WiPP: Workflow for improved Peak Picking for Gas Chromatography-Mass Spectrometry (GC-MS) data

Nico Borgsmüller[1,2,3,‡], Yoann Gloaguen[1,2,3,‡], Tobias Opialla[2,3], Eric Blanc[1,4], Emilie Sicard[6], Anne-Lise Royer[5], Bruno Le Bizec[5], Stéphanie Durand[6], Carole Migné[6], Mélanie Pétéra[6], Estelle Pujos-Guillot[6], Franck Giacomoni[6], Yann Guitton[5], Dieter Beule[1,3,4], Jennifer Kirwan[2,3,*]

[1] Core Unit Bioinformatics, Berlin Institute of Health, 10178 Berlin, Germany
[2] Berlin Institute of Health Metabolomics Platform, 10178 Berlin, Germany
[3] Max Delbrück Center for Molecular Medicine in the Helmholtz Association, 13125 Berlin, Germany
[4] Charité – Universitätsmedizin Berlin, 10178 Berlin, Germany
[5] LABERCA, Oniris, INRA, Université Bretagne-Loire, 44307, Nantes, France
[6] Université Clermont Auvergne, INRA, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, 63000 Clermont-Ferrand, France
[‡] These authors contributed equally.
[*] Correspondence: Jennifer.Kirwan@mdc-berlin.de;

Abstract: Lack of reliable peak detection impedes automated analysis of large scale GC-MS metabolomics datasets. Performance and outcome of individual peak-picking algorithms can differ widely depending on both algorithmic approach and parameters as well as data acquisition method. Comparing and contrasting between algorithms is thus difficult. Here we present a workflow for improved peak picking (WiPP), a parameter optimising, multi-algorithm peak detection for GC-MS metabolomics. WiPP evaluates the quality of detected peaks using a machine learning-based classification scheme based on seven peak classes. The quality information returned by the classifier for each individual peak is merged with results from different peak detection algorithms to create one final high quality peak set for immediate down stream analysis. Medium and low quality peaks are kept for further inspection. By applying WiPP to standard compound mixes and a complex biological dataset we demonstrate that peak detection is improved through the novel way to assign peak quality, an automated parameter optimisation, and results integration across different embedded peak picking algorithms. Furthermore, our approach can provide an impartial performance comparison of different peak picking algorithms. WiPP is freely available on GitHub (https://github.com/bihealth/WiPP) under MIT licence.

1. Introduction

Metabolomics and related sciences use a combination of analytical and statistical approaches to qualitatively and quantitatively analyse the small molecules in a cell or biological system to answer biological questions [1,2]. Metabolomics benefits from maximizing the number of compounds detected in any individual analysis, while requiring concurrently that the results are robust and reproducible. Gas chromatography-mass spectrometry (GC-MS) is a common technology used in metabolomics research and contains information in both the chromatographic and the mass spectral space of the data [3]. Before downstream statistical and functional analysis, the data must first be pre-processed, such that individual peaks are identified, retained and catalogued in a numerical format, while irrelevant noise data should be removed. Attempting this by hand is a laborious process unsuited for epidemiological size datasets and also impedes reproducibility of this important data analysis step. Instead, this is commonly achieved by using one of a number of software options on the market e.g. XCMS [4], metaMS [5], MetAlign [6], mzMine [7], ADAP-GC [8,9], PyMS [10] and eRah [11]. It is commonly understood that there are still certain conditions in which these automated methods are sub-optimal, and the user-defined settings as well as some of the hard coded features of each software will have a large impact on the results [12]. However, each tends to have its strengths and weaknesses and will result in a slightly different result [8,12]. In this study, we sought to combine the strengths of different algorithm while minimizing the weaknesses.

In order to benefit from the strengths of each individual peak picking algorithm, we adopted a machine learning approach to classify peaks, enabling the automatic optimisation of user definable parameters for each algorithm and the combining of results. Machine learning uses statistical and pattern recognition strategies to progressively improve in their learning of data interpretation without requiring specific data interpretation programming [13]. Various forms of machine learning have previously been used for metabolomics studies including least squares-support vector machine, support vector machine regression, random forest and artificial neural networks [14–16]. Support vector machines (SVM) are a well known supervised machine learning models which are well suited for classification analysis [17]. Supervised learning uses an existing classified dataset(s) to train the model. This has the advantage that the resulting model is easy to optimise and validate [13].

We present a novel approach to automate peak picking in GC-MS data in order to optimise the reproducibility, accuracy and quality of the process by combining the strengths of multiple existing or new peak picking algorithms. We apply a visualization strategy combined with a support vector machine (SVM) based supervised machine learning approach to assess and learn peak quality. The learned model enables us to perform automatic optimisation of parameters for different peak picking algorithms. This is achieved by scoring the algorithm results on a representative data set. The learned model is also used to evaluate and integrate peaks suggested by two or more embedded (pre-existing) peak picking algorithms. The workflow result is a single high quality, high confidence peak set, which is suitable for immediate further downstream analysis. Once WiPP has been trained for a particular setting (resolution and sample type) it can quickly process large data sets automatically without further user involvement. The WiPP output format is designed to be easily searched by mass spectral libraries software such as NIST and can be subsequently analysed using standard statistical tools. It consists of a .csv and a .msp file detailing the individual peaks as chromatographic retention time or retention index - mass spectral fragment groups and their associated individual absolute mass spectral peak intensities.

2. Results

To demonstrate and evaluate the performance of the WiPP method we determined the recovery of compounds from a known mix at different concentrations (section 2.1). To quantify the added value we also compared WiPP results with those obtained from the individual algorithms used within WiPP. Furthermore we performed a case study using a complex biological data set: We compare results obtained using WiPP to a published independent analysis by third party (section 2.2). Readers less familiar with peak detection, algorithm benchmarking and machine learning approaches might benefit from first reading section 4.1 of Material and Methods which explains important concepts and defines relevant terms and necessary technical jargon.

2.1. Validation and Benchmarking

A known mix of commercially available standards was analysed at three different concentrations. Datasets were acquired on two different GC-MS instruments using different (high and low) resolutions using multiple replicates each (section 4.3.1). High quality peak sets were generated using the WiPP implementation utilising centWave and matchedFilter peak picking algorithms. The full pipeline was run, including a manual classification of training sets to generate SVM peak classifiers, parameter optimisation for both algorithms, and the filtering step as described by the methods section. The set of optimal parameters determined for each peak picking algorithm and both dataset is available in Table S1. For comparison the parameter optimisation of matchedFilter was also performed on the high resolution data using IPO [18], a tool optimising parameters based on isotopolgue detection. Overall, similar optimal parameters were found with one notable exceptions as shown in Table S1. Possible reasons for these exceptions are explored in the discussion section.

WiPP removes peaks classified with insufficient and intermediate quality from the final high quality peak set but always keeps them accessible to the user in a separate file. The following analysis was conducted only on the high quality peak set, i.e. does not consider the peaks of intermediate quality classes, because we do not consider their quantification fit for immediate down stream processing.

First, we compared the performance of WiPP with the embedded matchedFilter and centWave algorithm alone, always using the algorithm parameters determined by WiPP. The total number of detected peaks and their classification into either high quality or intermediate or low quality were analysed and contrasted (Figure 1). For example, in the high concentration, low resolution dataset 1, the total number of peaks detected was 137 for matchedFilter and 238 for centWave, of which 88 (59.9%) and 144 (60.5%) unique peaks were respectively classified as high quality. By contrast, in the high concentration, high resolution dataset 2, the total number of peaks detected was 2153 for matchedFilter and 997 for centWave, of which 280 (13.0%) and 181 (18.2%) unique peaks were respectively classified as high quality. Figure 1.A shows that in low resolution data, the number of peaks annotated as high quality and detected by centWave is on average 47% higher than the number of high quality peaks detected by matchedFilter. However, the output of the two algorithms do not entirely overlap and there are cases where matchedFilter detects high quality peaks which centWave does not and vice versa. Therefore, the number of peaks classified as high quality increases by merging the results of both algorithms. As the compound mix concentration increases, the number of high quality peaks found by both individual algorithms and in WiPP increases due to reduced number of peaks near the noise level. The number of peaks filtered out by WiPP (Figure 1.B) in low resolution data for matchedFilter and centWave represent on average 40% and 43% of the total number of peaks reported by the two algorithms respectively.

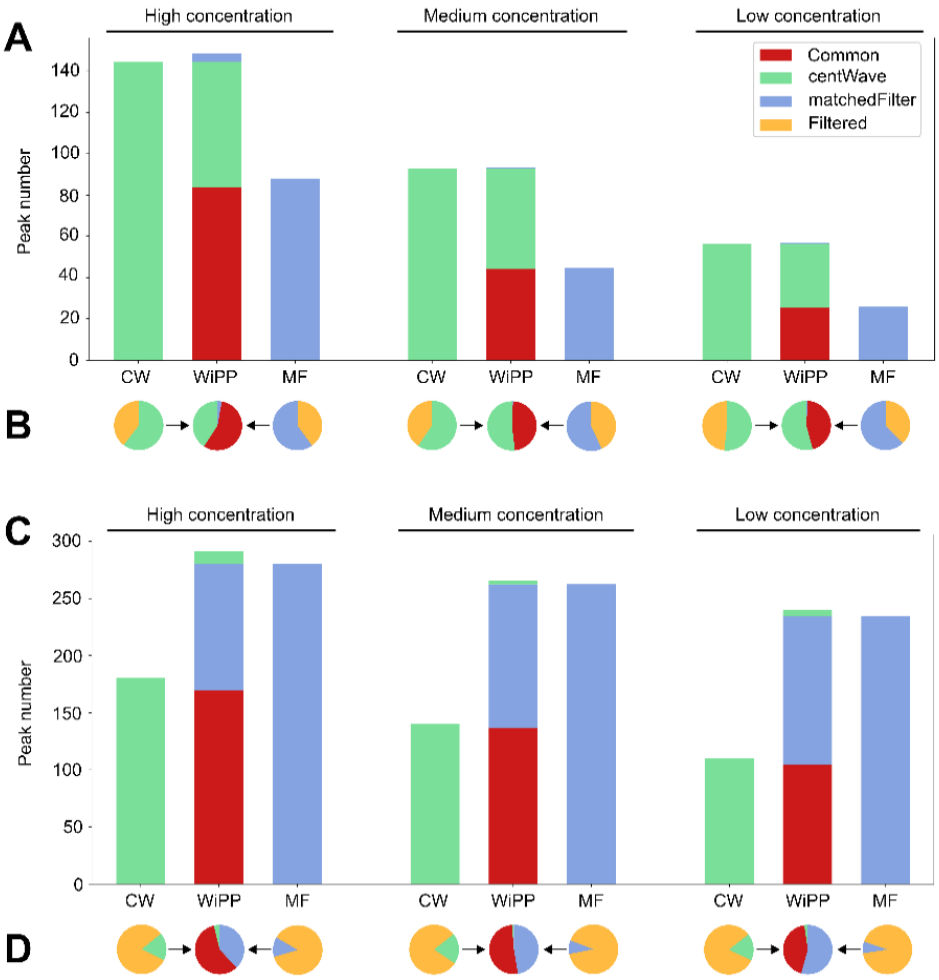

Figure 1. Number of peaks detected by individual algorithms on the high, medium and low concentrations of the standards mix dataset 1 in low resolution (A, B) and high resolution (C, D). A & C: Number of unique high quality peaks as classified by WiPP and their algorithm of origin. B & D: Proportion of peaks detected by centWave and matchedFilter rejected by at least one of the quality filters in WiPP.

Similarly, as for low resolution data, the use of centWave and matchedFilter together increases the number of high quality peaks detected in high resolution datasets. However, matchedFilter detects 44% more high quality peaks in comparison to centWave (Figure 1.C). The number of filtered peaks is very high in comparison to low resolution data, as they represent an average of 90% of the total number of peaks detected by matchedFilter and 80% of the total number of peaks detected by centWave (Figure 1.D).

Figure S2 explores the performance of the SVM classifier with respect to the size of the training data set. Please note that standard ROC curves are not applicable for the classifier evaluation due to the multi-class classification approach, i.e. classification does not depend on a single threshold that could be varied to generate different sensitivity and recall rates. Therefore, we generated confusion matrices (Figure S3) to assess if certain peak classes were misclassified systematically. The highest number of constantly misclassified peaks was found for matchedFilter in high resolution data, where 9% of peaks were wrongly classified as merged/shouldering and 10% of noise peaks were classified differently. The misclassified peaks were evenly distributed among the other classes, which lowers the overall result but does not distort it systematically.

Compound detection is not a built-in feature of WiPP, but the WiPP output format enables easy library matching using existing compound libraries. To test the ability of WiPP to report true compound related peaks, peaks detected and classified as high quality were annotated using our own internal library corresponding to the compound mix using reverse matching (supplementary materials). The output of the automated annotation implemented in WiPP was separated into two categories: high confidence annotation, requiring both the retention index (R.I.) to be within a 1.5 R.I. window and a spectra similarity score higher than 0.9 (supplementary materials), and low confidence annotation requiring only a spectral match to the internal library within the R.I. window. Manual annotation of datasets 1 and 2 was performed by an experienced mass spectrometrist and used as a gold standard to assess the ability of WiPP to detect known compounds. For comparison a manual annotation was generated according to our labs best practice. It consisted of data pre-processing and peak detection using Chromatof (Leco), followed by manual annotation using the in-house software Maui-via [19]. Parameters used in Chromatof for data pre-processing are available in Table S3.

The automated WiPP workflow achieves comparable performances to the manual annotation for medium and high concentration as summarized in Figure 2 (full details for all compounds and concentration are shown in Figure S4). WiPP delivers 95% of the manually annotated compounds in the high concentration samples and 86% of the compounds in the medium concentration samples. WiPP shows however, some limitation with low concentration data as only 42% of the metabolites are recovered.

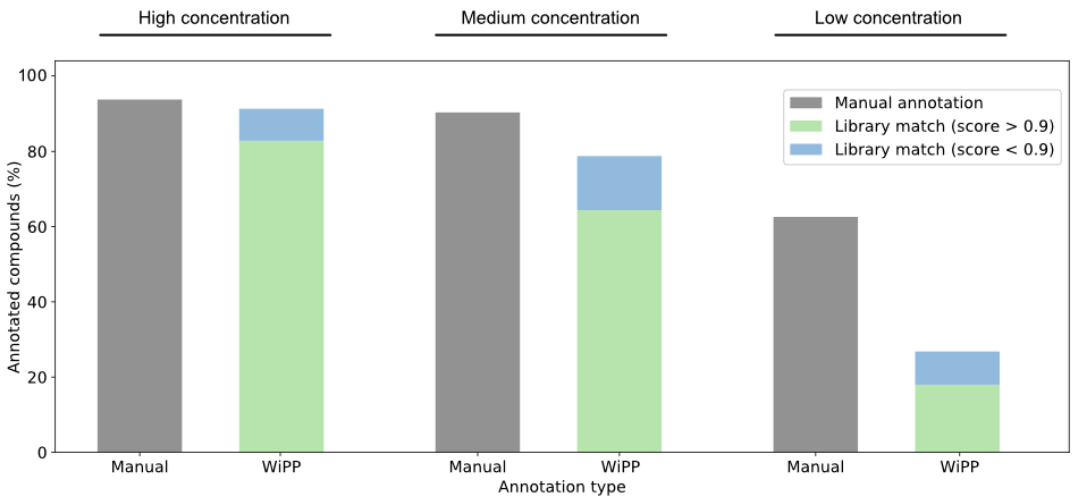

Figure 2. Number of metabolites detected and annotated manually or automatically by WiPP compared to the number compounds present in the three concentration of dataset 1.

2.2. Case Study

To further validate the results produced by WiPP, we ran the full workflow, including the generation of classifiers, on a publicly available biological dataset and results were compared to original results reported by the study. The selected study and data are introduced in section 4.3.2. WiPP classifiers were trained using a subset of the biological samples (2 samples from each biological conditions) as no pooled samples were available, and peak picking algorithms parameters were optimised using a different subset of the dataset (2 samples from each biological conditions). The final high quality peak set obtained by WiPP was annotated using the same spectral matching similarity score, reference masses, and intensities, as the original study [20]. Nine analytes were confirmed and manually validated in this study through targeted analysis. Using WiPP, six out of the nine analytes could be identified automatically with comparable calculated fold changes and

corresponding p-values (Table 1). The remaining three analytes were identified by WIPP but labelled and flagged as shouldering peaks, requiring user attention. For these we do not report fold changes and p-values because they are not automatically calculated by WiPP. Finally, one hexose, not reported in the original study, was found to be significantly different by WiPP. Our classifier labelled 100% of the peaks as high quality in all samples.

Table 1. Comparison of the results found by the published study to the one produced using WiPP automated workflow. (X) Data could not be automatically computed. (–) Missing data

ID	Identified (WiPP)	p-value (study)	Fold change (study)	p-value (WiPP)	Fold change (WiPP)
Glutamic acid	+	5.5×10^{-8}	1.9	1.5×10^{-4}	1.89
α-tocopherol	+	1.2×10^{-3}	1.5	7.7×10^{-3}	1.36
Valine	+	3.3×10^{-3}	1.5	3.0×10^{-2}	1.52
Citric acid	+	9.5×10^{-3}	-1.3	8.6×10^{-3}	-1.20
Sorbose	+	1.3×10^{-2}	-2.4	3.3×10^{-2}	-1.66
Cholesterol	+	3.5×10^{-2}	1.1	2.4×10^{-2}	1.10
Lactic acid	+	2.8×10^{-3}	-1.3	X	X
Leucine	+	1.8×10^{-2}	1.6	X	X
Isoleucine	+	4.2×10^{-2}	1.5	X	X
Hexose	+	–	–	4.0×10^{-2}	-1.61

3. Discussion

In this study, we present WiPP, a machine learning-based pipeline that enables the optimisation, combination and comparison of existing peak picking algorithms applied to GC-MS data. WiPP integrates machine learning classifiers to automatically evaluate the performance of peak picking algorithms and their selected parameters. WiPP also offers to the community a new approach to compare the performances of different peak picking algorithms, not only based on quantity but also on quality, and enables an automated parameter optimisation.

Our results show that WiPP produces comparable outputs to manually curation of data in an automated and thus more reproducible and scalable manner. For low concentration peaks the comparability is less convincing then for medium and high concentrations. This result indicates either a general incapability of embedded algorithms to detect low intensity peaks or algorithm parameters to be less suitable for low intensity peaks. The latter might be the case in our setting as parameters were optimised based on samples from all three concentration ratios, therefore fitting best for the entire dataset. Specific training and optimisation for detection of low concentration peaks would however be feasible with WiPP. We consider this use case to be less relevant because

low quality and low intensity peaks are also difficult to quantify and therefore add unwanted noise to statistical down stream modelling. Our main use case for WiPP currently is the automatic generation of a comprehensive set of reliably quantified high quality peaks that is suitable for immediate down stream analysis, as needed for large scale high sample number studies. Manual inspection of intermediate quality peaks is feasible and maybe useful to fully utilize a given data set as shown in the case study. However, our cases study also shows that an automated pipeline may discover features that may have escaped the attention of human specialists.

The bar plots in Figure S2 show the number of peaks reported by centWave and matchFilter that are considered noise upon human inspection. At least these peaks have to be considered false positives and are of high abundance especially in the high resolution data set. The yellow section of the bar chart in Figure 1.B and 1.D shows that substantial parts of the peaks generated by centWave and matchFilter are rejected for quality reasons. Our approach of using only high quality peaks (methods section) aims to achieve low false positive rates. Peaks classified by WiPP with intermediate quality are more likely to be false positive, while peaks classified as noise are likely to be false positives.

Automated classification of peak picking provides a novel way to assess and compare the performance of peak picking algorithms

We have developed a peak quality classification system that enables algorithm-identified peaks to be classified based on user trained peak characteristics related to both peak quality and accurate quantification e.g. apex shifted to a side, shoulder peak. We have set the number of distinct classes used for classification to 7. This is subjective but represents a balance between having enough classes to suitably differentiate between detected peaks while avoiding excessive manual annotation. Importantly, the current classifiers enable the reporting of complex peaks such as shouldering peaks to the user for manual inspection, avoiding potential loss of data. The number of classes could be altered to suit requirements if necessary but requires some changes in the WiPP source code and will also affect the time it takes to create the training set. The time taken to manually annotate the training set is an important consideration in the functional operation of WiPP and increases proportionally with the number of classes. The manual classification of peaks still has an element of user subjectivity to it, especially where a peak may fit into more than one category (e.g. too wide and skewed). We would recommend users to be consistent in their training classification of such peaks. Future versions of WiPP may seek to address this by enabling selection of multiple peak categories for an individual peak. Model training should be carried out by someone with a good knowledge of mass spectrometry data analysis. The user will imprint his know-how and judgement into the classification model.

Our peak classification method allows to assess the quality of peaks picked by individual algorithms, and thus enables a comparison of the relative performance of different peak picking algorithms. Peak detection in GC-MS data is a challenging and long-lasting problem. New approaches and tools emerge every year, yet there is still no established procedure to evaluate their performances objectively, and simple comparisons such as total number of peaks detected is not a robust metric for benchmarking purposes [12]. It is also influenced by the selection of algorithm specific parameters which leads to a certain subjective component when assessing each algorithm. We have demonstrated that WiPP can objectively assess the performance of multiple peak picking algorithms and is flexible enough that new algorithms can be added by the user, thus enabling future algorithm developers to objectively rate their algorithms against competitors.

Optimising parameters for peak picking

Currently, most peak picking algorithms require manual optimisation of parameters for every analysis. This is laborious and if not done can lead to suboptimal parameters being used to process datasets having a strong effect on the selected peaks [18]. It is noteworthy that the heatmap figures that illustrate the parameter optimisation strategy (Figure S1) also highlight the fact that the best parameters found for matchedFilter on the considered samples do not correspond to the parameters

that find the maximum number of peaks. In this specific case, the parameters displaying the highest number of peaks also find the highest number of high quality peaks. It comes, however, at the cost of an increased number of poor quality or false positive peaks compared to the best parameters returned by WiPP. An important consideration when dealing with poor quality peaks can be the accuracy of their integration for statistical purposes. We would argue that in most cases where statistical analysis is being conducted on the results, it is better to have a smaller number of robust and accurately quantified peaks than a larger number of peaks with high technical variation. Thus, we have optimised the balance between choosing the maximum number of high quality peaks while minimising the selection of poor quality peaks. The user can decide which approach to take for themselves by changing the weighting parameters of the scoring function. Optimal parameters returned by IPO are similar to those determined by WiPP with the notable exception of the FWHM value which is much greater in IPO. As the average full peak width of manually annotated peaks is 4 seconds, the FWHM value of 1 returned by WiPP appears to be more appropriate than the 8.8 value returned by IPO. A possible explanation is the technical differences between liquid and gas chromatography. Gas chromatography often suffers from column "bleed" at the end of an analytical run where large amounts of chemical substances elute from the column, seen as a characteristic increase in chemical baseline noise at the end of the run. (We speculate that this well-known characteristic of gas chromatography may be distorting the ability of IPO to find an appropriate FWHM value). As IPO has been designed for LC-MS data, it is not equipped to deal with characteristics that are specific to GC data. In our analysis, the vast majority of peaks detected after 2000 seconds are associated with noise and therefore penalised by the WiPP optimisation approach.

Improving overall quality of the final picked peak list

Interestingly, when optimised, centWave detects a higher number of true positive peaks than matchedFilter on low resolution data while the opposite is true for high resolution data. It is, however, important to note that the vast majority of peaks detected by matchedFilter on high resolution data are irrelevant (noise, duplicates or presenting less than 3 characteristic *m*/*z*) and increases as the concentration decreases. MatchedFilter therefore seems better at detecting low concentration peaks but at the expense of a higher poor quality peak or noise selection whereas CentWave algorithm is better at avoiding the selection of poor quality or noise peaks, but with a potential loss of sensitivity to peaks near the signal to noise threshold. The combination of both algorithms as implemented in WiPP shows in both low and high resolution data, a significant improvement on the coverage of peaks and compounds detected. These results clearly argue towards the use of several peak picking algorithms over a single one as previously shown [21].

While only centWave and matchedFilter were integrated so far, it is possible to integrate any peak picking algorithm to the workflow to further improve the coverage of detected high quality peaks. The modular architecture of WiPP, based on the python workflow framework Snakemake, enables new peak picking algorithms integration with little effort. The more peak picking algorithms are used, the longer the workflow runtime will be. Based on dataset 1 and 2 presented here, we estimate a 4-hour manual peak labelling process to generate the training data per algorithm, which must only be done once. The total runtime of the workflow is highly dependent on the computing power available and the range of parameters tested. For example, the full processing for dataset 1 can be completed overnight using 4 cores. This time can be further reduced by narrowing the parameter search space. For high resolution data, we recommend to use a high performance computing cluster (HPC) as the number of parameters tested increases significantly.

The overall results from the benchmarking process on a known mix of commercial standards and the replication of the workflow using a publicly available dataset shows that WiPP brings automated data analysis closer to the current gold standard that is manual curation using exclusively existing peak picking algorithms. In a context where large studies become routinely run in metabolomics laboratories, it is crucial to develop automated tools that can match manually validated standards. In this respect, these results also highlight that the shortest way to automation may lie in better using existing tools than creating new ones.

We have shown that WiPP improves current automated detection of peaks by:

1. Providing a novel way to classify peaks based on seven classes and thus objectively assess their quality.
2. Enabling objective performance comparison of different peak picking algorithms.
3. Enabling automated parameter optimisation for individual peak picking algorithm.
4. Enabling a final, improved high quality peak list to be generated for further analyses.
5. Reducing the operator-time required by packaging WiPP within a fully automated workflow (once the initial training of data is completed).

For the definition of a peak and the deconvolution process we need to distinguish WiPP from the embedded peak picking algorithms. For the classification WiPP operates on the full compound spectra, taking into account all measured m/z traces of the raw data within the peak retention time, but no deconvolution is applied at this stage. The definition of a peak and de-convolution used in the peak picking algorithms may vary for different algorithms, but WiPP is not affected by it. Indeed, this actually enables the combination of several peak picking approaches into WiPP. In the current implementation of WiPP, XCMS centWave and matchedFilter are used for peak detection and CAMERA [22] for the deconvolution of reported peaks. We hope that the permissive licence encourages the community to contribute by integrating additional peak picking algorithm. Compound identification is currently not a feature of WiPP and presents separate challenges. Instead WiPP generates output files in .msp format that can be used by common library matching based identification tools. Expanding WiPP to support Liquid Chromatography – Mass Spectrometry (LC-MS) data is desirable for future versions of WiPP, as this is the technology of choice for most untargeted analysis.

4. Materials and Methods

4.1. Peak Detection

We started the WiPP workflow definition by discussing the problems of peak detection. Myers et al. have shown, that, when applied to the same dataset, different peak picking algorithms return two different yet overlapping peak sets [8,12]. To assess the performance of a peak picking algorithm, we would like to consider the true peak set, which is unknown. The true peak set can be defined as the full set of peaks corresponding to all metabolites or contaminants (including metabolites adducts and fragments) present in a sample. The accurate quantification of a peak requires precise measurement of the peak area, which necessitates knowledge of peak centre and boundaries. A multitude of effects including but not limited to neighbouring peaks and technical drifts can transform and blur the peak shape making peak quantification problematic. This also means that there is a grey area between what is a true peak (i.e. not noise, but a distinct signal caused by a chemical) and what should be selected for further analysis. Figure 3 illustrates that algorithms select a proportion of the true peak set with a varying degree of success. For the purposes of this paper, we are defining a robust/high quality peak as a peak where the peak boundaries are accurately identified and demarcated and both the signal to noise and the intensity of the peak are sufficiently high to enable accurate peak intensity measurement, allowing for robust statistical downstream analysis. Algorithms can report lesser quality measurement of true peaks (e.g. by reporting two peaks as a single peak or a single peak as two, or incorrect assessment of peak boundaries). Furthermore, each algorithm will also report "peaks" that do not correspond to actual chemical signal (i.e. noise). Ultimately, before starting downstream analysis, a user defined filtering of what is considered "high enough" quality peaks must be defined. The schematic in Figure 3.A represents chromatographic peaks being accepted or rejected by two different peak picking algorithms. Figure 3.B and C illustrates that the peak sets returned by two different algorithms depends on the algorithm parameters used and have different overlap with the actual true peak set, which, due to the addition of chemical signals from contaminants, is normally unknown, even if working with known chemical standard mixes. Maximization of the coverage of the true peak set

can be achieved through an optimisation approach of the parameters of the peak detection algorithms (Figure 3.C). In the same manner, the number of false positive peaks (reported by the algorithm but not corresponding to chemical signal) returned by a peak detection algorithm also depends on the algorithm parameters. There is a trade-off between the number of false positives and false negative peaks and users may have different preferences depending on the kind of analysis they are working on.

The quality of the picked peaks may also be parameter dependent, for example, a true peak can be reported as two separate peaks (peak splitting) or two true peaks can be reported as a single peak (peak merging). Figure 3.D illustrates the objectives of WiPP which consists of optimising parameters for multiple peak picking algorithms, classifying of the reported peaks, and combining outputs of different algorithms to automatically produce a high quality peak set for further analysis.

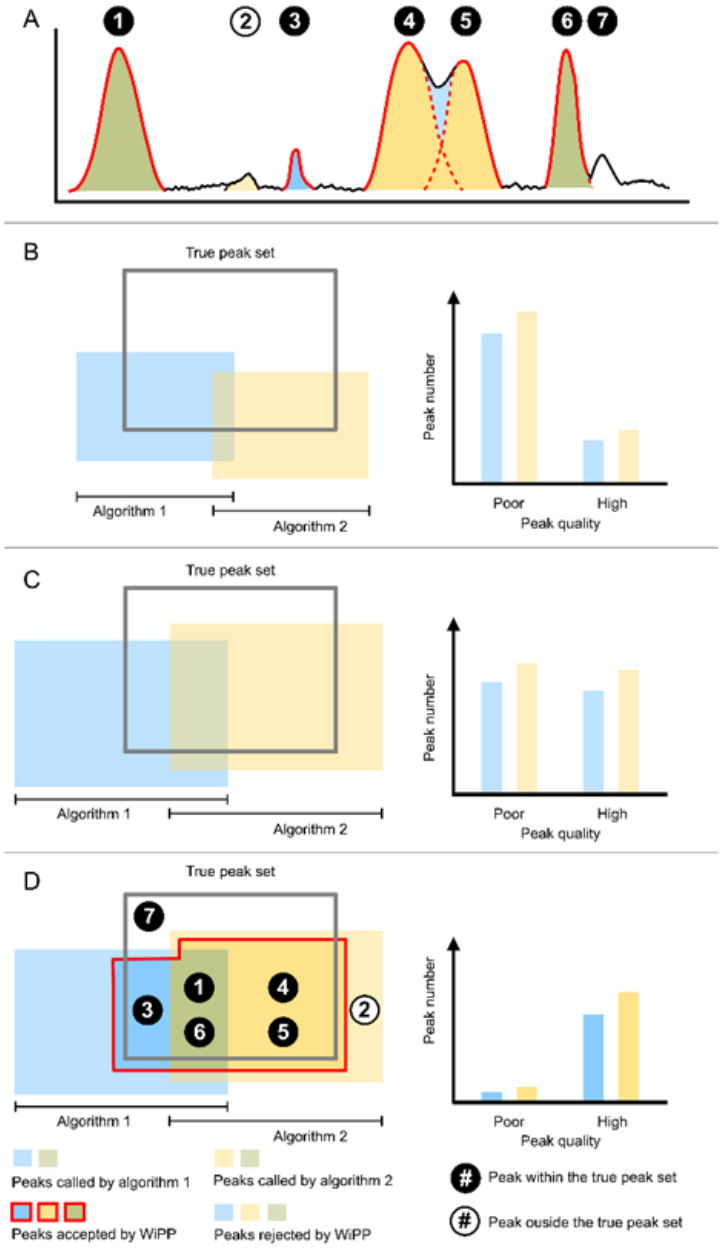

Figure 3. Schematic representation of the peak sets. **A.** Chromatographic representation of peaks detected by the two peak picking algorithms and accepted or rejected by WiPP. Peak 4 and 5 are erroneously detected by algorithm 1 as one single merged peak, hence the light blue colour between the distinct peaks properly detected by algorithm 2. **B.** Peak called by peak picking algorithm 1 and 2 compared to the true peak set of a dataset before parameter optimisation. **C.** Peak called by algorithm

1 and 2 after parameter optimisation. **D.** Peaks accepted and rejected by WiPP compared to the true peak set. Numbers represent the peak id from figure A and are placed in their respective regions in peak space.

4.1.2 Peak Classes

For the purpose of our workflow we define seven classes of peaks (Figure 4). Many criteria can be considered to define the different peak classes, we focus on the peak shape and peak boundaries. While Figure 4 shows schematic representation of a single *m/z* trace, WiPP operates on the full compound spectra, taking into account all measured *m/z* traces within the peak retention time. Special attention was paid to the boundaries as this heavily influences both the risk of inaccurate quantification (if peak area is used) and the risk of peak splitting and peak merging. Figure 4 shows a schematic representation of the seven classes established in WiPP based on the selected criteria. Each of these classes carries qualitative information about the peaks. We have designated classes A, B and C as being "high quality" peaks, class D describes noise signal and is considered as a false positive, while the last 3 classes, E, F and G represent intermediate quality true peaks which we consider not to be robust enough for downstream analysis. In WiPP, classes E, F and G are reported to the user for manual attention.

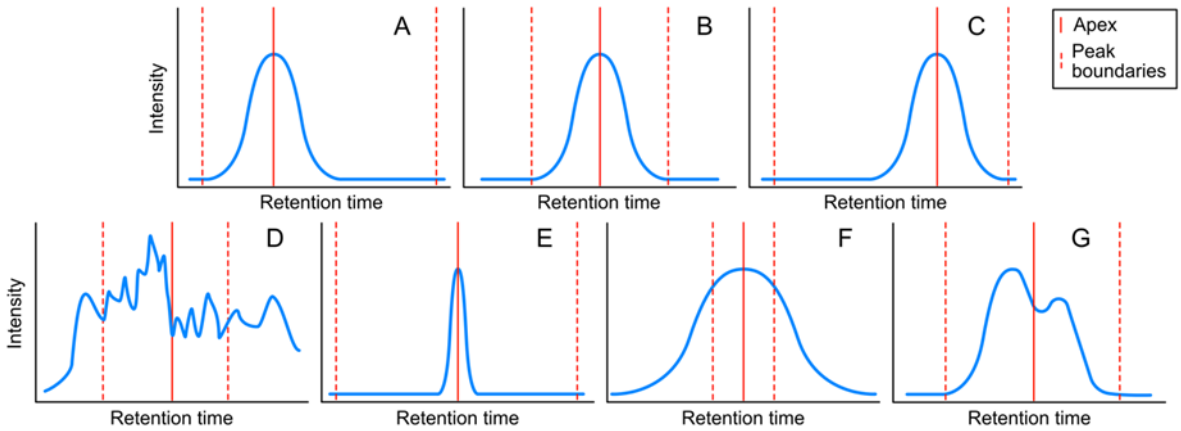

Figure 4. Schematic representation of the seven peak classes defined in WiPP. For clarity purposes, only one *m/z* is represented here. **A.** Apex shifted to the left. **B.** Centred apex. **C.** Apex shifted to the right. **D.** Noise. **E.** Peak with wide margins to window borders. **F.** Peak exceeds window borders. **G.** Merged/shoulder peak.

4.1.3 WIPP Workflow and Model

The proposed workflow (Figure 5) is composed of two main distinct parts, the training of the classifiers, and the high quality peak set generation. The supervised classifier training involves manual interaction. It should be performed at least once per instrument and sample type (i.e. blood, specific tissue, cell extract) but the same training dataset can then be used for all other analyses performed of this type. The final output of this first part is an instrument/sample type specific classifier for each individual peak detection algorithms. The second part of the workflow uses the trained classifier for unsupervised optimisation of the peak detection algorithms parameters. Subsequently, it generates a high quality peak set based on integrating results from the individual algorithms.

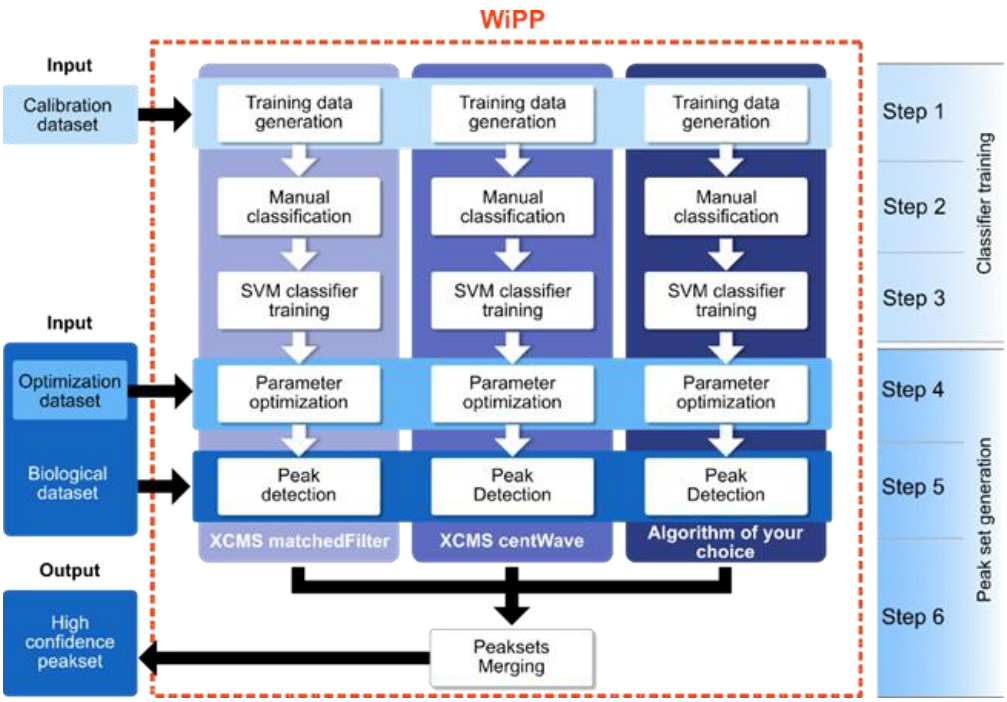

Figure 5. Flowchart of the WiPP method consisting of 6 steps. **Steps 1 to 3** consist of generating the training data and training the classifiers using a calibration dataset. **Step 4** optimises the parameters of individual algorithms using an optimisation dataset and the trained classifiers. **Step 5** run the optimised peak detection algorithms on the full biological dataset. **Step 6** classifies, filters and merged the outputs of individual peak picking algorithms to generate a high quality peak set.

The first step of the workflow aims at generating a training peak set containing a large variety of peaks differing in quality and intensity. For this purpose, we recommend to apply the algorithms to pooled or quality control samples using a wide range of parameters, and to manually classify a minimum of 700 peaks (which in our datasets equated to a minimum of 7 peaks in the smallest class) to generate the training set (Figure S2), henceforth called the calibration dataset. The parameter search ranges are user defined and should be set by an experienced user with prior knowledge on both the data produced by the instrument and the algorithm they are applying (see Table S2 for the parameters used in this study). A representative set of peaks for the supervised training is generated in step 2 by sampling algorithm parameters and retention time ranges (supplementary materials). WiPP provides an efficient peak visualization tool (Figure S5) allowing users to label each detected peak with one of the seven classes described in Figure 4. This important step allows expert users to imprint their individual peak classification know-how into the machine learning model. The labelled peaks form the training dataset that is used in step 3 of the workflow to train SVM classifiers. Every peak is described by an array of intensities within a certain *m/z* and retention time window. The peaks are baseline corrected, scaled and flattened to meet the input format required by the classifiers. During training, hyper-parameter optimisation [23] is performed using stratified cross-validation to avoid over-fitting.

The fourth step of the workflow performs an unsupervised optimisation of the algorithm-specific parameters for each of the peak picking algorithms. For this purpose, the number of peaks within each class is determined and a scoring function (supplementary materials) is applied that rewards high quality peaks while penalizing low quality peaks. Relative weighting can be user-defined to cater for different use cases, e.g. discovery studies or diagnostic studies (supplementary materials). To perform this unsupervised optimisation, WiPP generates a new peak set containing a large variety of peaks differing in quality and intensity. We recommend to apply the peak picking algorithm using different pooled or quality control samples than the one used for the training data generation (with a minimum of two samples). We call this peak set the optimisation dataset (Figure 5). The peaks detected by every single parameter set are classified using the

algorithm specific classifier and scored using the scoring function. We apply a simple grid search approach to determine the parameters returning the highest score. Those parameters are considered as optimal. We consider this method preferable to other alternatives; descent methods may lead to suboptimal solutions if the algorithm is trapped in local minima, and annealing methods are potentially computationally costly. As minima are generally shallow and broad, there is very little benefit in using more computationally costly methods. An example of the results for parameter optimisation performed by WiPP for matchedFilter and based on optimising the number of peaks per quality class is shown in Figure S1.

The following step (step 5) consists of running the peak detection algorithms with their optimal parameters on the full biological dataset.

Finally, a high quality peak set is generated in step 6 through a series of sub steps. First, the peaks detected by the different algorithms are classified using their respective classifiers. Next, simple filters such as class-based removal of duplicate peaks or rejection of peaks presenting less than n *m/z* are applied (n is set to 3 by default and can be user defined). The resulting algorithm specific peak sets are then merged, removing duplicate peaks where the peak sets overlap. The final peak set is composed of only high quality peaks, and the peaks predicted as low and intermediate quality are kept aside for optional further manual inspection.

4.2. Implementation and availability

The pipeline is implemented in python 3 using Snakemake [24], a reproducible and scalable workflow management system. WiPP offers a modular design to allow the addition of other existing or newly developed peak picking algorithms written in common programming languages (Java, python, R…). Currently centWave and matchedFilter peak picking algorithms are available in WiPP, eRah was also considered as an additional algorithm, but, in our hands, had memory issues when processing large datasets and was thus omitted from the current release. The pipeline can be run on local computers as well as on a high performance cluster. WiPP supports mzML, mzData and NetCDF input formats. It was tested with Ubuntu 16 and CentOS 7.6.1810 (Core). A comprehensive user manual and quick start guide are available on the GitHub repository. WiPP is released under the permissive MIT open source license and freely available at https://github.com/bihealth/WiPP, contributions and bug reports are welcome.

4.3. Data

4.3.1. Dataset 1 & 2

The first two datasets are made of an identical three-point dilution series (designated high, medium and low concentration) of a compound mix of 69 metabolites in known concentrations [25]. 9 samples of each dilution (1:1, 1:10 and 1:100) for a total of 27 samples were used to form the first dataset. These samples were prepared in duplicates to be run on two instruments with different resolutions (Pegasus 4D-TOF-MS-System: RP(FWHM) = 1290 at *m/z* = 219, and 7200 Q-TOF: RP(FWHM) = 14299 at *m/z* = 271,9867, see supplementary materials for details). Sample preparation and data acquisition details are available in the supplementary materials.

4.3.2. Dataset 3

The third dataset was collected by Ranjbar et al. and is publicly available in Metabolights repository [26] at https://www.ebi.ac.uk/metabolights/MTBLS105. The study evaluates changes in metabolite levels in hepatocellular carcinoma (HCC) cases vs. patients with liver cirrhosis by analysis of human blood plasma using GC-MS [20]. Briefly, data was collected using a GC-qMS (Agilent 5975C MSD coupled to an Agilent 7890A GC) equipped with an Agilent J&W DB-5MS column (30 m x 0.25 mm x 0.25 µm film 95% dimethyl/5% diphenyl polysiloxane) with a 10 m Duragard Capillary column with a 10 minute analysis using a temperature gradient from 60 °C. to 325 °C. Only 89 files generated in selected ion monitoring (SIM) mode were used for validation purposes here. Although SIM normally simplifies peak detection, in this dataset, there were often

multiple peaks detected for the same *m/z* meaning that there was still a peak detection issue to be addressed.

References

1. De Livera, A.M.; Dias, D.A.; De Souza, D.; Rupasinghe, T.; Pyke, J.; Tull, D.; Roessner, U.; McConville, M.; Speed, T.P. Normalizing and Integrating Metabolomics Data. *Analytical Chemistry* **2012**, *84*, 10768–10776.
2. Dunn, W.B.; Broadhurst, D.; Begley, P.; Zelena, E.; Francis-McIntyre, S.; Anderson, N.; Brown, M.; Knowles, J.D.; Halsall, A.; Haselden, J.N.; et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols* **2011**, *6*, 1060–1083.
3. Haggarty, J.; Burgess, K.E. Recent advances in liquid and gas chromatography methodology for extending coverage of the metabolome. *Current Opinion in Biotechnology* **2017**, *43*, 77–85.
4. Smith, C.A.; Want, E.J.; Maille, G.O.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry* **2006**, *78*, 779–787.
5. Wehrens, R.; Weingart, G.; Mattivi, F. metaMS: An open-source pipeline for GC–MS-based untargeted metabolomics. *Journal of Chromatography B* **2014**, *966*, 109–116.
6. Lommen, A. MetAlign: Interface-Driven, Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing. *Analytical Chemistry* **2009**, *81*, 3079–3086.
7. Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **2010**, *11*, 395.
8. Myers, O.D.; Sumner, S.J.; Li, S.; Barnes, S.; Du, X. One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks. *Analytical Chemistry* **2017**, *89*, 8696–8703.
9. Smirnov, A.; Jia, W.; Walker, D.I.; Jones, D.P.; Du, X. ADAP-GC 3.2: Graphical Software Tool for Efficient Spectral Deconvolution of Gas Chromatography–High-Resolution Mass Spectrometry Metabolomics Data. *Journal of Proteome Research* **2018**, *17*, 470–478.
10. O'Callaghan, S.; De Souza, D.P.; Isaac, A.; Wang, Q.; Hodkinson, L.; Olshansky, M.; Erwin, T.; Appelbe, B.; Tull, D.L.; Roessner, U.; et al. PyMS: a Python toolkit for processing of gas chromatography-mass spectrometry (GC-MS) data. Application and comparative study of selected tools. *BMC bioinformatics* **2012**, *13*, 115.
11. Domingo-Almenara, X.; Brezmes, J.; Vinaixa, M.; Samino, S.; Ramirez, N.; Ramon-Krauel, M.; Lerin, C.; Díaz, M.; Ibáñez, L.; Correig, X.; et al. eRah: A Computational Tool Integrating Spectral Deconvolution and Alignment with Quantification and Identification of Metabolites in GC/MS-Based Metabolomics. *Analytical Chemistry* **2016**, *88*, 9821–9829.
12. Myers, O.D.; Sumner, S.J.; Li, S.; Barnes, S.; Du, X. Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data. *Analytical Chemistry* **2017**, *89*, 8689–8695.
13. Friedman, J.; Hastie, T.; Tibshirani, R. *The elements of statistical learning*; Springer series in statistics New York, 2001; Vol. 1;.
14. Zhou, Z.; Tu, J.; Zhu, Z.-J. Advancing the large-scale CCS database for metabolomics and lipidomics at the machine-learning era. *Current Opinion in Chemical Biology* **2018**, *42*, 34–41.

15. Zheng, H.; Zheng, P.; Zhao, L.; Jia, J.; Tang, S.; Xu, P.; Xie, P.; Gao, H. Predictive diagnosis of major depression using NMR-based metabolomics and least-squares support vector machine. *Clinica Chimica Acta* **2017**, *464*, 223–227.
16. Khitan, Z.; Shapiro, A.P.; Shah, P.T.; Sanabria, J.R.; Santhanam, P.; Sodhi, K.; Abraham, N.G.; Shapiro, J.I. Predicting Adverse Outcomes in Chronic Kidney Disease Using Machine Learning Methods: Data from the Modification of Diet in Renal Disease. *Marshall Journal of Medicine* **2017**, *3*.
17. Hsu, C.-W.; Chang, C.-C.; Lin, C.-J. *A Practical Guide to Support Vector Classification*; 2003;
18. Libiseller, G.; Dvorzak, M.; Kleb, U.; Gander, E.; Eisenberg, T.; Madeo, F.; Neumann, S.; Trausinger, G.; Sinner, F.; Pieber, T.; et al. IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinformatics* **2015**, *16*, 118.
19. Kuich, P.H.J.L.; Hoffmann, N.; Kempa, S. Maui-VIA: A User-Friendly Software for Visual Identification, Alignment, Correction, and Quantification of Gas Chromatography–Mass Spectrometry Data. *Front. Bioeng. Biotechnol.* **2015**, *2*.
20. Nezami Ranjbar, M.R.; Luo, Y.; Di Poto, C.; Varghese, R.S.; Ferrarini, A.; Zhang, C.; Sarhan, N.I.; Soliman, H.; Tadesse, M.G.; Ziada, D.H.; et al. GC-MS Based Plasma Metabolomics for Identification of Candidate Biomarkers for Hepatocellular Carcinoma in Egyptian Cohort. *PLOS ONE* **2015**, *10*, e0127299.
21. Coble, J.B.; Fraga, C.G. Comparative evaluation of preprocessing freeware on chromatography/mass spectrometry data for signature discovery. *Journal of Chromatography A* **2014**, *1358*, 155–164.
22. Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T.R.; Neumann, S. CAMERA: An integrated strategy for compound spectra extraction and annotation of LC/MS data sets. *Anal Chem* **2012**, *84*, 283–289.
23. Chicco, D. Ten quick tips for machine learning in computational biology. *BioData Mining* **2017**, *10*, 35.
24. Köster, J.; Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **2012**, *28*, 2520–2522.
25. Pietzke, M.; Zasada, C.; Mudrich, S.; Kempa, S. Decoding the dynamics of cellular metabolism and the action of 3-bromopyruvate and 2-deoxyglucose using pulsed stable isotope-resolved metabolomics. *Cancer & metabolism* **2014**, *2*, 9.
26. Haug, K.; Salek, R.M.; Conesa, P.; Hastings, J.; de Matos, P.; Rijnbeek, M.; Mahendraker, T.; Williams, M.; Neumann, S.; Rocca-Serra, P.; et al. MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res* **2013**, *41*, D781–D786.