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
[‡] Authors contributed equally
[*] Corresponding author

Abstract: High false positive rates in GC-MS metabolomics peak detection is a common issue that impedes automated analysis of large-scale datasets. There is a growing need for improving the reliability and scalability of data analysis workflows. Many algorithms are available for peak detection [1], a crucial step for the data analysis, but performance and outcome can differ widely depending on both algorithmic approach and data acquisition method. This makes it difficult to compare and contrast between algorithms without extensive manual intervention.

We present a workflow for improved peak picking (WiPP), a parameter optimizing, multi-algorithm peak detection workflow for GC-MS metabolomics, which automatically evaluates the quality of detected peaks using machine learning-based classification. First, the classifier is trained to distinguish between real compound related peaks and false positive peaks. Then the algorithm parameters are scored based on the quality of detected peaks and optimized accordingly. This procedure is repeated for two peak detection algorithms and subsequently both algorithms are run in parallel on the entire data set with the optimized parameters. The qualitative information returned by the classifier for every peak is then used to merge individual algorithm results into one final high confidence peak set.

Using this approach, we show that automated detection and evaluation of peak quality is improved. The additional quantitative and qualitative information generated by the classifier allows:

1. a novel way to classify peaks based on seven classes and thus objectively to assess their quality
2. impartial performance comparison of different peak picking algorithms
3. automated parameter optimization for each individual peak picking algorithm
4. a final, improved high quality peak list to be generated for statistical or further analyses.

It achieves this while minimising the operator-time required by packaging this within a fully automated workflow. The modular design allows extension, adjustment and improvement of the workflow using different or additional peak detection algorithms and classifiers. Importantly, due to the fully automated implementation, the workflow is suitable for large-scale studies.

The pipeline supports mzML, mzData and NetCDF formats and is implemented in python using snakemake, a reproducible and scalable workflow management system, it is available on GitHub (https://github.com/bihealth/WiPP).

1. Introduction

We present a novel approach to automate peak picking in gas chromatography-mass spectrometry (GC-MS) data in order to optimize the 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 learning approach to automatically assess peak quality. This enables automatic optimization of parameters for peak picking on future datasets following a relatively brief manual training stage on test data acquired by the same instrument. The generated model is used for two or more peak picking algorithms whose results can then themselves be scored for quality. The results of this scoring allow the merging of the resulting peak lists to result in a final high quality dataset which maximizes peak number while minimizing false peak discovery. This enables large numbers of analyses to be processed automatically within a short time period with minimal user involvement.

Metabolomics and related sciences use a combination of analytical and statistical approaches to qualitatively and quantitatively analyze the small molecules in a cell or biological system to answer biological questions [2,3]. 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 is a common technology used in metabolomics research and contains information in both the chromatographic and the mass spectral space of the data [4].

In order that this information can be used for statistical 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 is ignored. Attempting this by hand is a laborious process unsuited for epidemiological size datasets. Instead, this is commonly achieved by using one of a number of software options on the market e.g. XCMS [5], metaMS [6], MetAlign [7], mzMine [8], ADAP-GC [9,10], PyMS [11] and eRah [12].

It is commonly understood that there are still certain conditions in which these automated methods are sub-optimal and the user-defined settings and some of the hard coding in the software will have a large impact on the results [13]. However, each tends to have its strengths and weaknesses and will result in a slightly different result [9,13]. In this study, we sought to optimize the strengths of each 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 optimization of user defined 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 [14]. 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 [15–17]. Support vector machines (SVM) is a well known machine learning method that uses a supervised learning approach that is well suited for classification analysis [18]. Supervised learning requires an existing classified dataset(s) to train the model. This has the advantage that the resulting model is easy to optimize and validate [14]. The training model is represented as datapoints in a mathematical n-dimensional space which can then be segregated into predetermined categories by the use of hyperplanes (in effect, classification decision boundaries that segment the space) [19,20].

The aims of this study were to (i) see whether machine learning approaches could be used to classify the quality of peak selection by existing automated methods and (ii) build an open-source software tool that could (a) use that knowledge to optimize parameter selection for individual existing algorithms so that each maximizes the quality and quantity of the data it finds in any individual analysis and (b) classify, score, combine and parse peaks identified by well constructed peak lists from individual existing algorithms to generate a final high quality master list. The final outputs consist 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 per mass fragment. The format is designed so that it can easily be searched

by mass spectral libraries such as NIST and can be subsequently analyzed using standard statistical tools.

2. Results

2.1. Validation and benchmarking

To evaluate the performance of the WiPP workflow detailed in the methods section (Figure 4), a known mix of commercially available standards was used and analysed at different concentrations and two resolutions (See datasets 1 and 2 in the methods section). High confidence peak sets for both high and low resolution datasets were generated in a demonstration of the standard WiPP workflow by using centWave and matchedFilter peak picking algorithms both of which are currently available in WiPP. Erah was also considered as an alternative peak picking algorithm, but, in our hands, had memory issues with processing large datasets. The full pipeline was run, including a manually annotated training set for the training of peak classifiers, parameter optimization for both algorithms, and a filtering step as described by the methods. An example of the results for parameter optimization as performed by WiPP for matchedFilter and based on optimizing the number of peaks per quality class is shown in Figure S1. The full set of optimal parameters found for each dataset and peak detection algorithm is available in Table S1. The accuracy of the SVM to recover manually annotated peaks against the number of peaks used for training is shown in Figure S2 (supplementary materials). ROC curves are not applicable for classifier evaluation due to the multi-class classification approach. Classification does not depend on a flexible threshold that could be varied to generate different sensitivity and recall rates. Therefore, we generated confusion matrices (Figure S3 in supplementary materials) to assess if certain peak classes were misclassified systematically. Parameters optimization of matchedFilter algorithm was also performed on the high-resolution data using IPO [21]. Overall, similar optimal parameters were found with notable exceptions as shown in Table S1 (supplementary materials). Possible reasons for these exceptions are explored further in the discussion. Peaks of insufficient or of intermediate quality were removed from the final peak set but kept accessible to the user in a separate file. The following analysis was conducted only on the high quality peak set as defined by peaks that were classified as belonging to groups A-C.

First, we compared the performance of matchedFilter and centWave after both algorithms had been optimised by WiPP. The total number of detected peaks and their classification of either high quality or insufficient (intermediate or low) quality were analysed and contrasted (Figure 1.A). 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 by WiPP's classifier and detected by centWave has a mean percentage 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 ranked as high quality increases when WiPP is used to merge the results of both algorithms. As the compound mix concentration increases, the number of high quality peaks found by both algorithms increases as may be expected 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.

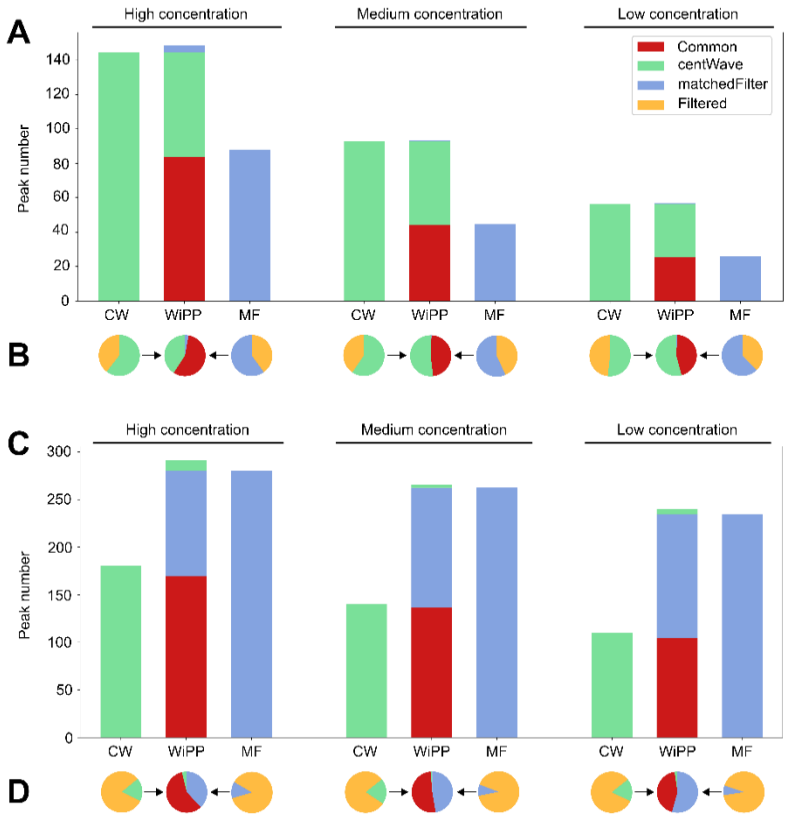

Figure 1. Number of peaks detected by individual algorithms on the high, medium and low concentrations of the standards mix dataset 1 in low resolution (A, B) and high resolution (C, D). A & C: Number of unique high quality peaks as classified by WiPP and their algorithm of origin. B & D: Proportion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

The automated annotation of high quality peaks reported by WiPP were then directly compared to the same data following manual annotation using the same library. The automated WiPP workflow achieves comparable performances to the manual annotation for medium and high concentration (See Figure S4). 95% of the manually annotated compounds are found by the pipeline in high concentration samples while 86% of the compounds are automatically recovered in medium concentration samples. WiPP shows however, some limitation with low concentration data as only 42% of the metabolites are recovered. However, since there is high probability that the missing compounds in question may not be able to be accurately quantified due to their low concentrations and therefore may add extra unwanted noise to statistical modelling, the reader can decide whether this may be beneficial to reduce the overall peak set in these conditions.

2.2. Case study

To further validate the results produced by WiPP, it was applied to a publicly available biological dataset and results were compared to original results reported by the study. The selected study is introduced in the data section as dataset number 3.

Nine analytes were confirmed and manually validated in the study by the targeted analysis. Using WiPP, six out of the nine analytes could be identified automatically and gave similar calculated corresponding p-values and fold changes to the original data (average and maximum absolute log fold change difference: 0.168 and 0.74. See Table 1 for details). The remaining three analytes were identified as the same analytes but WiPP labelled them as shouldering peaks and flagged them as requiring user attention, the fold changes and p-values are not reported here as they could not be automatically calculated. This is being addressed in version 2. Finally, one hexose, not reported in the study, was found to be significantly different by WiPP, whose classifier's labelled 100% of the peaks as high quality with no missing values.

Table 1. Comparison of the results found by the 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 optimization, combination and comparison of existing peak picking algorithms applied to GC-MS data. WiPP integrates a machine learning classifier to automatically evaluate the performance of a peak detection algorithm and its selected parameters. Our results show that WiPP produces comparable outputs to manually generated data in an automated manner. WiPP also offers to the community a new approach to compare the performances of different peak picking algorithms and enable an automated parameter optimization.

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 whether they display a number of common peak characteristics related to both peak quality and accurate quantification e.g. apex shifted to a side, merged or shoulder peak. We have set the number of classes used for the WiPP classifier at 7 distinct classes. This is subjective but represents a balance between having enough classes to suitably classify 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 can be altered to suit requirements if necessary but requires minor changes in the code which will affect the time it takes to create the training set. The time taken to manually annotate the original training set is an important consideration in the functional operation of WiPP and increases proportionally with the number of classes used to classify data. 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 for example). We would recommend users are consistent in their treatment of such peaks and this part is carried out by someone with a good knowledge of mass spectrometry. Future versions of the program may seek to address this by enabling selection of multiple peak categories for an individual peak or automating this process more.

This peak classifying method allows the quality of peaks picked by individual algorithms to be classified and, in addition, it makes a comparison of the relative performance of different peak algorithms possible. Peak detection in Gas Chromatography Mass Spectrometry 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 [13]. It is also influenced by the selection of appropriate 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 be able to objectively rate their algorithms against the market leaders.

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. Parameter selection has previously been shown to have a strong effect on the selected peaks [21]. It is noteworthy that the heatmap figures that illustrate the parameter optimization strategy (Figure S1) also highlight the fact that the best parameters found for matchedFilter and 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 scoring function. 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 and 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 algorithm 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 due to 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 LCMS 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 optimization 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 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 [22].

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 work to programmers and bioinformaticians. 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 runtime of dataset 1 on 4 cores can be completed overnight. This time can be brought down by narrowing the parameter search as a large one was used in our example for demonstration purposes. For high resolution data, we recommend to use an high performance computing cluster (HPC) as the number of parameters tested can increase 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, and this using exclusively existing tools. 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 to assess their quality
2. Enabling objective performance comparison of different peak picking algorithms
3. Enabling automated parameter optimization for each individual peak picking algorithm
4. Enabling a final, improved high quality peak list to be generated for statistical or further analyses.
5. Reducing the operator-time required to achieve this by packaging this within a fully automated workflow (once the initial training of data is completed).

4. Materials and Methods

We started the workflow development by defining 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 [9,13]. To assess the performance of a peak picking algorithm, we must also 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 one sample. For accurate quantification of the peak it requires precise measurement of the peak area which necessitates knowledge of the peak center and boundaries. This 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 bioinformatical and statistical analysis. The

schematic in Figure 2 demonstrates that the algorithms select a proportion of the true peakset with a varying degree of success as to how robust the definition of each individual peak is. 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 analysis. Algorithms can report lesser quality measurement of true peaks (eg. 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 (eg. noise). Ultimately before starting downstream analysis, a user defined threshold of what is considered "high enough" quality peaks must be determined. The schematic in Figure 2.A represents chromatographic peaks detected by two different peak picking algorithms that are accepted or rejected by our Workflow for improved Peak Picking (WiPP). Figure 2.B and C illustrates that the high quality peak sets returned by two different algorithms depends on the algorithm parameters used and has different overlap with the actual true peakset, 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 optimization approach of the parameters of the peak detection algorithms (Figure 2.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. The quality of the resulting picked peaks may also be parameter dependent, for example, the likelihood that a true peak will be reported as two separate peaks (peak splitting) or two true peaks reported only as a single peak (peak merging). Figure 2.D illustrates the objectives of the Workflow for improved Peak Picking (WiPP) which consists of optimizing parameters for initial peak reporting for multiple peak selection algorithm, including classifying the reported peaks, and ultimately combining outputs of different algorithms to produce a high quality peakset for further analysis.

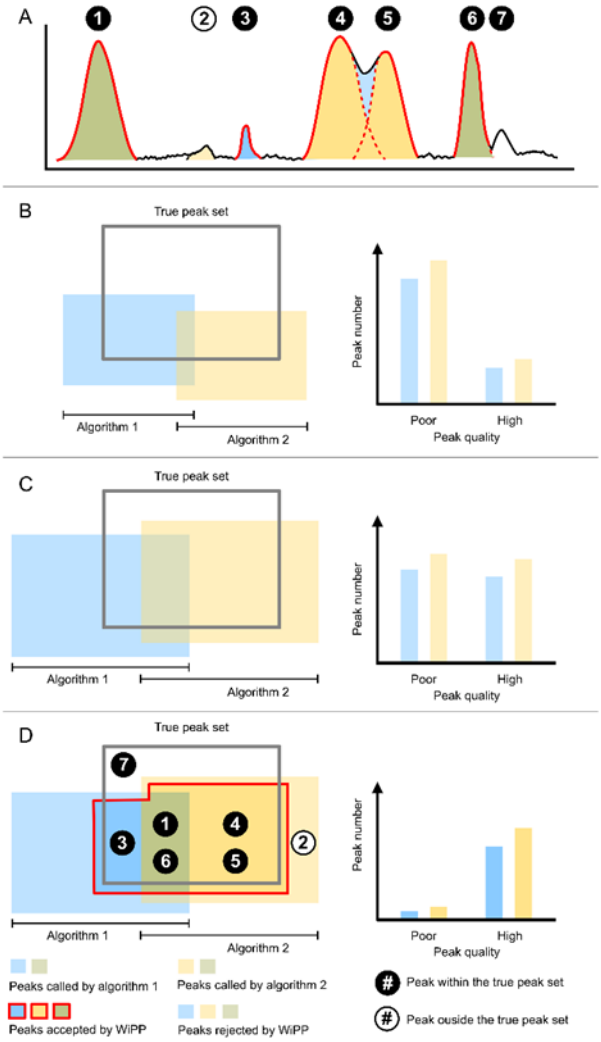

Figure 2. Schematic representation of the peak sets. **A.** Chromatographic representation of peaks detected by the two peak picking algorithms and accepted or rejected by WiPP. Peak 4 and 5 are erroneously detected by algorithm 1 as one single merged peak, hence the light blue color between the distinct 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 optimization. **C.** Peak called by algorithm 1 and 2 after parameter optimization. **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.

For the purpose of our workflow we define seven classes of peaks (Figure 3). Many criteria can be considered to define the different peak classes, we focus on the peak shape and peak boundaries. While Figure 3 shows schematic representation of a single *m/z* trace, our workflow 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 3 shows a schematic representation of the seven classes established in WiPP based on the selected criteria. Each of these classes carry qualitative information about the peaks which enable the assessment of the peak set returned by peak detection algorithms. 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.

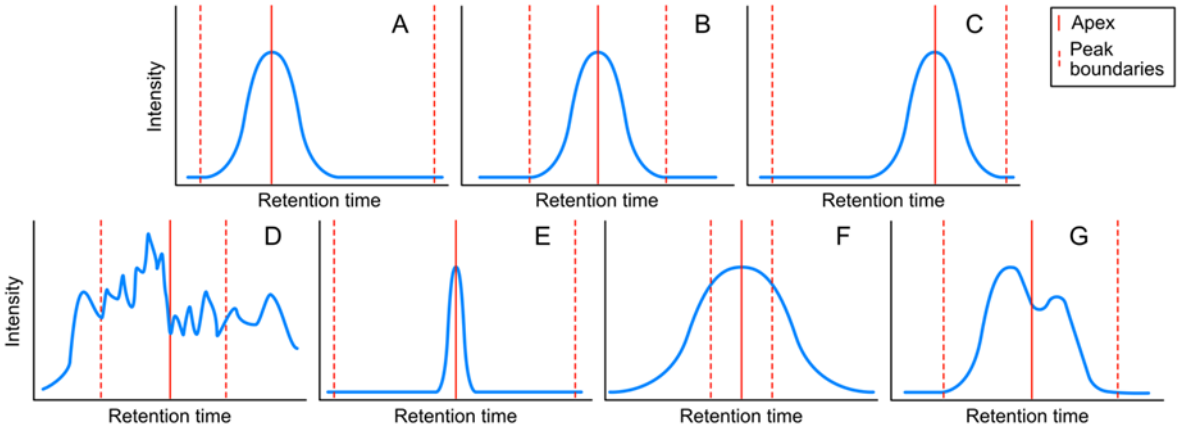

Figure 3. Schematic representation of the seven peak classes defined in WiPP. For clarity purposes, only one *m/z* is represented here. **A.** Apex shifted to the left. **B.** Centered apex. **C.** Apex shifted to the right. **D.** Noise. **E.** Peak with wide margins to window borders. **F.** Peak exceed window borders. **G.** Merged/shoulder peak.

4.2. Workflow and model

The proposed workflow is composed of two main distinct parts, the training of the classifiers (one for each algorithm), and the peak set generation (Figure 4). The supervised classifier training involves manual interaction to create a training dataset required to generate the peak classifier. 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 optimization of the peak detection algorithms parameters. Furthermore, it generates a high confidence peak set based on integrating results from multiple algorithms.

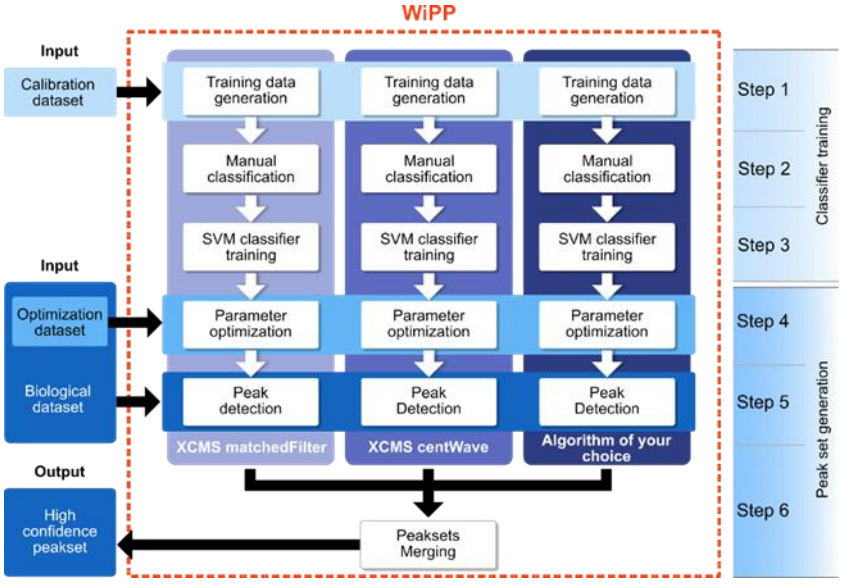

Figure 4. Flowchart of the WiPP method consisting of 6 steps. **Steps 1 to 3** consist of generating the training data and training the classifiers using a calibration dataset. **Step 4** optimizes the parameters of individual algorithms using an optimization dataset and the trained classifiers. **Step 5** run the optimized 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 manually annotate 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 (see Figure S2 in supplementary material), 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 in supplementary material for our choice for the dataset 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 (See supplementary material for details). WiPP embeds a peak visualization tool (see Figure S5 in supplementary material) allowing users to label each individually presented peak with one of the seven classes described in Figure 3. 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, hyperparameter optimization [23] is performed using stratified cross-validation to avoid over-fitting.

The fourth step of the workflow performs an unsupervised optimization of the algorithm-specific parameters for each of the peak picking algorithms. For this purpose, the number of peaks for the individual classes is determined and a scoring function 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 (See supplementary materials for details). To perform this unsupervised optimization, 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 optimization dataset (See Figure 4). The output peaks generated by every single parameter set are classified using the algorithm specific classifier and a score is assigned using the scoring function. We apply a simple grid search approach to determine the parameters returning the highest score. Those parameters are then 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 shallow and broad, there is very little to gain in using more computationally costly methods.

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

Finally, a high confidence 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 high quality peaks, and the peaks predicted as low and intermediate quality are kept aside for optional further manual inspection.

4.3. Implementation and availability

The pipeline is implemented in python 3 using Snakemake [24], a reproducible and scalable workflow management system, and is available on GitHub (https://github.com/bihealth/WiPP). Connection with R is enabled through system calls from Snakemake enabled to run R-based peak detection algorithms. 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…). The pipeline can be run on local computer as well as on HPC for big datasets. WiPP supports mzML, mzData and NetCDF formats. It was tested under Ubuntu 16 and CentOS 7.6.1810 (Core). A comprehensive user manual and quick start guide are available on the GitHub repository for detailed instructions on how to use the pipeline. WiPP is released under the permissive MIT License.

4.4. Data

As a proof of principle to demonstrate the function of the workflow, two datasets comprising of commercially bought standards acquired on two different GC-MS instruments using different resolutions were used. A complex biological dataset collected and analyzed independently was reanalyzed using WiPP to validate the workflow with more complex sample matrices. Further details on the datasets are given below.

4.4.1. 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.

Compound detection is not a built-in feature of WiPP, but the WiPP output enables it to be easily searched using existing compound libraries. For testing the ability of WiPP to detect known peaks, peaks detected and classified as true positive were annotated using our own internal library corresponding to the compound mix using reverse matching (see supp methods).

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 900/1000 (see supplementary materials for details), and low confidence annotation requiring only a spectral match to the internal library.

Manual annotation: 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 peaks. The manual annotation consisted of data pre-processing and peak detection using Chromatof (Leco), followed by manual annotation using an in-house script Maui-via [26]. Parameters used in Chromatof for data pre-processing are available in Table S3.

4.1.2. Dataset 3

The third dataset was collected by Ranjbar et al. and was taken by us from the publicly available Metabolights repository [27] 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 gas chromatography coupled with mass spectrometry (GC-MS) [28]. The full details and protocol of the sample preparation and data acquisition methods have been taken by us from the Metabolights repository and are available with the data files in the same place. 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.

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 (centWave and matchedFilter) parameters were optimized using a different subset of the dataset (2 samples from each biological conditions). The standard WiPP workflow of parameter optimizing peak picking, machine learning peak quality classification and finally merging the results of the two algorithms was applied to result in a final high quality peak set. Finally, peaks were then identified, independent of the WiPP workflow, using the same spectral matching similarity score, reference masses, and intensities, as the original study [28].

References

1. Spicer, R.; Salek, R.M.; Moreno, P.; Cañueto, D.; Steinbeck, C. Navigating freely-available software tools for metabolomics analysis. *Metabolomics : Official journal of the Metabolomic Society* **2017**, *13*, 106.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. Lommen, A. MetAlign: Interface-Driven, Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing. *Analytical Chemistry* **2009**, *81*, 3079–3086.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. Friedman, J.; Hastie, T.; Tibshirani, R. *The elements of statistical learning*; Springer series in statistics New York, 2001; Vol. 1;.
15. 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.

16. 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.
17. 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*.
18. Hsu, C.-W.; Chang, C.-C.; Lin, C.-J. *A Practical Guide to Support Vector Classification*; 2003;
19. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A training algorithm for optimal margin classifiers. In Proceedings of the Proceedings of the fifth annual workshop on Computational learning theory; ACM, 1992; pp. 144–152.
20. Shawe-Taylor, J.; Sun, S. A review of optimization methodologies in support vector machines. *Neurocomputing* **2011**, *74*, 3609–3618.
21. 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.
22. 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.
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. 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*.
27. 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.
28. 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.