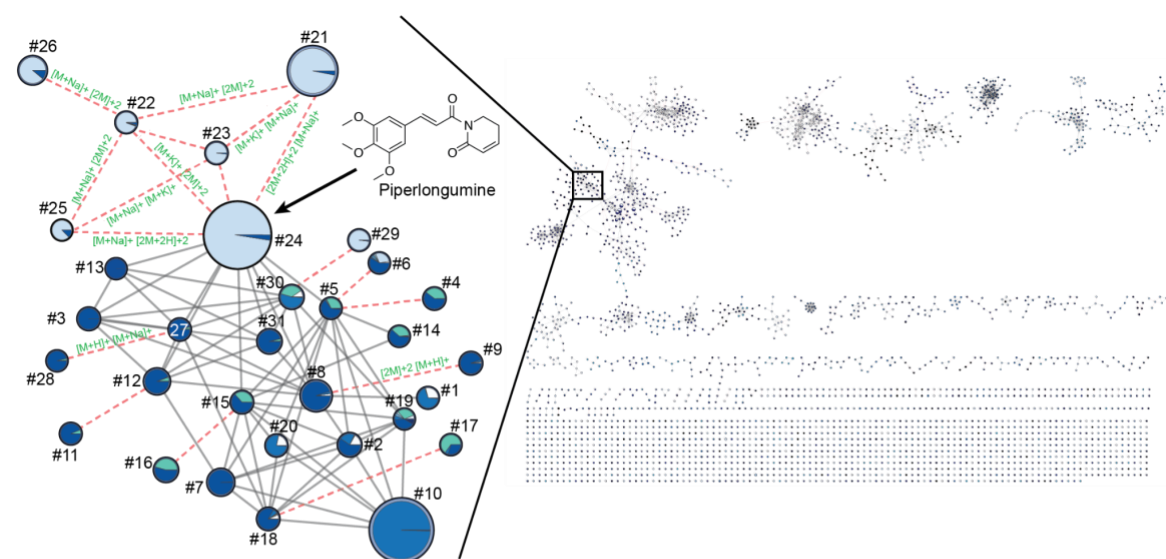
1148 **22. (Optional) Export for feature-based molecular networking and ion identity** 1149 **molecular networking (IIMN)**

1150 Feature-based molecular networking (FBMN) is a computational approach to increase metabolite
1151 annotation rates based on MS² spectral similarity (see **Box 7**). As described below, MZmine 3 allows the
1152 direct export of quantification table and MS² spectral list files necessary to perform FBMN using third-
1153 party tools such as GNPS, MetGem or FERMO.^{46,47,58,59} Moreover, feature correlation information from
1154 the 'Ion identity networking' (IIN) module ([Step 17](#)) can also be exported to perform ion identity
1155 molecular networking (IIMN)⁶⁰ on the GNPS platform.

- 1156
- 1157
- 1158
- 1159
- 1160
- 1161
- 1162
- 1163
- 1164
- 1165
- 1166
- 1167
- 1168
- 1169
- 1170
- 1171
- 1172
- 1173
- 1174
- a. Navigate to **'Feature list methods → Export feature list → Molecular networking (GNPS, FBMN, IIMN, MetGem)'**.
 - b. Specify the **'Feature lists'** to export (see [Step 5](#)).
 - c. Click the **'Select'** button and set the 'Filename' to a suitable file path in your computer filesystem (e.g. 'C:\Data\project_gnps' on Windows) for the export of the results. Two separate files are exported: a feature quantification table (CSV format) and an MS² spectral list (MGF format). If the IIN module was run (see [Step 17](#)), MZmine exports an additional file (*edges_msannotation.csv*) containing edges connecting *features* annotated as multiple adducts of the same molecule (see **Box 7**).
 - d. Disable the **'Merge MS/MS (experimental)'** option.
 - e. Select **'Filter rows = ONLY WITH MS2'** to export only features associated with an MS2 spectrum (see [Step 5](#)). If 'Ion identity networking' was performed in Step 16, select **'MS2 OR ION IDENTITY'**.
 - f. Set the **'Feature intensity'** measure (i.e. peak area or height) to use in the quantification table being exported.
 - g. Set the **'CSV export = SIMPLE'**.
 - h. Disable the **'Submit to GNPS'** checkbox. This option enables the direct submission of the files to GNPS for launching a FBMN job. However, the launched job cannot be saved to your GNPS account.

1175

Molecular networking is strategy for untargeted MS data clustering, annotation propagation and visualisation. Briefly, it organises untargeted MS data into networks where ions sharing similar MS² spectra appear as connected nodes. FBMN expands the concept of classical molecular networking⁶¹ by including the *feature detection* information (e.g. RT, peak area) in the molecular network construction.⁴⁶ Moreover, ion species of the same compound that do not connect in the network due to different fragmentation behaviour (e.g. [M+H]⁺ vs [M+Na]⁺) can be highlighted using IIMN, which includes the IIN information ([Step 17](#)) in the FBMN workflow.⁶⁰ Although various solutions exist,^{47,59} the most widely used platform to perform FBMN is the Global Natural Products Social Molecular Networking (GNPS) ecosystem.¹⁷ The main advantage of using the GNPS platform is the possibility to perform spectral library search against the GNPS spectral data repositories and to use a range of other computational tools for feature annotation integrated in the GNPS ecosystem (e.g. MASST⁶², network annotation propagation⁶³). An example of a molecular network generated using the example dataset provided for this procedure is shown in the figure below. The piperlongumine sub-network is highlighted. MS2 similarity edges are shown as solid grey lines and IIN edges are shown as dashed red lines. Nodes are shown as pie charts representing the intensity of each feature in the different samples (node size is proportional to the summed signal intensity). Each node is numbered and the corresponding aligned EICs are shown in the table below.



1178

1179

1180 23. (Optional) SIRIUS export

1181 SIRIUS is a software suite that combines automated interpretation of MS² spectra and machine learning
1182 to annotate each detected *feature* with an elemental formula, chemical structure and compound class.
1183 As described below, MZmine 3 enables the export of MS² spectral lists for the direct input into SIRIUS.

- 1184 a. Navigate to **'Feature list methods → Export feature list → SIRIUS / CSI-FingerID'**
- 1185 b. Specify the **'Feature lists'** to export (see [Step 5](#)).
- 1186 c. Click the **'Select'** button and set the **'Filename'** to a suitable file path in your computer
1187 filesystem (e.g. 'C:\Data\project_sirius' on Windows) for the export of the results. A MS²
1188 spectral list file (MGF format) will be exported.
- 1189 d. Disable the **'Merge MS/MS (experimental)'** option.
- 1190 e. Set the **'m/z tolerance'**. This tolerance is used to remove duplicate signals that were detected,
1191 for example, as an isotope and as a correlated feature at the same time.
- 1192 f. Disable all the remaining options.

1193

1194 24. (Optional) Export for statistical analysis

1195 Multiple downstream statistical workflows exist to further analyse results from MZmine and FBMN
1196 results. Although spreadsheet tools (e.g. Microsoft Excel) can be used for statistical analysis and data
1197 visualisation, these tools normally offer only a set of basic univariate tests and suffer limitations in terms
1198 of reproducibility and scalability. Alternatively, widely used programming languages such as R, Python,
1199 or Matlab offer more extensive analysis capabilities. Here, the aligned feature table from MZmine is
1200 typically imported and formatted for the desired statistical tests to perform. For efficient statistical
1201 analysis, various scripted pipelines and GUI-based web tools are available, such as the FBMN-STATS
1202 pipeline.⁶⁶ A widely used web platform that provides a user-friendly interface for statistical analysis of
1203 metabolomics data is MetaboAnalyst.⁶⁴ As described below, MZmine 3 enables the export of aligned
1204 feature tables in the format required for uploading to MetaboAnalyst.⁶⁵ Since MetaboAnalyst requires
1205 sample information (i.e. metadata) to be included in the exported *feature table*, a metadata file has to
1206 be either imported or created in MZmine first (see **Box 8**). We provide an example dataset (see the
1207 **Required data** section) together with sample metadata (*metadata_metaboanalyst.tsv*) and batch
1208 (*batch_metaboanalyst.xml*) files to perform the untargeted feature detection and export the aligned
1209 feature table in a MetaboAnalyst-compatible format.

- 1210 ● Navigate to **'Project → Sample Metadata'**. This will open a tab to add, edit or import metadata
1211 from an external file (TXT or TSV format).

- 1212 ● Click **'Import parameters'** and select the metadata file to import. For the provided example
1213 dataset, use the metadata file *'metadata_metaboanalyst.tsv'* (provided).

1214 ▲ **CRITICAL** For this to work, raw data files must be already imported in MZmine as the
1215 software will automatically try to match the metadata file with the raw file names.

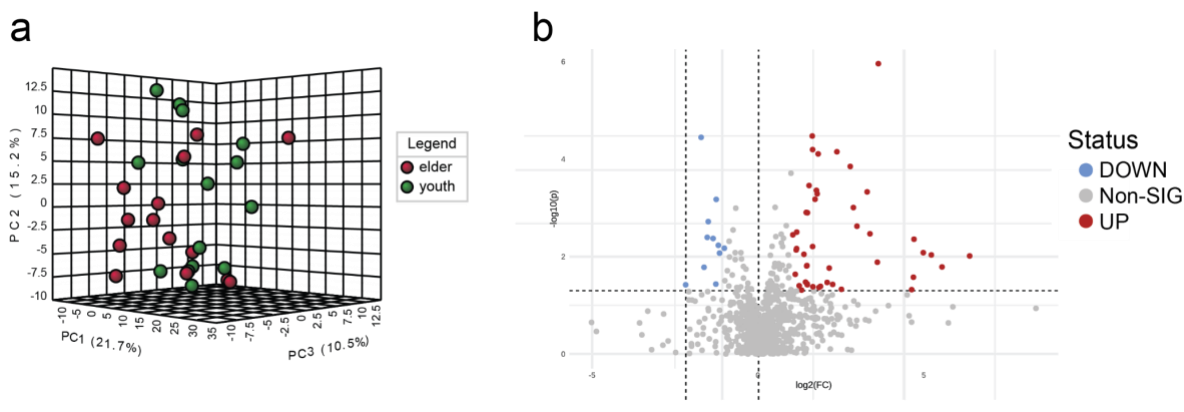
- 1216 ● Navigate to **'Feature list methods → Export feature list → Statistics Export (MetaboAnalyst)'**.
- 1217 ● Select the **'Feature lists'** to export (see [Step 5](#)).
- 1218 ● Choose the metadata grouping (e.g. 'Age').

- 1219
- 1220
- 1221
- Click the **'Select'** button and set the 'Filename' to a suitable file path in your computer filesystem (e.g. 'C:\Data\project_metaboanalyst' on Windows) for the results export. A quantification table (CSV format) including metadata information will be exported.

1222 **Box 8 – Downstream statistical analysis**

As mentioned in [Step 25](#), different tools and strategies can be used for downstream statistical analysis of MZmine feature tables and FBMN results; these include spreadsheet tools and programming scripts. A recent pipeline for downstream statistical analysis of MZmine and, in particular, FBMN results is "The Hitchhiker's Guide for Statistical Analysis for Feature-Based Molecular Network" (FBMN-STATS), which contains modules for data clean-up, batch-correction, as well as multivariate and univariate analysis. The code is available in multiple scripting languages (R, Python and Qiime2) as Jupyter and Google Colab notebooks. In addition to the scripting version, a GUI (FBMN-STATS-GUIde) is available.⁶⁶ Another widely used web-based platform is MetaboAnalyst, a popular tool for the post-processing of metabolomics data, including enrichment analysis, biomarker analysis and statistical analysis. MZmine enables the export of feature lists integrated with sample metadata needed for comparative statistical analysis in MetaboAnalyst. Metadata can be either imported in MZmine from an external file (TXT or TSV format) or created directly through the GUI. Importing metadata files requires a specific template, which can be obtained by exporting a blank metadata table directly from the MZmine 'Sample metadata' module. For further information on metadata creation and exporting, see [Step 25](#).

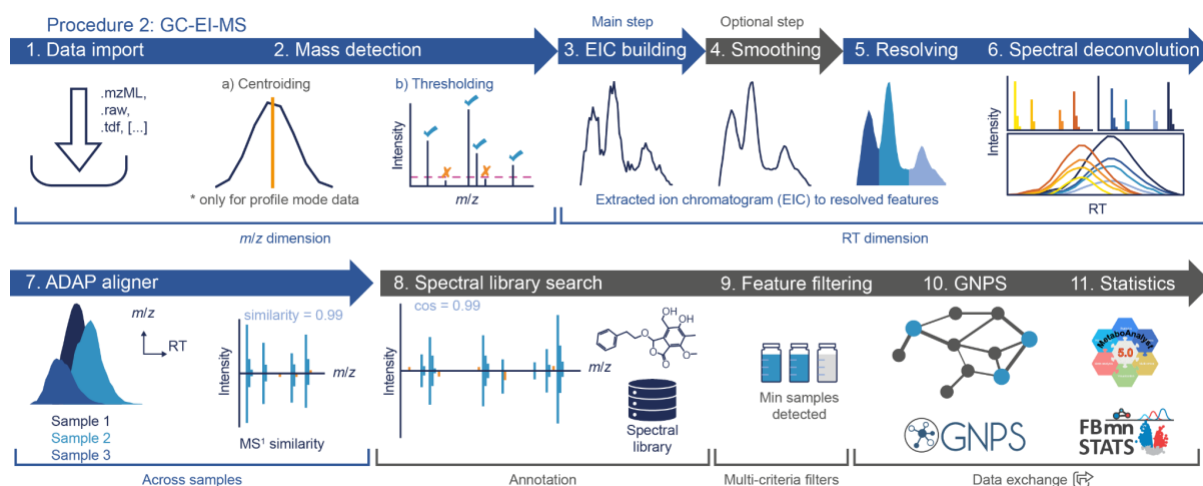
Here we showcase the capabilities of MetaboAnalyst when combined with MZmine. Example plots generated with MetaboAnalyst 5.0 using the provided dataset and metadata file are shown in the figure below. Samples were split into two equal groups (i.e. youth vs elder) and the MZmine quantification table was filtered and transformed in MetaboAnalyst. More information about MetaboAnalyst 5.0 can be found in the corresponding publication.⁶⁵



1223

1224 **Procedure 2: GC-EI-MS**

1225



1226

1227 **Figure 6: Schematic representation of the GC–EI–MS workflow described in Procedure 2.** A graphical reference for each
 1228 step (see numbers) summarises steps required for the GC–EI–MS workflow in blue. Additional optional steps (in grey) may
 1229 be applied to improve the input into the next steps or to provide additional annotations and results.

1230

1231 **1. Import MS data**

1232 The data import step can be performed as described in [Procedure 1 – Step 1](#).

1233 **2. Mass detection**

1234 The mass detection step can be performed as described in [Procedure 1 – Step 2](#). A more pronounced
 1235 background noise is often observed at higher GC temperature due to increased column bleeding. For
 1236 this reason, we recommend applying a higher noise level towards the end of the GC run. This can be
 1237 done by running the mass detection on two different RT range separately: one for the first and the
 1238 other for the second part of the GC run (see e.g. batch file *'batch_procedure-2.xml'*).²¹

1239 **3. EIC building with 'ADAP chromatogram builder'**

1240 The EIC building step can be performed as described in [Procedure 1 – Step 4](#).

1241 **4. (Optional) Chromatogram smoothing**

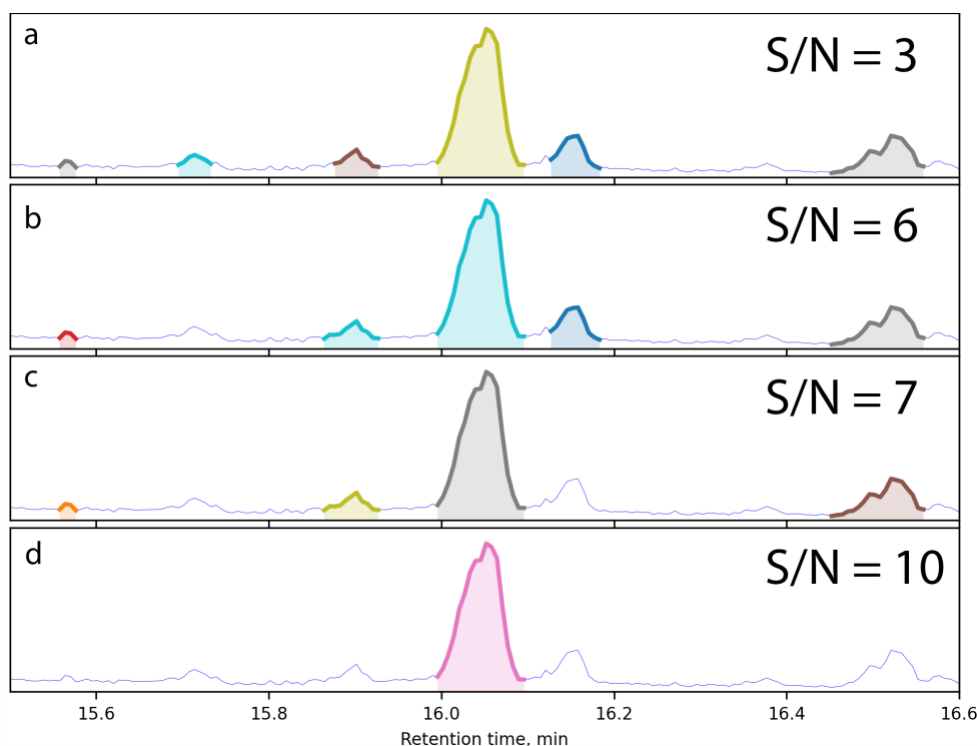
1242 The EIC smoothing step can be performed as described in [Procedure 1 – Step 5](#).

1243 **5. EIC resolving with the ADAP resolver**

1244 As explained in [Procedure 1 – Step 6](#), the EIC traces built in the previous steps might contain multiple
 1245 chromatographic peaks that need to be resolved into individual features. Although various EIC resolving
 1246 algorithms are available, we recommend using the *ADAP resolver* module when processing GC–EI–MS
 1247 data. In contrast to the *Local minimum resolver* (described in [Procedure 1 – Step 6](#)), which resolves EIC
 1248 traces based on local minima, the *ADAP resolver* uses the continuous wavelet transform algorithm to

1249 detect peaks in EIC traces.⁶⁷ A detailed description of this module is provided in the [online](#)
1250 [documentation](#).

- 1251 ● Navigate to **'Feature detection → Chromatogram resolving → ADAP resolver'**.
- 1252 ● Specify the **'Feature lists'** to process (see [Procedure 1 – Step 5](#)).
- 1253 ● Specify how to handle the **'Original feature list'**. This option determines whether to 'KEEP' in
1254 memory or 'REMOVE' the input feature lists once the processing is completed. We recommend
1255 using the 'KEEP' option during parameter optimization (see **Processing large datasets section**).
- 1256 ● Disable the **'MS/MS scan pairing'** option when processing GC–EI–MS data.
- 1257 ● Select **'Dimension → Retention time'**;
- 1258 ● Set the **'(S/N) threshold'**. This is the minimum signal-to-noise (S/N) a feature has to exhibit
1259 after resolving to be considered valid. The S/N ratio is the ratio between the signal of the peak
1260 and the nearby background. High S/N ratios (e.g. ≥ 10) of S/N ratio are normally associated
1261 with 'real' features whereas 'noisy' features (i.e. hardly distinguishable from the nearby
1262 background, see **Fig. 7**) normally exhibit low S/N ratios (e.g. ≤ 3). We recommend a starting
1263 value of 6.
- 1264 ● Specify the **'S/N estimator'**. This is the algorithm used to estimate the S/N ratio of each peak
1265 detected during the resolving. Two options are available. The 'Intensity Window S/N'
1266 (preferred in most cases), uses the feature height as the signal level and the standard deviation
1267 of the data immediately before and after the feature as the noise level. Because of this, the
1268 S/N estimation might not be accurate for closely-eluting features (see **Fig. 7** panel **b** and **c**).
- 1269 ● Set the **'Coefficient/area threshold'**. This is the minimum ratio between the highest wavelet
1270 coefficient of a peak and its area. The parameter is designed to assist the detection of low-
1271 intensity peaks when the noise level is high (e.g. as high as the real signal). This can be done
1272 by setting a lower 'S/N threshold' (e.g. 3–5) and a high 'Coefficient/area threshold' (e.g.
1273 ≥ 300). We recommend disabling this parameter (i.e. set it to 0) when the noise level is low.
- 1274 ● Specify the **'min feature height'**. This is the minimum signal intensity a peak needs to reach to
1275 be retained as a feature after the resolving. We recommend using the same value set for the
1276 'Minimum absolute height' in the *ADAP chromatogram building* module ([Step 3](#)).
- 1277 ● Define the allowed **'Peak duration range'**. This is the acceptable duration of a
1278 chromatographic peak to be retained as a feature after the resolving. This parameter can be
1279 used to filter out noisy features based on their overly short, or long, duration.
- 1280 ● Set the **'RT wavelet range'**. This is the range of wavelets RT width used for detecting peaks. It
1281 must be noted that this parameter is used to detect peaks, while the 'Peak duration range' is
1282 used to filter out noisy peaks based on their overly short, or long, duration.
1283 ▲ **CRITICAL** The *ADAP Resolver* algorithm is very sensitive to the upper limit of this parameter,
1284 which we recommend to set to approximately half a typical peak width. The lower limit can be
1285 set to 0.
- 1286 ● Provide a **'Suffix'** (e.g. '_ADAP-res') to name the newly created feature lists (see [Procedure 1](#)
1287 [– Step 5](#)).



1288

1289 **Figure 7: The S/N threshold parameter of ADAP Resolver.** Peaks produced by ADAP Resolver with the parameter 'S/N
 1290 estimator' set to 'Intensity Window S/N' and various values of the parameter 'S/N threshold'. **a**, S/N = 3: Too many peaks are
 1291 detected; **c**, S/N = 6: An optimal number of peaks is detected; **b**, S/N = 7: Most peaks are detected except some peaks in
 1292 close proximity to large peaks; **d**, S/N = 10: Too few peaks are detected.

1293 6. Spectral deconvolution

1294 *Spectral deconvolution* is a crucial step in the feature detection of GC–EI–MS due to the extensive *in-*
 1295 *source* fragmentation caused by EI. In fact, EI-produced spectra can contain fragment ions originating
 1296 from different co-eluting metabolites.⁶⁸ Therefore, *spectral deconvolution* is necessary to
 1297 computationally reconstruct fragmentation mass spectra for *features* not fully resolved by
 1298 chromatography. The so-reconstructed spectra are then used during the feature alignment and, most
 1299 importantly, in the feature annotation step. For this reason, the fine-tuning of the *deconvolution*
 1300 parameters is crucial (see **Box 9**).

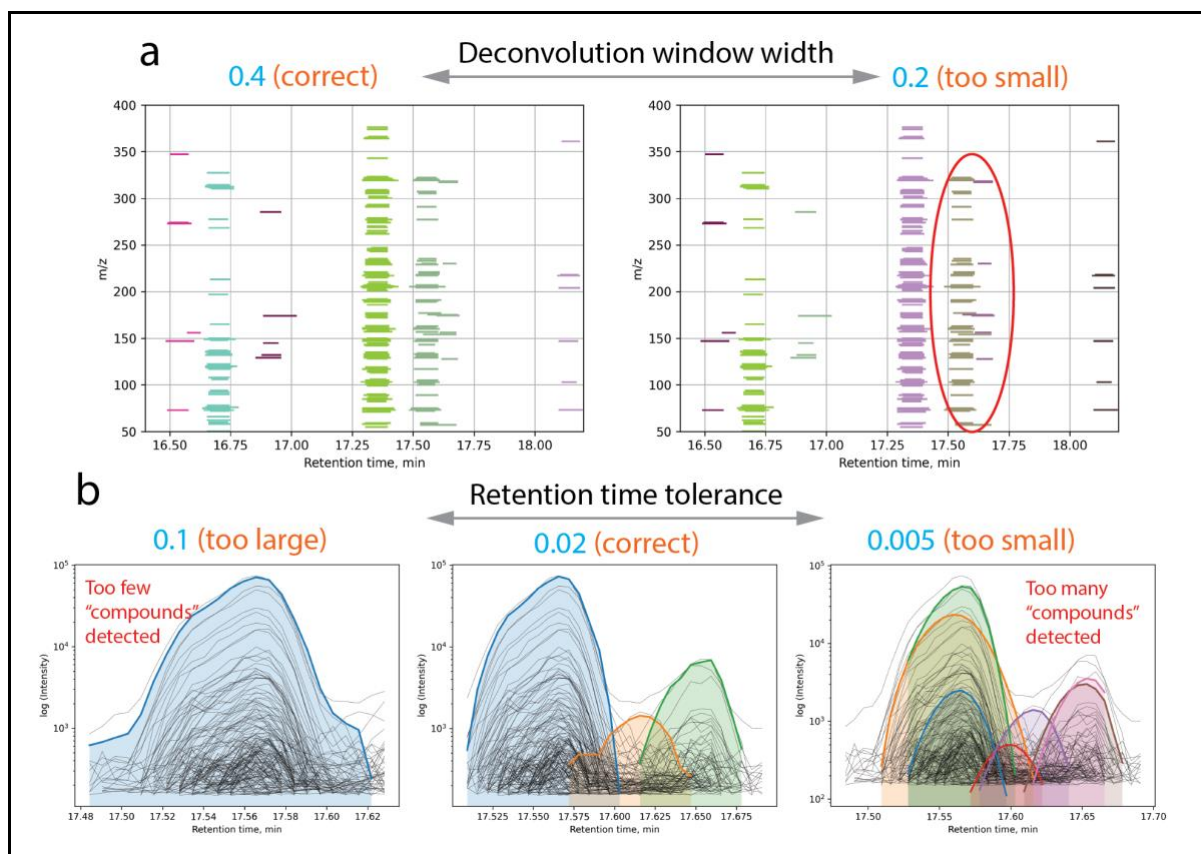
- 1301 ● Navigate to 'Feature list methods → Spectral deconvolution (GC) → Multivariate curve
 1302 resolution'. Although two algorithms are available, we recommend the 'Multivariate curve
 1303 resolution' for its simplicity.
- 1304 ● Specify the 'Features' and 'Chromatograms' list to process. This algorithm requires both EICs
 1305 constructed in [Step 3](#) (i.e. 'Chromatograms') and peaks detected in [Step 5](#) (i.e. 'Features'). To
 1306 do so, enable the 'Specific feature lists' option from the drop-down menu, click the 'Select'
 1307 button and manually select the feature lists produced by the 'ADAP Chromatogram Builder'
 1308 and the 'ADAP Resolver', respectively. As an alternative, name patterns in the feature lists can
 1309 be used to automatize the selection (e.g. batch mode). To do so, choose the 'Feature list name
 1310 pattern' option from the drop-down menu, click the select button and type a suitable name
 1311 pattern. For example, 'Chromatograms' and 'Features' can be selected by typing the '*'
 1312 character, followed by the suffix used to name the feature lists created in the respective step
 1313 (e.g. '*_eic' for 'Chromatograms' and '*_ADAP-res' for 'Features').
- 1314 ● Set the 'Deconvolution window width'. This is the maximum width of a deconvolution
 1315 window. Overall, the optimal deconvolution window should be wide enough to contain co-

1316 eluting peaks within the same window, and small enough to allow a fast execution of the
1317 algorithm (see **Box 9**).
1318 **▲ CRITICAL** This parameter directly affects the *spectral deconvolution* performance. Overly
1319 narrow deconvolution windows can cause suboptimal *feature detection* and/or inaccurate
1320 reconstruction of the fragmentation spectra. Overly wide deconvolution windows can
1321 significantly increase the computing time.

- 1322 ● Specify '**Retention time tolerance**' for the peak grouping. This is the maximum allowed RT
1323 deviation between any two peaks being grouped.
- 1324 ● Specify the '**Minimum number of peaks**' in a group to be considered valid. Groups with fewer
1325 peaks are discarded. This parameter is dataset-specific and mainly depends on the number of
1326 peaks detected by the *ADAP resolver* module ([Step 5](#)). Typically, values between 1 and 10 (or
1327 more) are used. By setting this parameter to 1, all groups are allowed.
- 1328 ● (Optional) Enable the parameter '**Adjust apex retention time**' if most peaks consist of few data
1329 points (e.g. 4–8). When this option is enabled, the algorithm fits a parabola into each EIC peak
1330 to determine its apex and calculate the RT.⁶⁹

1331 **Box 9 – Spectral deconvolution**

The *spectral deconvolution* constructs fragmentation mass spectra of GC–EI–MS features in three steps.⁶⁹ In the first step, the entire RT range is split into disjoint deconvolution windows, which separate the detected GC–EI–MS peaks into non-overlapping interval (see panel **a** in the figure). Each interval represents a detected GC–EI–MS peak, and the different colours denote the produced clusters. The fine-tuning of the 'Deconvolution window width' parameter is crucial to achieve optimal clustering. In the second step, peaks within each deconvolution window are clustered based on their chromatographic shapes to infer the number of GC–EI–MS features within each window (see coloured peaks in panel **b** of the figure). Their number can be adjusted by changing the 'Retention time tolerance' parameter. When the RT tolerance is too large, low-intensity features can be missed. On the other hand, when the retention time tolerance is too small, the algorithm will produce false features with inaccurate fragmentation mass spectra. The third step in the *spectral deconvolution* is building fragmentation mass spectra of the GC–EI–MS features by decomposing every EIC into a linear combination of the shapes of the features inferred by the second step (coloured peaks in the **b** panel).



1332 7. ADAP feature alignment

1333 Although two alignment algorithms are available for GC–EI–MS data, we recommend using the *ADAP*
 1334 *aligner* module. This module aligns deconvoluted *features* across multiple samples based on the RT
 1335 proximity and similarity of their reconstructed fragmentation mass spectra.

- 1336
- 1337
- 1338
- 1339
- 1340
- 1341
- 1342
- 1343
- 1344
- 1345
- 1346
- 1347
- 1348
- 1349
- 1350
- 1351
- 1352
- 1353
- 1354
- 1355
- Navigate to **'Feature list methods → Alignment → ADAP aligner (GC)'**.
 - Specify the **'Feature lists'** to process.
 - Specify the **'Min confidence'**. This is the minimum fraction of samples a feature should be detected in to be retained during the alignment. For example, if a *feature* is expected in at least N out of M samples, set this parameter to N/M. Set it to 0 to ignore this parameter.
 - Set the **'Retention time tolerance'**. This is the maximum allowed RT deviation between *features* being aligned.
 - Specify **'m/z tolerance (sample-to-sample)'**. This is the maximum allowed *m/z* deviation between the samples for feature alignment. This is a sample-to-sample tolerance and largely depends on the performance (stability) of the MS analyser over time.
 - Set the **'Score threshold'**. This is the minimum spectral similarity for features being aligned.
 - Specify **'Score weight'**. When multiple features fall within the defined tolerances, this parameter defines the contribution of the RT proximity and spectral similarity in calculating the total alignment score. When set to 0, only the spectrum similarity is considered. When set to 1, only the RT difference is considered. When a value between 0 and 1 is set, a weighted combination of spectral similarity and the RT difference is used. The default value of this parameter is 0.1.
 - Set the **'Retention time similarity = Retention time difference (fast)'**. This is the algorithm for calculating the RT similarity for the alignment. Although two options are available, we recommend using the 'Retention time difference' option.

1356 Feature annotation and data export

1357 8. (Optional) Spectral library search

1358 Feature annotation based on spectral matching can be performed as described in [Procedure 1 – Step](#)
1359 [18](#) and [Step 19](#) with the following adjustments:

- 1360 ● Import a library of GC–EI–MS spectra. If you are using the example dataset provided for this
1361 procedure, we recommend downloading the ‘GC-MS spectra’ library from the MoNA website
1362 (<https://mona.fiehnlab.ucdavis.edu/downloads>).
- 1363 ● In the ‘Spectral library search’ module, set the ‘Scans for matching = MS1’ and
- 1364 ● Select ‘Similarity = Composite cosine identity’ algorithm. This algorithm considers the relative
1365 intensity of neighbouring signals in the similarity calculation and is recommended for GC–EI–
1366 MS data. It is used to calculate the similarity between experimental and library spectra.

1367

1368 9. (Optional) Feature list rows filter

1369 The feature filtering step can be performed as described in [Procedure 1 – Step 22](#). A detailed description
1370 for each filter is provided in the [online documentation](#).

1371 10. (Optional) Export for feature-based molecular networking

1372 To export the feature quantification table and MS² spectral list for FBMN, a different module from the
1373 one described in Procedure 1 – Step 22 should be used when dealing with GC–EI–MS data.

- 1374 ● Navigate to ‘Feature list methods → Export feature list → GNPS–GC–MS (with ADAP)’.
- 1375 ● Specify the ‘Feature lists’ to process.
- 1376 ● Click the ‘Select’ button and set the ‘Filename’ to a suitable file path in your computer
1377 filesystem for the export of the results (e.g. ‘C:\Data\project_gnps’). Two separate files are
1378 exported: a feature quantification table (CSV format) and an MS² spectral list (MGF format)
- 1379 ● Select the ‘Representative m/z = As in feature table’. This is the m/z assigned to each feature
1380 in the MGF file.
- 1381 ● Set the ‘Feature intensity’ measure (i.e. peak area or height) to use in the quantification table
1382 being exported.

1383

1384 11. (Optional) Export for statistics (MetaboAnalyst)

1385 Export of the aligned feature table for statistical analysis in MetaboAnalyst can be done as described in
1386 [Procedure 1 – Step 24](#).

1387

1388

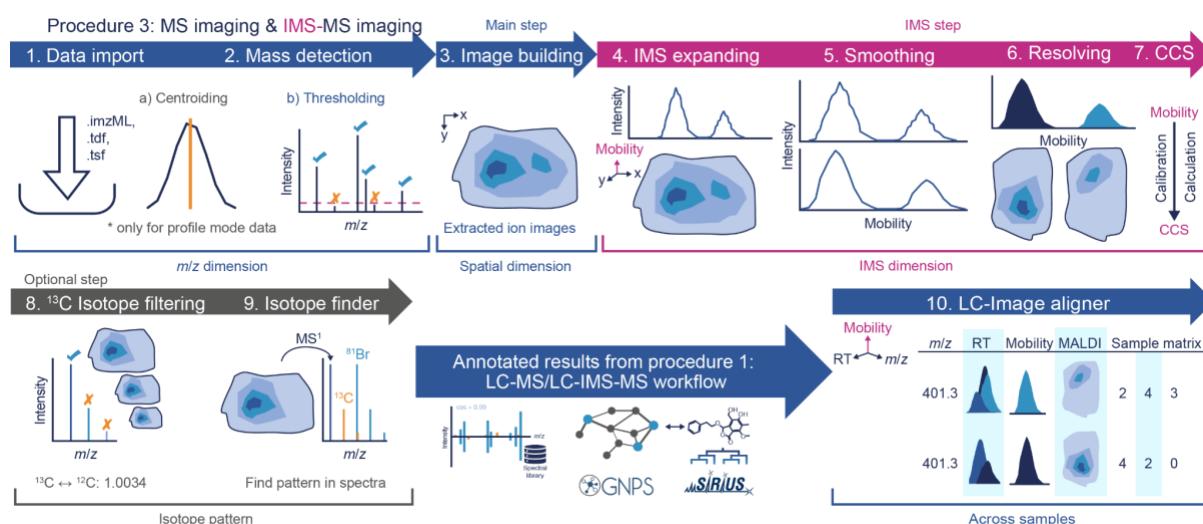
1389

1390

1391

1392 Procedure 3: MS imaging and IMS–MS imaging

1393



1394

1395 **Figure 8: Schematic representation of the MS imaging and IMS-MS imaging workflows described in Procedure 3.** A
 1396 graphical reference for each step (see numbers), summarising the steps required for the MS imaging workflow in blue.
 1397 Additional steps required for IMS–MS imaging data are highlighted in magenta and optional steps for both workflows, in
 1398 grey, may be applied to improve the input into the next steps or to provide additional annotations and results.

1399 1. Import MS data

1400 The data import step can be performed as described in [Procedure 1 – Step 1](#).

1401 **▲ CRITICAL** Although we recommend converting MS imaging data to the .imzML format prior to
 1402 processing, at the time of writing this format does not support IMS. Therefore, IMS–MS imaging data
 1403 can only be imported into MZmine using the native Bruker format (.tdf).

1404 2. Mass detection

1405 The mass detection step can be performed as described in [Procedure 1 – Step 2](#). When processing IMS–
 1406 MS imaging data, mass detection has to be run on both *mobility scans* and *frames spectra*. Two different
 1407 noise levels can be applied (see [Procedure 1 – Step 2](#)).

1408 Feature image detection and resolving

1409 3. Image builder

1410 This step builds an image for each *m/z* value detected over a minimum number of adjacent MS¹ scans
 1411 (pixel) in the MS imaging analysis. Images matching a set of user-defined requirements (e.g. minimum
 1412 number of data points and intensity) are stored as features in a *feature list*.

- 1413 ● Navigate to **'Feature detection → Imaging → Image builder'**.
- 1414 ● Specify the **'Raw data files'** to process;
- 1415 ● Set the **'Scan filters'**. Enable the checkbox, click the **'Show'** button and set the **'MS level filter'**
 1416 as **'MS1, level = 1'**. When fragmentation data are acquired, images can be built also for the
 1417 MS² level.
- 1418 ● Set the **'m/z tolerance (scan-to-scan)'**. As imaging experiments typically require longer
 1419 analysis time than LC–MS, mass accuracy drift may occur during the measurement (especially

- 1420 when using TOF instruments). For this reason, we recommend using larger tolerances
1421 compared to LC–MS data processing (e.g. 0.005 m/z or 20 ppm).
- 1422 ● Set the **‘Min consecutive scans’** as the minimum number of consecutive pixels where an *m/z*
1423 must be detected for the corresponding image to be considered valid. This parameter mainly
1424 depends on the spatial resolution of your instrument and the size of the sample.
 - 1425 ● Set the **‘Minimum absolute height’** as the minimum intensity the most intense pixel in the
1426 image must exceed for the corresponding image to be considered valid.
 - 1427 ● Set the **‘Minimum total signals’**. This is the minimum number of pixels an ion image must
1428 contain to be considered valid.
 - 1429 ● Provide a **‘Suffix’** to name the newly created feature lists (e.g. ‘_img’).

1430

1431 4. (Only IMS data) IMS expander

1432 As explained in [Procedure 1 – Step 7](#), in this step the individual *mobility scans* are inspected to create
1433 *IMS-resolved features*.

- 1434 ● Navigate to **‘Feature detection → LC-IMS-MS → Ims expander’**.
 - 1435 ● Enable and set the **‘m/z tolerance’**. This is the maximum allowed deviation between the *m/z*
1436 of the image feature (frame scans only) and the *m/z* signals in the individual *mobility scans*.
 - 1437 ● Disable the **‘Raw data instead of thresholded’** parameter unless low-intensity compounds are
1438 of interest. Enabling this option increases computation cost.
 - 1439 ● Disable **‘Override default mobility bin width (scans)’** to use MZmine’s default binning of
1440 mobility scans.
 - 1441 ● Enable and set the **‘Maximum features per thread’**. This parameter controls thread
1442 parallelization (i.e. number of images processed at the same time), which affects RAM
1443 consumption and processing time
- 1444 ▲ **CRITICAL** Processing MS imaging data is much more computational demanding than LC–MS
1445 data processing. For this reason, we recommend setting a small number (e.g. 5–10). If the
1446 software crashes at this step, lower the value.

1447

1448 5. (Only IMS data, optional) EIM smoothing

1449 The mobilogram smoothing step can be performed as described in [Procedure 1 – Step 8](#).

1450 6. (Only IMS data) EIM resolving

1451 The mobilogram resolving step can be performed as described in [Procedure 1 – Step 9](#).

1452 7. (Only IMS data, optional) CCS calibration and calculation

1453 The calibration and calculation of CCS values can be performed as described in [Procedure 1 - Step 10](#).

1454 8. (Optional) ¹³C isotope filter

1455 While applying the ¹³C isotope filter is recommended in most cases when processing LC–MS data, extra
1456 attention should be paid when using this module on MS imaging data. This is because isobaric overlap
1457 of ¹³C isotopic signals is much more frequent in MS imaging data. Consider applying this filter based on

1458 the resolving power of your MS instruments. The module can be used as described in [Procedure 1 –](#)
1459 [Step 11](#) with the following adjustment:

- 1460 - Set the 'RT tolerance' to a high value (e.g. 1.0E4). This is needed to ignore the acquisition time
1461 associated with each feature. In fact, although there is no chromatographic separation, a total
1462 ion current (TIC) is still acquired over time in MS imaging experiments.

1463

1464 9. Isotope pattern finder

1465 Isotopes can be annotated with the isotope pattern finder module as described in [Procedure 1 – Step](#)
1466 [12](#).

1467

1468 Alignment with LC–MS data

1469 10. (Optional) LC-Image-Aligner

1470 If the same samples were analysed by means of MS–imaging and LC–MS, the *feature detection* results
1471 from both datasets can be aligned into a single *feature list* to increase annotation confidence.⁹ This can
1472 be done using the 'LC-Image-Aligner' module, which uses an alignment scoring system similar to the
1473 Join aligner algorithm.

- 1474 ● Navigate to 'Feature list methods → Alignment → LC-Image-Aligner'.
- 1475 ● Specify the 'Feature lists' to process.

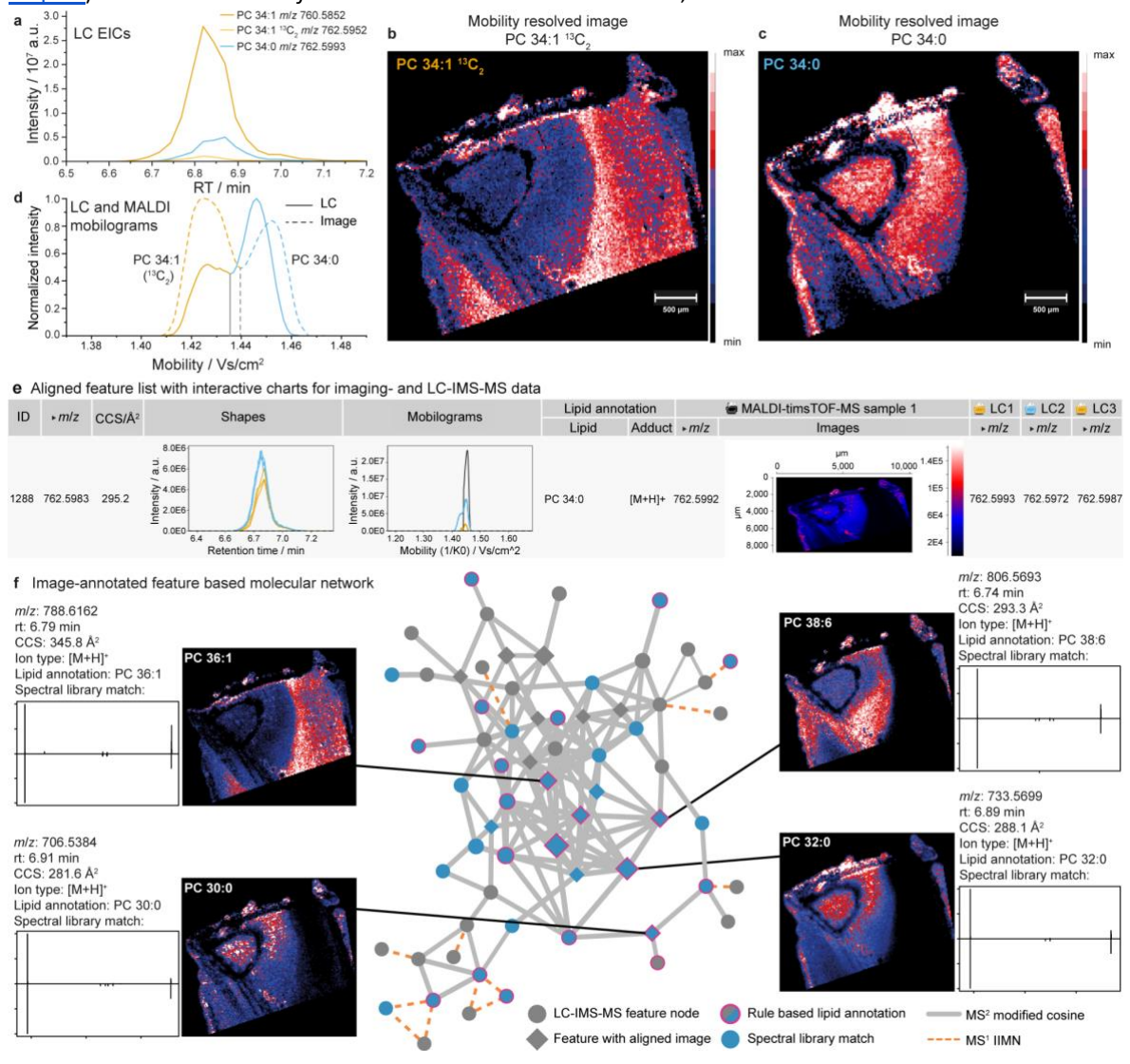
1476 ▲ **CRITICAL** The LC feature list is obtained by processing the LC–(IMS–)MS data according to
1477 [Procedure 1](#). The final aligned feature list should be used here.

- 1478 ● Set the 'm/z tolerance'. This is the maximum allowed *m/z* deviation between the imaging
1479 experiment and the aligned LC–(IMS–)MS *feature list*. This tolerance should take into account
1480 potential mass accuracy drifts occurred during the measurements and largely depends on the
1481 stability of the MS analyser.
- 1482 ● Set the 'm/z weight' and 'Mobility weight' (see [Procedure 1 – Step 13](#)). If the 'mobility
1483 tolerance' option (step e) is disabled, the 'mobility weight' will be ignored.
- 1484 ● (Only IMS data) Enable and set a 'mobility tolerance' if IMS–MS imaging and LC–IMS–MS data
1485 are being aligned.
- 1486 ● Provide a 'Feature list name' to name the newly created aligned feature list. Use '{lc}' to insert
1487 the name of the feature LC feature list.

1488 Box 10 – Alignment of LC–IMS–MS and IMS–MS imaging data

MZmine 3 enables the alignment of LC–(IMS–)MS and (IMS–)MS imaging data into a single, aligned feature list. By doing so, MS²-based identifications from the LC dataset can be used to increase the confidence in the annotation of imaging features. Such a workflow requires LC–MS and MS imaging data to be acquired from the same sample and processed according to [Procedure 1: LC–\(IMS–\)MS](#) and [Procedure 3: \(IMS–\)MS–imaging](#), respectively. The obtained feature lists are then aligned over the *m/z* and mobility (if available) dimension. **The figure below** illustrates an example of alignment of features detected in LC–IMS–MS and IMS–MS imaging data. In particular, panel a shows EICs of PC 34:1 and isobarically overlapping PC 34:1 13C2 and PC 34:0 (mobility-resolved). Panels b and c, show mobility-resolved images of PC 34:1 13C2 and PC 34:0. Panel d displays Overlaid mobilograms of LC–IMS–MS and MALDI–IMS–MS imaging data, solid and dashed lines,

respectively, of m/z 762.5945 ± 0.01 , which includes both lipids in the isobaric overlap. The solid or dashed grey line indicates where the EIMs were split in the LC or MALDI data to create the mobility-resolved features. Finally, aligned results can be visualised in a molecular network after performing FBMN (see [Procedure 1 – Step 23](#)) on the LC–IMS–MS *features* as described in Schmid et al., 2023⁹.



1489

1490

1491

1492 **Troubleshooting**

1493
 1494 Troubleshooting advice can be found in **Table 2**. We also recommend checking the MZmine website, where the
 1495 latest news are posted. Common issues and solutions are also described in the [online documentation](#).

1496
 1497 **Table 2.** Troubleshooting table.

Step	Problem	Possible reason and/or solution
All	User faces an issue not described in this protocol.	Go to the MZmine GitHub page and open an issue (https://github.com/mzmine/mzmine3/issues) that describes the problem. We recommend also including the MZmine log file.
All	The processing results obtained with the latest version of MZmine are inconsistent with those described in this protocol	This could be due to changes/updates in newer MZmine versions. To fully reproduce the results described in this protocol, download and use MZmine 3.5.0 (which was used in the preparation of this protocol)
All	The GUI freezes during remote desktop connection sessions.	This is a known bug of the Java GUI and will be resolved in future versions. Close and re-open MZmine solves the issue. If no errors are encountered, the processing is typically completed in the background.
All	The software uses all PC memory and crashes at various stages of the pipeline when processing large datasets.	When processing large datasets, we recommend applying a few measures (see Processing large datasets section) to minimise the memory consumption during the most computational-demanding steps.
Procedure 1-3 - Step 1	Data import fails.	Make sure the file format being imported is supported by MZmine (see online documentation). If so, ensure the original data files are not corrupted and/or no error is introduced during the file conversion.
Procedure 1 – Step 7	The software crashes at this step.	Enable the ‘Maximum features per thread’ option and set a small number (e.g. 10). Gradually increase the value, if needed.
Procedure 1 – Step 18	Library matches are expected, but none are retrieved	Ensure the correct MS level is used for the library search (e.g. GC–EI–MS data are generally stored as ‘MS level = 1’). To do so, double-click on the data files to open the ‘Raw data overview’. The table in the bottom panel

		contains information for every scan in the data file (see Extended Data Figure 2). The MS level is displayed in the 'MSn' column.
Procedure 3 - Step 1	Impossible to import <i>.imzML</i> files	The <i>.imzML</i> file converter used by MZmine requires an internet connection. In case the internet connection is working, ensure there is no proxy and that you are running MZmine with the administrative rights.

1498

1499 **Timing**

1500

1501 The time required to perform MS data processing can be divided into:

- 1502 - pipeline design and optimization of the processing parameters, which may take from a few minutes to
- 1503 several hours, based on the user's expertise, prior experience with the software, etc.; and
- 1504 - actual computing time, which mainly depends on the chosen pipeline, processing parameters, and
- 1505 hardware resources (e.g. the number of cores and RAM memory available).

1506 Therefore, the time required to perform *feature detection* on MS data cannot, in general, be estimated. In this
 1507 protocol, we provide example datasets and corresponding batch files to help non-experts replicating the
 1508 described procedures (see the [Reproducing the procedures with the 'Batch mode'](#) section). We anticipate this
 1509 to take up to one hour for new MZmine users.

1510

1511 **Anticipated results**

1512

1513 In this protocol, we describe how to use MZmine 3 to perform untargeted feature detection and annotation on
 1514 example datasets from three different MS platforms (i.e. LC-IMS-MS, GC-EI-MS, and IMS-MS imaging). A
 1515 batch file optimised for each example dataset is provided to reproduce the data processing described in each
 1516 procedure (see the **Required data** section). Although the same batch files cannot be used to process different
 1517 datasets without adaptation, they represent a good reference for new users and a starting point for parameter
 1518 optimization.

1519 The main outputs generated during feature detection and annotation in MZmine are represented by aligned
 1520 feature intensity tables (CSV format) and MS² spectral lists (MGF format). Aligned feature tables contain
 1521 information about the abundance of each feature across the different samples, as well as other chemical
 1522 annotations (e.g. isotopic pattern, adduct type, spectral library match). Feature list can be visualised and
 1523 explored in MZmine (**Fig. 9**). MS² spectral lists represent a summary of the fragmentation spectra associated
 1524 with each feature. Both these outputs are used by other third-party tools for further downstream analysis (see
 1525 the **Data export** section and [Procedure 1 – Steps 23, 24, and 25](#)). All batch files and corresponding output files
 1526 (feature lists and MS² spectral lists) produced by processing the example datasets are available in the
 1527 Supplementary Information.



1528
 1529 **Figure 9: Screenshot of a feature list visualised in MZmine.** The displayed columns can be changed by clicking the button in
 1530 the top-right corner, and the search filters can be used to control the displayed features.
 1531

1532 **Data availability**

1533
 1534 All example datasets used in this protocol are publicly available through the GNPS-MassIVE, MetaboLights and
 1535 Metabolomics Workbench repositories under the following accession numbers: [MSV000091634](https://www.ebi.ac.uk/metabolomics/workbench/studies/MSV000091634), Procedure 1,
 1536 LC-IMS-MS; [ST000981](https://www.ebi.ac.uk/metabolomics/workbench/studies/ST000981), Procedure 2, GC-EI-MS; [MSV000090328](https://www.ebi.ac.uk/metabolomics/workbench/studies/MSV000090328), Procedure 3, IMS-MS imaging;
 1537 [MSV000091642](https://www.ebi.ac.uk/metabolomics/workbench/studies/MSV000091642), lipid annotation (Procedure 1 – Step 20), LC-IMS-MS; [MTBLS265](https://www.ebi.ac.uk/metabolomics/workbench/studies/MTBLS265), export for statistics
 1538 (Procedure 1 – Step 20), LC-MS. The FBMN results can be accessed on GNPS at:
 1539 <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=ffd5aee568b54d9da1f3b771c459ebe5>.
 1540

1541 **Code availability**

1542
 1543
 1544 The latest release of MZmine can be downloaded from <https://www.mzmine.org>. The complete source code is
 1545 available at <https://github.com/mzmine/mzmine3/> under the MIT licence. The MZmine documentation is
 1546 hosted on GitHub and available at
 1547 <https://www.mzmine.org/documentation>.
 1548

1549 **Supplementary information**

1550
1551

1552 **Supplementary data**

1553 A ZIP archive file containing all batch files optimised for each example dataset and the corresponding data
1554 processing outputs (feature lists and MS² spectral lists).
1555

1556 **Acknowledgements**

1557

1558 T.P. is supported by the Czech Science Foundation (GA CR) grant 21-11563M and by the European Union's
1559 Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No. 891397.
1560 T.D. is supported by the European Regional Development Fund, Programme Johannes Amos Comenius (PJAC),
1561 project "IOCB MSCA PF Mobility" No. CZ.02.01.01/00/22_010/0002733. C.B. is supported by the Czech Academy
1562 of Sciences Program to Support Prospective Human Resources. A.S. and X.D. are supported by the National
1563 Institutes of Health grant U01CA235507. P.C.D. is supported by R01GM107550, R03OD034493, R01DK136117
1564 and NSF 2152526. We thank Fred Rooks and Michael FitzGerald for editing this manuscript.

1565 **Disclosures:** P.C.D. served as a consultant for DSM animal Health in 2023, is an advisor and has equity in Cybele,
1566 and is a co-founder, advisor and holds equity in Ometa, Enveda and Arome with prior approval by UC-San Diego.
1567 J.J.J.vdH. is a member of the Scientific Advisory Board of NAICONs Srl., Milano, Italy.

1568

1569 **References**

1570

- 1571 1. Alseekh, S. *et al.* Mass spectrometry-based metabolomics: a guide for annotation, quantification and best
1572 reporting practices. *Nat. Methods* **18**, 747–756 (2021).
- 1573 2. Da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proceedings*
1574 *of the National Academy of Sciences of the United States of America* vol. 112 12549–12550 Preprint at
1575 <https://doi.org/10.1073/pnas.1516878112> (2015).
- 1576 3. Müller, C., Binder, U., Bracher, F. & Giera, M. Antifungal drug testing by combining minimal inhibitory
1577 concentration testing with target identification by gas chromatography–mass spectrometry. *Nat. Protoc.*
1578 **12**, 947–963 (2017).
- 1579 4. Lisec, J., Schauer, N., Kopka, J., Willmitzer, L. & Fernie, A. R. Gas chromatography mass spectrometry–
1580 based metabolite profiling in plants. *Nat. Protoc.* **1**, 387–396 (2006).
- 1581 5. Chan, E. C. Y., Pasikanti, K. K. & Nicholson, J. K. Global urinary metabolic profiling procedures using gas
1582 chromatography–mass spectrometry. *Nat. Protoc.* **6**, 1483–1499 (2011).
- 1583 6. Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G. & Kell, D. B. Metabolomics by numbers:
1584 acquiring and understanding global metabolite data. *Trends Biotechnol.* **22**, 245–252 (2004).
- 1585 7. Aksenov, A. A., da Silva, R., Knight, R., Lopes, N. P. & Dorrestein, P. C. Global chemical analysis of biology
1586 by mass spectrometry. *Nature Reviews Chemistry* **1**, 1–20 (2017).
- 1587 8. Kompauer, M., Heiles, S. & Spengler, B. Atmospheric pressure MALDI mass spectrometry imaging of
1588 tissues and cells at 1.4- μm lateral resolution. *Nat. Methods* **14**, 90–96 (2017).
- 1589 9. Schmid, R. *et al.* Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nat. Biotechnol.*
1590 (2023) doi:10.1038/s41587-023-01690-2.
- 1591 10. Paglia, G., Smith, A. J. & Astarita, G. Ion mobility mass spectrometry in the omics era: Challenges and
1592 opportunities for metabolomics and lipidomics. *Mass Spectrom. Rev.* (2021) doi:10.1002/mas.21686.
- 1593 11. Vasilopoulou, C. G. *et al.* Trapped ion mobility spectrometry and PASEF enable in-depth lipidomics from
1594 minimal sample amounts. *Nat. Commun.* **11**, 331 (2020).
- 1595 12. Chang, H.-Y. *et al.* A Practical Guide to Metabolomics Software Development. *Anal. Chem.* **93**, 1912–1923
1596 (2021).
- 1597 13. Tautenhahn, R., Böttcher, C. & Neumann, S. Highly sensitive feature detection for high resolution LC/MS.
1598 *BMC Bioinformatics* **9**, 504 (2008).
- 1599 14. De Vijlder, T. *et al.* A tutorial in small molecule identification via electrospray ionization-mass
1600 spectrometry: The practical art of structural elucidation. *Mass Spectrom. Rev.* **37**, 607–629 (2018).
- 1601 15. Korf, A., Jeck, V., Schmid, R., Helmer, P. O. & Hayen, H. Lipid Species Annotation at Double Bond Position
1602 Level with Custom Databases by Extension of the MZmine 2 Open-Source Software Package. *Anal. Chem.*
1603 **91**, 5098–5105 (2019).
- 1604 16. Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure
1605 information. *Nat. Methods* **16**, 299–302 (2019).
- 1606 17. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products
1607 Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
- 1608 18. Pluskal, T. *et al.* CHAPTER 7:Metabolomics Data Analysis Using MZmine. in *Processing Metabolomics and*
1609 *Proteomics Data with Open Software* 232–254 (2020).
- 1610 19. Hammann, S., Korf, A., Bull, I. D., Hayen, H. & Cramp, L. J. E. Lipid profiling and analytical discrimination of
1611 seven cereals using high temperature gas chromatography coupled to high resolution quadrupole time-
1612 of-flight mass spectrometry. *Food Chem.* **282**, 27–35 (2019).
- 1613 20. Simon, C. *et al.* Mass Difference Matching Unfolds Hidden Molecular Structures of Dissolved Organic
1614 Matter. *Environ. Sci. Technol.* **56**, 11027–11040 (2022).
- 1615 21. Korf, A. *et al.* Digging deeper - A new data mining workflow for improved processing and interpretation of
1616 high resolution GC-Q-TOF MS data in archaeological research. *Sci. Rep.* **10**, 767 (2020).

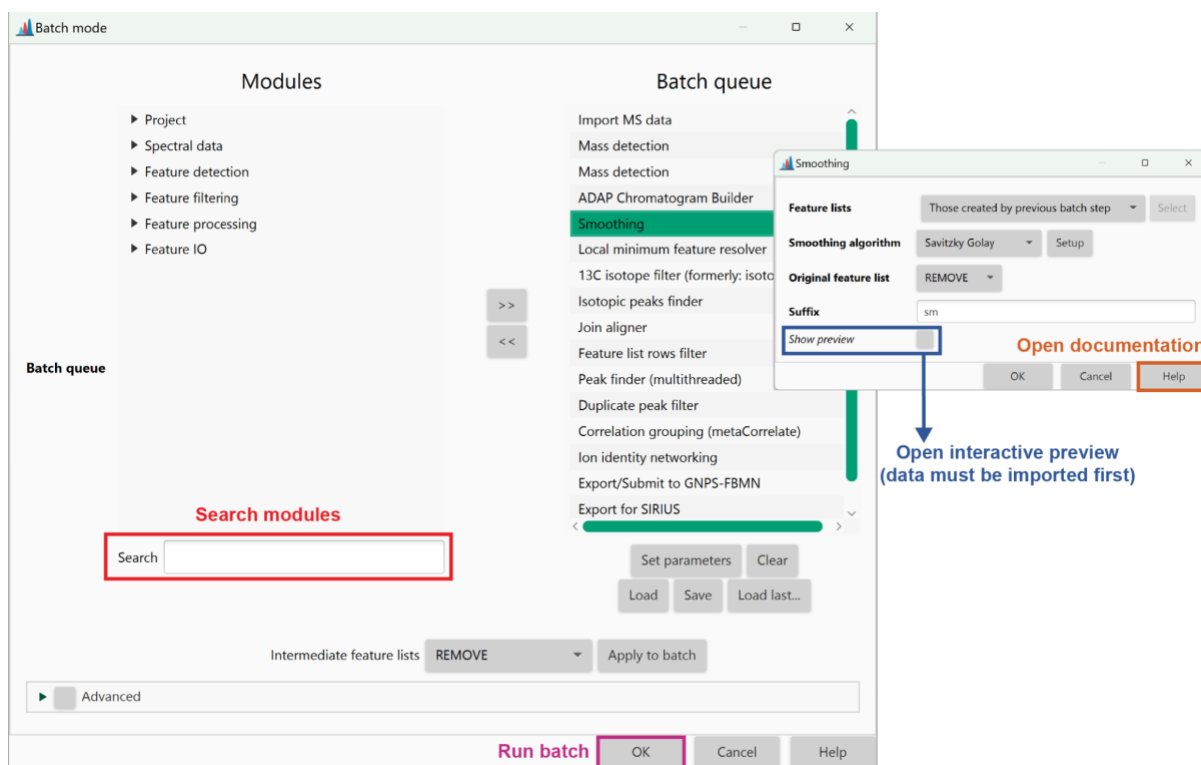
- 1617 22. Brungs, C. *et al.* Tattoo Pigment Identification in Inks and Skin Biopsies of Adverse Reactions by
1618 Complementary Elemental and Molecular Bioimaging with Mass Spectral Library Matching. *Anal. Chem.*
1619 **94**, 3581–3589 (2022).
- 1620 23. Wolf, C. *et al.* Mobility-resolved broadband dissociation and parallel reaction monitoring for laser
1621 desorption/ionization-mass spectrometry - Tattoo pigment identification supported by trapped ion
1622 mobility spectrometry. *Analytica Chimica Acta* vol. 1242 340796 Preprint at
1623 <https://doi.org/10.1016/j.aca.2023.340796> (2023).
- 1624 24. Deutsch, E. W. Mass spectrometer output file format mzML. *Methods Mol. Biol.* **604**, 319–331 (2010).
- 1625 25. Pedrioli, P. G. A. *et al.* A common open representation of mass spectrometry data and its application to
1626 proteomics research. *Nat. Biotechnol.* **22**, 1459–1466 (2004).
- 1627 26. Römpp, A. *et al.* imzML: Imaging Mass Spectrometry Markup Language: A common data format for mass
1628 spectrometry imaging. *Methods Mol. Biol.* **696**, 205–224 (2011).
- 1629 27. Rew, R. & Davis, G. NetCDF: an interface for scientific data access. *IEEE Comput. Graph. Appl.* **10**, 76–82
1630 (1990).
- 1631 28. Lu, M., An, S., Wang, R., Wang, J. & Yu, C. Aird: a computation-oriented mass spectrometry data format
1632 enables a higher compression ratio and less decoding time. *BMC Bioinformatics* **23**, 35 (2022).
- 1633 29. Adusumilli, R. & Mallick, P. Data Conversion with ProteoWizard msConvert. *Methods Mol. Biol.* **1550**,
1634 339–368 (2017).
- 1635 30. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.*
1636 **30**, 918–920 (2012).
- 1637 31. Smith, C. A., Want, E. J., O’Maille, G., Abagyan, R. & Siuzdak, G. XCMS: processing mass spectrometry data
1638 for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **78**,
1639 779–787 (2006).
- 1640 32. Röst, H. L. *et al.* OpenMS: a flexible open-source software platform for mass spectrometry data analysis.
1641 *Nat. Methods* **13**, 741–748 (2016).
- 1642 33. Tsugawa, H. *et al.* MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome
1643 analysis. *Nat. Methods* **12**, 523–526 (2015).
- 1644 34. Katajamaa, M., Miettinen, J. & Oresic, M. MZmine: toolbox for processing and visualization of mass
1645 spectrometry based molecular profile data. *Bioinformatics* **22**, 634–636 (2006).
- 1646 35. Kirkwood, K. I. *et al.* Utilizing Skyline to analyze lipidomics data containing liquid chromatography, ion
1647 mobility spectrometry and mass spectrometry dimensions. *Nat. Protoc.* **17**, 2415–2430 (2022).
- 1648 36. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci*
1649 *Data* **3**, 160018 (2016).
- 1650 37. Barker, M. *et al.* Introducing the FAIR Principles for research software. *Sci Data* **9**, 622 (2022).
- 1651 38. Haug, K. *et al.* MetaboLights--an open-access general-purpose repository for metabolomics studies and
1652 associated meta-data. *Nucleic Acids Res.* **41**, D781–6 (2013).
- 1653 39. Sud, M. *et al.* Metabolomics Workbench: An international repository for metabolomics data and
1654 metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.*
1655 **44**, D463–70 (2016).
- 1656 40. Meier, F. *et al.* Online Parallel Accumulation-Serial Fragmentation (PASEF) with a Novel Trapped Ion
1657 Mobility Mass Spectrometer. *Mol. Cell. Proteomics* **17**, 2534–2545 (2018).
- 1658 41. Whittemore, J. C., Stokes, J. E., Laia, N. L., Price, J. M. & Suchodolski, J. S. Short and long-term effects of a
1659 synbiotic on clinical signs, the fecal microbiome, and metabolomic profiles in healthy research cats
1660 receiving clindamycin: a randomized, controlled trial. *PeerJ* vol. 6 e5130 Preprint at
1661 <https://doi.org/10.7717/peerj.5130> (2018).
- 1662 42. Matyash, V., Liebisch, G., Kurzchalia, T. V., Shevchenko, A. & Schwudke, D. Lipid extraction by methyl-
1663 tert-butyl ether for high-throughput lipidomics. *J. Lipid Res.* **49**, 1137–1146 (2008).
- 1664 43. Chaleckis, R., Murakami, I., Takada, J., Kondoh, H. & Yanagida, M. Individual variability in human blood
1665 metabolites identifies age-related differences. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 4252–4259 (2016).

- 1666 44. Smith, R., Ventura, D. & Prince, J. T. LC-MS alignment in theory and practice: a comprehensive algorithmic
1667 review. *Brief. Bioinform.* **16**, 104–117 (2015).
- 1668 45. Pluskal, T., Uehara, T. & Yanagida, M. Highly accurate chemical formula prediction tool utilizing high-
1669 resolution mass spectra, MS/MS fragmentation, heuristic rules, and isotope pattern matching. *Anal.*
1670 *Chem.* **84**, 4396–4403 (2012).
- 1671 46. Nothias, L. F. *et al.* Feature-based molecular networking in the GNPS analysis environment. *Nat. Methods*
1672 **17**, 905–908 (2020).
- 1673 47. Olivon, F. *et al.* MetGem Software for the Generation of Molecular Networks Based on the t-SNE
1674 Algorithm. *Anal. Chem.* **90**, 13900–13908 (2018).
- 1675 48. Beniddir, M. A. *et al.* Advances in decomposing complex metabolite mixtures using substructure- and
1676 network-based computational metabolomics approaches. *Nat. Prod. Rep.* **38**, 1967–1993 (2021).
- 1677 49. Dührkop, K. *et al.* Systematic classification of unknown metabolites using high-resolution fragmentation
1678 mass spectra. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0740-8.
- 1679 50. Renai, L. *et al.* Combining Feature-Based Molecular Networking and Contextual Mass Spectral Libraries to
1680 Decipher Nutrimetabolomics Profiles. *Metabolites* **12**, (2022).
- 1681 51. Bazsó, F. L. *et al.* Quantitative Comparison of Tandem Mass Spectra Obtained on Various Instruments. *J.*
1682 *Am. Soc. Mass Spectrom.* **27**, 1357–1365 (2016).
- 1683 52. Lerno, L. A., Jr, German, J. B. & Lebrilla, C. B. Method for the identification of lipid classes based on
1684 referenced Kendrick mass analysis. *Anal. Chem.* **82**, 4236–4245 (2010).
- 1685 53. Sleno, L. The use of mass defect in modern mass spectrometry. *J. Mass Spectrom.* **47**, 226–236 (2012).
- 1686 54. Helmer, P. O., Korf, A. & Hayen, H. Analysis of artificially oxidized cardiolipins and monolyso-cardiolipins
1687 via liquid chromatography/high-resolution mass spectrometry and Kendrick mass defect plots after
1688 hydrophilic interaction liquid chromatography based sample preparation. *Rapid Commun. Mass*
1689 *Spectrom.* **34**, e8566 (2020).
- 1690 55. Müller, W. H. *et al.* Dual-polarity SALDI FT-ICR MS imaging and Kendrick mass defect data filtering for lipid
1691 analysis. *Anal. Bioanal. Chem.* **413**, 2821–2830 (2021).
- 1692 56. Korf, A. *et al.* Three-dimensional Kendrick mass plots as a tool for graphical lipid identification. *Rapid*
1693 *Commun. Mass Spectrom.* **32**, 981–991 (2018).
- 1694 57. Korf, A., Fouquet, T., Schmid, R., Hayen, H. & Hagenhoff, S. Expanding the Kendrick Mass Plot Toolbox in
1695 MZmine 2 to Enable Rapid Polymer Characterization in Liquid Chromatography–Mass Spectrometry Data
1696 Sets. *Analytical Chemistry* vol. 92 628–633 Preprint at <https://doi.org/10.1021/acs.analchem.9b03863>
1697 (2020).
- 1698 58. Elie, N., Santerre, C. & Touboul, D. Generation of a Molecular Network from Electron Ionization Mass
1699 Spectrometry Data by Combining MZmine2 and MetGem Software. *Anal. Chem.* **91**, 11489–11492 (2019).
- 1700 59. Zdouc, M. M. *et al.* FERMO: a Dashboard for Streamlined Rationalized Prioritization of Molecular Features
1701 from Mass Spectrometry Data. *bioRxiv* 2022.12.21.521422 (2022) doi:10.1101/2022.12.21.521422.
- 1702 60. Schmid, R. *et al.* Ion identity molecular networking for mass spectrometry-based metabolomics in the
1703 GNPS environment. *Nat. Commun.* **12**, 3832 (2021).
- 1704 61. Aron, A. T. *et al.* Reproducible molecular networking of untargeted mass spectrometry data using GNPS.
1705 *Nature Protocols* **15**, 1954–1991 (2020).
- 1706 62. Wang, M. *et al.* Mass spectrometry searches using MASST. *Nat. Biotechnol.* **38**, 23–26 (2020).
- 1707 63. da Silva, R. R. *et al.* Propagating annotations of molecular networks using in silico fragmentation. *PLoS*
1708 *Comput. Biol.* **14**, e1006089 (2018).
- 1709 64. Pang, Z. *et al.* MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights.
1710 *Nucleic Acids Res.* **49**, W388–W396 (2021).
- 1711 65. Pang, Z. *et al.* Using MetaboAnalyst 5.0 for LC–HRMS spectra processing, multi-omics integration and
1712 covariate adjustment of global metabolomics data. *Nat. Protoc.* **17**, 1735–1761 (2022).
- 1713 66. *FBMN-STATS: A hitchhiker's guide to statistical analysis of Feature-based Molecular Networks*
1714 (<https://chemrxiv.org/engage/chemrxiv/article-details/6540eb2548dad23120c52242>).

- 1715 67. Myers, O. D., Sumner, S. J., Li, S., Barnes, S. & Du, X. One Step Forward for Reducing False Positive and
1716 False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms
1717 for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks. *Anal. Chem.* **89**,
1718 8696–8703 (2017).
- 1719 68. Du, X., Smirnov, A., Pluskal, T., Jia, W. & Sumner, S. Metabolomics Data Preprocessing Using ADAP and
1720 MZmine 2. in *Computational Methods and Data Analysis for Metabolomics* (ed. Li, S.) 25–48 (Springer US,
1721 2020).
- 1722 69. Smirnov, A. *et al.* ADAP-GC 4.0: Application of Clustering-Assisted Multivariate Curve Resolution to
1723 Spectral Deconvolution of Gas Chromatography–Mass Spectrometry Metabolomics Data. *Anal. Chem.* **91**,
1724 9069–9077 (2019).
- 1725

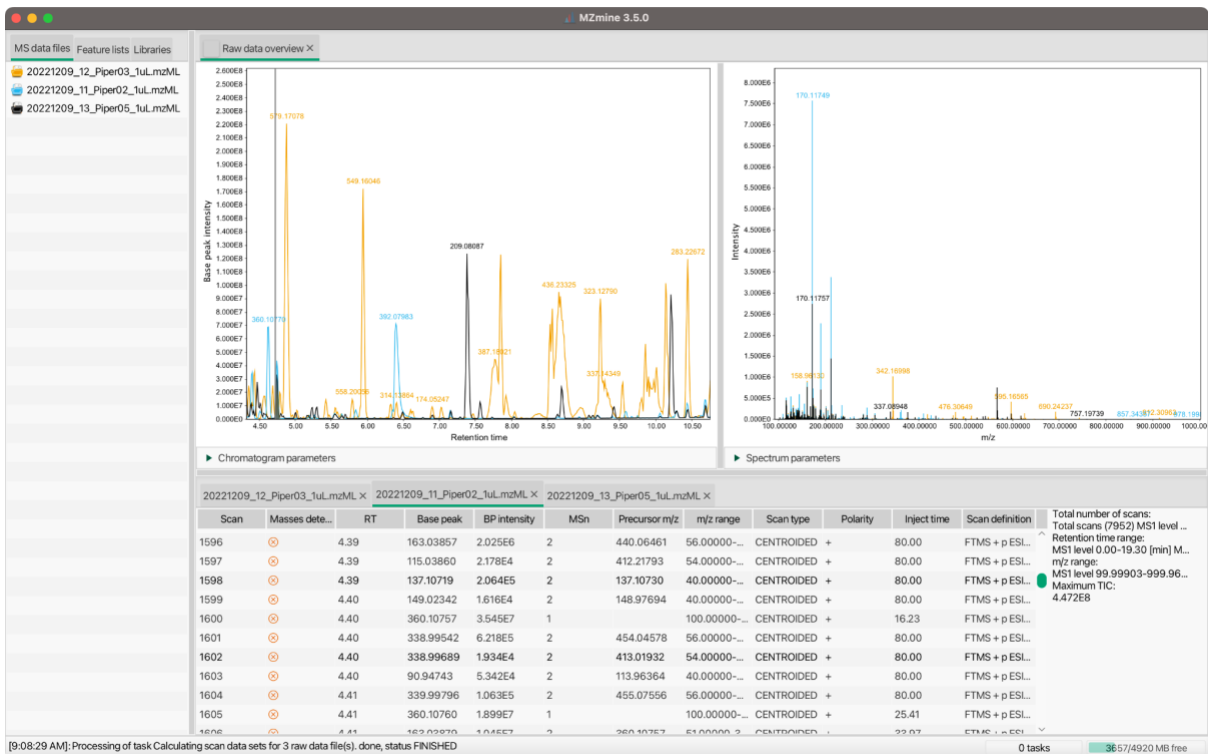
1726 **Extended data**

1727



1728

1729 **Extended Data Figure 1: Screenshot of the batch mode dialogue box. The current processing steps are**
1730 **displayed in the 'Batch queue' panel.** Additional steps can be selected from the 'Modules' panel and included
1731 using the double-arrows buttons. The current batch file can be saved using the 'Save' button whereas other
1732 batch files can be imported using the 'Load' button. Some modules offer a 'Show preview' option that can be
1733 opened by ticking the corresponding checkbox. For the preview to work, data must be already imported in
1734 MZmine. The online documentation for each processing module can be opened using the 'Help' button.
1735



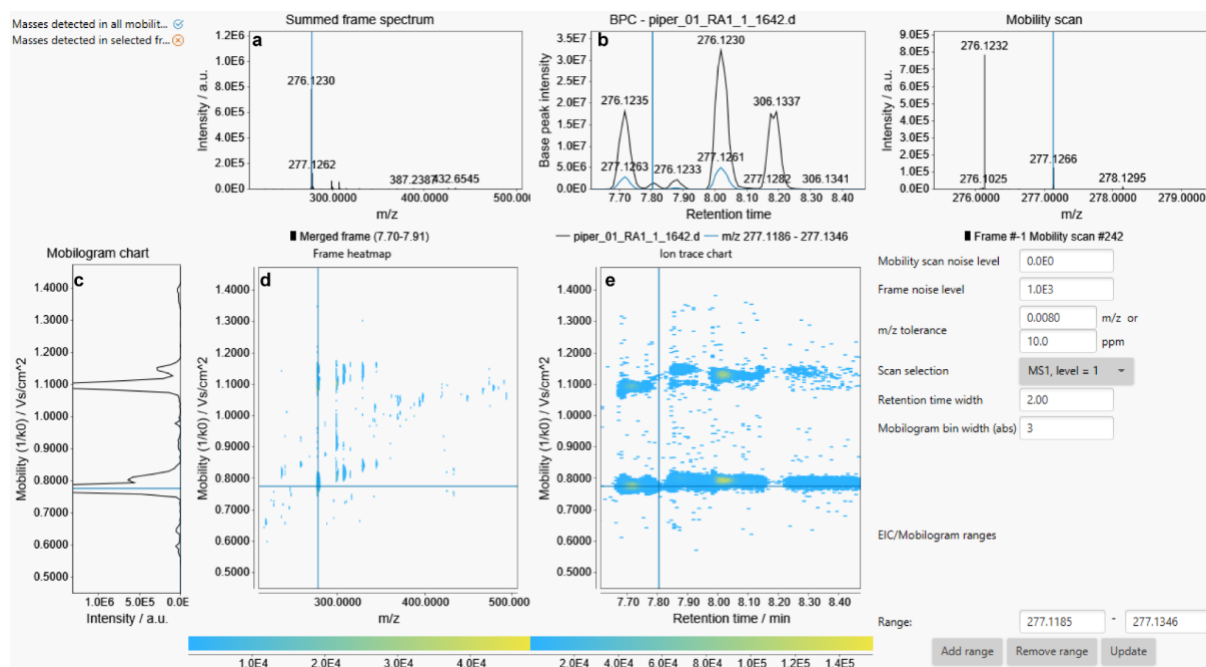
1736
1737
1738
1739
1740
1741

Extended Data Figure 2: Screenshot of the 'Raw data overview' module. The module displays three panels: chromatogram panel (left), mass spectrum panel (right) and scan information panel (bottom panel), which contains information for every scan in the data file.



1742
1743

1744 **Extended Data Figure 3: Screenshot of the interactive visualisation panel in the Local minimum resolver**
 1745 **module.** Two sub-panels are present: one for 'noisy' and one for 'good' EIC traces. The goal of the parameters
 1746 optimization is to ensure detection of true features while minimising 'noisy' peaks to be retained as features.
 1747 Feature lists and EIC traces to display can be chosen from the corresponding drop-down menus. Detected
 1748 features are colour-filled and resolved peaks are shown in different colours.
 1749



1750
1751
1752
1753
1754
1755

Extended Data Figure 4: Screenshot of the 'Ion mobility raw data overview' module. **a**, A summed frame spectrum with a blue indicator at the selected m/z. **b**, A chromatogram plot showing the BPC (black) and EIC (blue) of the selected m/z. The blue indicator shows the RT of the selected frame. **c**, A total ion mobilogram of the selected frame. **d**, A mobility vs. m/z heatmap of the selected frame. **e**, An ion mobility trace of the selected m/z in RT and mobility dimensions.