

This is the preprint version of the contribution published as:

Müller, E., Huber, C.E., Brack, W., Krauss, M., Schulze, T. (2020):

Symbolic aggregate approximation improves gap filling in high-resolution mass spectrometry data processing

Anal. Chem. **92** (15), 10425 – 10432

The publisher's version is available at:

<http://dx.doi.org/10.1021/acs.analchem.0c00899>

Symbolic aggregate approximation improves gap filling in high resolution mass spectrometry data processing

Erik Müller,^{*,†,‡} Carolin Elisabeth Huber,^{†,‡} Werner Brack,^{†,‡} Martin Krauss,[†] and
Tobias Schulze^{*,†}

[†]*UFZ - Helmholtz Centre for Environmental Research, Permoserstraße 15, 04318 Leipzig,
Germany*

[‡]*RWTH Aachen University, Institute for Environmental Research, Worringerweg 1, 52074
Aachen, Germany*

E-mail: erik.mueller@ufz.de; tobias.schulze@ufz.de

Phone: +49 341 235 4736; +49 341 235 1083

Abstract

Nontargeted mass spectrometry (MS) is widely used in life sciences and environmental chemistry to investigate large sets of samples. A major problem for larger-scale MS studies is data gaps or missing values in aligned data sets. The main causes for these data gaps are the absence of the compound from the sample, issues related to chromatography or mass spectrometry (for example broad peaks, early eluting peaks, ion suppression, low ionization efficiency), and issues related to software (mainly limitations of peak detection algorithms). While those algorithms are heuristic by necessity and should be used with strict settings to minimize the number of false positive and negative peaks in a data set, gap filling may be used to reduce missing data in single

samples remaining after peak detection. In this study, we present a new gap filling algorithm. The method is based on the symbolic aggregation approximation (SAX) algorithm that was developed for the evaluation and classification of time series in data mining studies. We adopted SAX for liquid chromatography high resolution MS non-target screening to support the detection of missing peaks in aligned mass spectral data sets. The SAX-based algorithm improves the detection efficiency considerably compared to existing gap filling methods including the Peak Finder algorithm provided in MZmine.

Introduction

Gas/liquid chromatography high resolution mass spectrometry (GC/LC-HRMS) is a ubiquitous technology in environmental sciences, metabolomics, lipidomics and other fields where natural or synthetic molecules need to be detected. In the past, the focus of mass spectral analysis was primarily on structure elucidation of single compounds and targeted quantitative analysis of substances with known specific chromatographic retention times and mass-to-charge ratios. However, already in the 1970s first studies using nontargeted analysis (NTS) were published.¹ The availability of modern and affordable high resolution and accuracy mass spectrometers accelerated the application of NTS in all relevant domains.² NTS is an analytical approach that does not require a priori knowledge on the precursor masses obtained during the MS data acquisition and strives to detect all compounds that are contained in the sample. In practice, it covers all substances that can be acquired under the experimental and instrumental setup in dependence on the intrinsic physico-chemical properties and concentrations of the compounds. NTS helps to detect and prioritize unknown substances and can be used in conjunction with other methods to gain a more holistic understanding on chemical signals related to biological effects of substances even if the identity of the effect drivers is unknown. A basic processing of chromatography-mass spectrometry data across multiple samples can be summed up in two main computational steps after ac-

quisition: Peak picking and alignment.³ Peak picking (or peak/feature detection) is done by utilizing an algorithm that tries to find chromatographic peaks in certain mass-to-charge windows. These peaks are caused by the presence of specific compounds. The intensity curve of a compound measured in these regions usually results in a characteristic shape resulting from the chromatographic separation and is called a peak. Alignment (sometimes called “grouping”) algorithms aggregate peaks with similar m/z and retention times across different samples by binning them into consensus m/z and retention times across all samples in the data set. The binning is usually controlled by the setting of retention time and mass difference limits as well as weighting factors. The goal of alignment is to get an overview on the presence and absence of single peaks over all samples in the analysis, which also includes mitigating the retention time shifts that occur when measuring a large number of samples over time.⁴ One of the main challenges of automated peak detection is that it is heuristic by necessity, because the underlying data are complex and can differ strongly even between samples collected in the same study (e.g. an environmental study having water samples from within highly populated cities as well as from nature reserves). While nontargeted peak detection algorithms produce results with a high precision, they usually also have a high false negative rate, but a low false positive rate.³ This means that, given a sensible parametrization, many peaks present in the data remain undetected, but those revealed are reliable true positive detections. The main reason for this is that most peak detection algorithms assume peaks to have a minimum length and an approximately Gaussian shape. While it is possible for mass spectrometry noise to create a Gaussian shape by accident, it is very unlikely that this happens over several seconds. This results in the algorithm usually correctly identifying Gaussian shapes in the chromatogram as peaks, meaning that the false positive rate is low. However, many peaks do not have a shape close enough to Gaussian to be detected by those algorithms. This happens especially at lower intensities, where the noise is higher and “masks” the underlying Gaussian shape, resulting in many false negatives in the aligned peak list. While different peak detection and alignment algorithms are implemented

in proprietary and open source software packages such as XCMS,^{3,5,6} MZmine,⁷ OpenMS⁸ or enviMass,⁹ only few of these packages provide procedures to refine the peak detection by exploiting the information from the aligned peak (or feature) lists (e.g., XCMS, MZmine). These so called gap filling approaches reanalyze those positions with gaps in single samples with correspondence to all samples with detected peaks in the same row of the alignment matrix. If the extracted information supports the assumption of a peak at that position, this missing peak is imputed into the aligned peak list independent from the peak detection settings. The heuristic nature of the peak picking algorithms does not allow for gap filling during peak detection, because peak pickers can a priori only obtain peaks in the strict limits of its settings in order to create reliable results. In combination with peak alignment, gap filling can be used to mitigate some of the problems stemming from the peak detection procedures. However, gap filling algorithms are rarely implemented and none of them have been published. The most intuitive and simple approach is to indiscriminately integrate the extracted ion chromatogram (EIC or XIC) region (i. e. in the m/z and retention time window) of a chromatogram at the same retention time window at which a peak is found in other samples and assume that any signal within this window is always caused by the same compound. Another approach for gap filling is the one used in the “Peak Finder” method included in MZmine,⁷ which represents a more sophisticated form of the simple peak detection; it does not integrate EICs indiscriminately, but with a simple set of criteria for peaks, such as a notable local maximum being present. A completely alternative method to gap filling is data imputation – filling data gaps without reevaluating the EICs, but based on the patterns present in the known values. Data imputation is a common approach to handle missing values specifically in metabolomics MS,¹⁰ using a wide array of different imputation strategies. For example, MetaboAnalyst¹¹ offers a variety of imputation methods based on several statistical and machine learning strategies such as singular valued decomposition (SVD) and the k-nearest neighbours algorithm (kNN).¹² Several studies have shown the effectiveness of imputation on metabolomics MS data,^{13,14} especially using Random forests^{10,15} but it is

clear that imputation approaches are only feasible as long as the initial number of missing values is low, since any model for the imputation of missing data necessarily becomes worse if the initial data basis for the extrapolation is too small. This means that while imputation approaches are useful for some data sets, they would be too inaccurate in cases where the number of missing values is too high. This especially holds for many environmental nontarget MS studies, which create data matrices with many more missing values than detected signals, since many substances are site- or sample-specific. It is mathematically implausible to impute realistic values for these substances if the data basis is too small to build a model on. In this study, we propose a new algorithm for filling information gaps in alignment data sets that is robust with regard to the false positive and false negative rate. This procedure is based on the symbolic aggregate approximation (SAX) algorithm described by Lin et al.^{16,17} Our approach differs from other “gap filling” methods by building a more in-depth model of the characteristic peak shape of each individual signal, thus also being more robust to noise or other fluctuations of the chromatographic peak shape. By assigning each EIC a letter sequence representing its shape in a simplified form, we can compare EICs at similar m/z and retention time positions to each other in a reasonable time without sacrificing either specificity or sensitivity. In order to evaluate the performance of the SAX-based algorithm, it was compared with the performance of simple gap filling and the only other non-simple gap filling algorithm, the “Peak Finder” algorithm implemented in MZmine.

Existing gap filling algorithms

One of the main challenges of describing and evaluating the new algorithm in a scientific context is that the other two previously existing “gap filling” algorithms have not been published or peer-reviewed in any journal before. Each come with certain drawbacks and strengths, that – to our knowledge – have not been mentioned in scientific publications before. Since the traits of these algorithms are part of the main motivation why we developed a new type of gap filling algorithm, we describe those traits in this publication.

Simple gap filling algorithms

Most existing gap filling algorithms, such as the “Same m/z and RT range gap filler” from MZmine⁷ or the “fillPeaks”-method from XCMS^{3,5,6} indiscriminately integrate intensities at some predicted m/z and retention time ranges for a certain sample at chromatogram positions where a peak is reasonably suspected. This means that all m/z and retention time ranges where a peak was found in any sample get integrated over all samples. This approach is based on the assumption that all substances are technically present in all samples and that by integrating all signals in the same m/z and retention time regions, these substances can be considered in the analysis. This approach results in many false positive detections across the samples, since LC-HRMS measurements often contain regions with string background noise resulting from the sample matrix. While this approach can be effective if the false positive rate of detection is far less important than the false negative rate, it falls short when a large set of peaks is post-processed automatically. For targeted mass spectrometry or other studies that concern a small set of substances and/or peaks, this approach is a good choice, since the drawbacks of this method do not get exacerbated by the data set. Most peaks in targeted mass spectrometry studies get manually checked, so a false positive detection can be easily rectified. The main problem caused by this form of gap filling is that – especially for NTS and its large number of aligned peaks – many false positive peaks are included by this process. This is caused by the fact that peak detection methods will still pick some noise regions as peaks in every sample. As this noise is not as often picked as peaks in other samples, this results – for most data sets – in an alignment with many ostensible peaks that only occur in a single sample. If these false peaks get integrated in all other samples as well (which is likely, as background noise occurring from the mass spectrometer is moderately consistent across samples) this results in a huge number of false positive peak detections caused by a small number of initially wrongly picked peaks.

Peak Finder

The Peak Finder algorithm is the only widely-used¹⁸⁻²¹ non-simple gap filling algorithm and it is integrated into MZmine. At its core, it follows the same principle as the simple gap filling algorithms (i.e. it re-examines EICs if one can reasonably suspect a previously unpicked peak to be there due to the information gained in the alignment), but it works differently from the simple gap filling in that it does not integrate the EIC regions indiscriminately, but with additional criteria. The Peak Finder first includes a check to see whether a local maximum can be found within the given EIC. After finding the highest local maximum, the Peak Finder algorithm determines peaks by following the local maximum in positive and negative retention time direction. By analyzing the intensity ratio from one point to the next for increase or decrease beyond a certain tolerance, peak width is determined. If the total summed intensity of this determined peak is above a certain intensity threshold, the peak gets added to the alignment. This results in a more informed decision about the EIC than a simple check of the presence of a signal at the specific retention time, but it is not too strict by allowing only shapes similar to bell curves. A big advantage of this method is that the reported intensity – if there is a peak – is likely very accurate, as the method itself determines where the peak begins and ends, meaning that the peak can have a variable length. However, this method also has two major disadvantages: First, peaks have such a wide range of shapes in LC-MS that the heuristic described here does not sufficiently apply to a large part of these shapes and second, the noise present in some LC-MS EICs may still have the initial appearance of a peak. This algorithm contains the underlying assumption that the local maximum found is distinct from the signals surrounding it. In practice, the Peak Finder often results in false positives in a noisy region because the zigzagging signal locally creates the appearance of a significant peak.

Methods

Study data set

The study data set was characterized in detail in a Data Descriptor,²² but basic sample and data processing information is available in the supporting information (B). It is a collection of manually preclassified EICs designed specifically for the evaluation of gap filling algorithms. The data set contains 255,000 EICs that have been manually assigned to either be a peak or not a peak, as well as EICs that could not be clearly classified. There were no strict criteria determined for the manual classification of the EICs and it relied on the assessment of three mass spectrometry experts. The underlying raw mass spectral data was reused from a real world nontargeted screening study by Beckers et al.²³ The mass spectral data was acquired of 51 water samples collected along the Holtemme river (Saxony-Anhalt, Germany). The samples represented the whole river transect from close to its source to the confluence with the Bode river. The uniqueness of the data set is that several sampling sites along the river were sampled at time intervals corresponding with flow velocity during one day so that the samples all represent the same “water package”. In addition to diffuse forest, urban and farmland sources of natural and artificial compounds, the river receives also effluents from two wastewater treatment plants as point sources of pollution. This means that the data set as a whole should be consistent as pertaining to the compounds present in each sample, while also containing sets of samples that are clearly distinct from the others. It was expected, for example, for a large set of substances to be consistently present in the samples that were taken downstream of the first and second waste water treatment plant. The manual classification of the LC-HRMS chromatograms resulted in 62,500 out of 255,000 EICs (24.5 %) being classified as detectable peaks, while 184,850 EICs (72.5 %) were classified as not representing a peak. 6,250 EICs (3 %) could not be definitively classified. Leaving out these inconclusive EICs, this approach results roughly in a 1:3 ratio of “peaks” to “not peaks”, which means that this data set is imbalanced in regards to the classification of EICs.

Furthermore, of the 212,933 EICs that have not been picked, here called “gaps”, only 26,786 EICs (12.6 %) were manually identified as peaks, resulting in a ratio of estimatingly 1:8. This means the classes of the gaps are more imbalanced than the total data set. A detailed analysis of the complete data set (Fig. 1) showed that about 16 % of the aligned m/z and RT positions contained no peak in any of the samples, meaning that in one or more samples a peak has been falsely picked at this position in the aligned data set. At approximately 13 % of the positions, only one out of the 51 samples contained the peak. The peak count by coverage generally follows (roughly) an exponential distribution, except for peaks that occur in 20 samples, and those that occur in (almost) all samples.

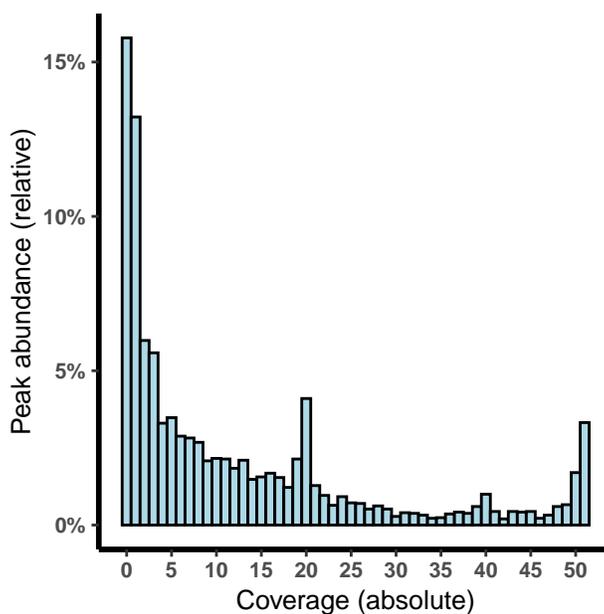


Figure 1: Relative peak abundance by coverage. The coverage is defined as the number of samples in which a peak is present. The relative peak abundance shows what percentage of peaks is present in any given number of samples in the study data set.

It is in agreement with expectations that many peaks occur in all samples, since there are substances that are known to occur in any surface water at measurable concentrations. The many peaks that occur in 20 of the samples are also explainable, since 20 samples were taken at sites that are located downstream of waste water treatment plants. Thus, these signals represent substances that are typical components of waste water treatment plants

effluents.

SAX

The Symbolic Aggregate ApproXimation (SAX) described Lin et al.¹⁷ is an algorithm generating a high level symbolic representation of time series specifically designed to reduce the dimensionality and numerosity of the data to make data mining more viable on any set of time series. SAX converts a time series into a string of letters by z-normalizing²⁴ the time series and splitting it into w segments of equal size and assigning each segment a letter according to the average of the function in that segment range. The value range is split into a separate segments so that a $N(0,1)$ normal distribution density function would be split into a ranges with an equal area under the curve. Figure 2 illustrates an example of how a peak is being converted into an SAX string.

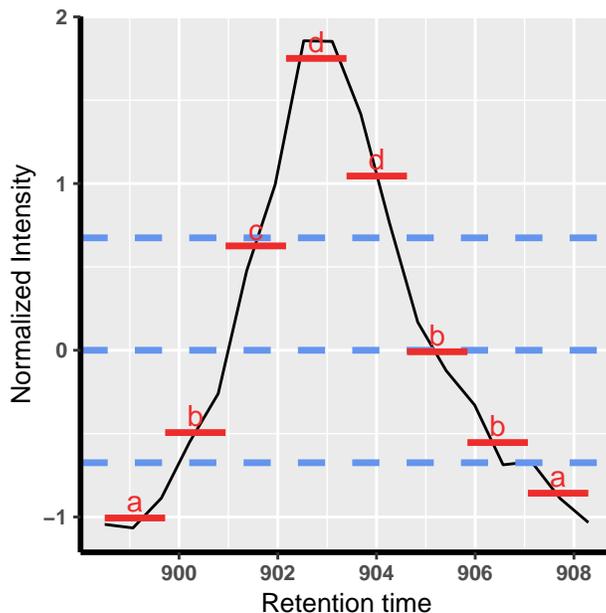


Figure 2: SAX-conversion of a time series with $w = 8$ and $a = 4$. The red bars show the position and width of each of the 8 equidistant segments with the letter above them denoting the letter that is assigned to the segment. The blue dashed lines show the positions where the value range is split so that each range would have the same area under a normal distribution. This time series is converted to the letter sequence “aabddcba”.

One main advantage of the SAX over other time series representations is that the letter

sequences, that result from the conversion of timer series data are to some degree interpretable even for humans. Another advantage is, that a lower bound distance measure called *MINDIST* is also defined, making it possible to compare two SAX strings (and henceforth time series) for their similarity without losing a lot of information compared to the original time series. For two SAX strings S_1 and S_2 , *MINDIST* is defined as:

$$MINDIST(S_1, S_2) = \sqrt{\frac{n}{w}} \sqrt{\sum_{i=1}^w (dist(S_{1_i}, S_{2_i}))}$$

With n being the number of initial data points before conversion, S_{k_i} being the letter of SAX string k at position i and $dist()$ being a lower-bound distance function for the SAX letters based on the break points of the value range. $dist()$ is usually implemented by utilizing a precalculated lookup table of the distance values between two letters.¹⁶

Implementation

The SAX-based and simple gap filling algorithms were implemented in R version 3.6.1.²⁵ The packages `openxlsx`,²⁶ `XLConnect`²⁷ and `readxl`²⁸ were used to interface with and process Excel files. The package `data.table`²⁹ was used for table processing. The package `XCMS`^{3,5,6} was used to read in `.mzML` files and to extract the EICs. The `caret`³⁰ package contained the functions for the generation of the confusion matrices for the results. The `progress`³¹ package was used to implement progress bars for several of the longer steps of the processing.

Results

SAX-based gap filling

The SAX-based gap filling algorithm was separated to work in four sequential steps:

1. Extracting and converting all EICs represented in the alignment table to the SAX-format.

2. Building a consensus string out of the SAX-strings resulting from the EICs of the originally picked peaks.
3. Comparing the SAX-strings of the EICs that have not been previously marked as definitive peaks to the consensus string and adding them if they show a high similarity.
4. Removal of low-coverage peaks if their shape does not occur in other EICs.

Extraction and conversion of EICs

The algorithm extracts raw EICs of all m/z and retention time positions specified by the aligned peak table across all S samples. The m/z range of the EICs needs to be set to the LC-MS instrument specific accuracy. To mitigate small retention time shifts, the EICs are centered around the weighted mean of the measurements of $\pm t$ seconds of the average retention time, with t being the average peak width of the chromatographic method. The EICs are then converted to the SAX-format, resulting in one string of letters for each EIC. Optimal parametrization a, w for the SAX and t for the EIC lengths are later evaluated in this study.

Building the consensus string

The consensus string of the SAX strings is built for each entry in the aligned peak table. The aligned peak table contains N rows (and therefore N aligned peak groups) with general peak information and S columns indicating the peak areas of the same peak (if present) in each sample. First, the SAX strings of each aligned peak are separated into two categories: Those strings in row x in the alignment table that represent EICs that have been picked by peak detection will be called

$$SAX_{P,x} = \{SAX_{P,x,1}, \dots, SAX_{P,x,N_{P,x}}\}$$

(with $N_{P,x}$ being the number of peaks that have been picked in the alignment table at

position x) and those that have not been picked will be called

$$SAX_{M,x} = \{SAX_{M,x,1}, \dots, SAX_{M,x,N_{M,x}}\}$$

(with $N_{M,x} = S - N_{P,x}$). The totality of $SAX_{P,x}$ and $SAX_{M,x}$ over all alignment positions will be called SAX_P and SAX_M respectively. The consensus string SAX_C, x of the peak x at row x is built out of the strings $SAX_{P,x}$ akin to a consensus sequence in genetics:³² The most frequently occurring letter at each position over all strings is taken as the canonical letter at that position. If there is no single most occurring letter, then this conflict is solved by calculating the mean of the most occurring letters by reinterpreting these letters as numbers in alphabetical order (i.e. "a" = 1, "b" = 2...) taking the average and then rounding the result if necessary.

```

abcdcbbc
abcdcbba
bbddcbbb
bbcdbcba
abbcdbbb
abddcbbc
abcdcbbb

```

Figure 3: Example of a SAX consensus string generated from 6 SAX strings with a conflict in the last position marked in red (a, b and c occur equally as often)

For example, as illustrated by Fig. 3, if the letter "a", "b" and "c" occur most often and equally as often in the same position SAX strings, the result will be $\frac{1+2+3}{3} = 2$, which means that the canonical letter in this example will be a "b".

Comparison of SAX-strings to the consensus string

To decide on whether the previously unpicked EICs get added to the alignment as a peak, the $SAX_{M,x}$ -strings are compared to the consensus string $SAX_{C,x}$ one by one. The maximum *allowed* distance of these strings to the consensus string is equal to the maximum distance of the picked strings $\max_k(MINDIST(SAX_{P,x,k}, SAX_{C,x}))$ to the consensus string. This means that not-picked EIC number m represented by the SAX string $SAX_{M,x,m}$ gets added to the alignment table only if

$$MINDIST(SAX_{M,x,m}, SAX_{C,x}) \leq \max_k(MINDIST(SAX_{P,x,k}, SAX_{C,x}))$$

Handling of low-coverage peaks

To handle low-coverage peaks (i.e. peaks that occur only below a certain percentage or in less than a set number of samples), all SAX strings contained in SAX_P that are not a representation of low-coverage picked peaks are aggregated and the occurrence of each string is subsequently counted. Afterwards the unique SAX strings are sorted by occurrence. A top percentile $p : 0 < p < 1$ of the unique SAX strings is then taken as a list of reference strings for the re-evaluation of low-coverage EICs. As an explanatory calculation, if $p = 0.3$ for a sample set with $n = 100$ peaks, the SAX strings are sorted cumulatively in decreasing order until a limit of 30 or higher.

All unique strings among those that are summed up are seen as reference strings for the evaluation of low-coverage peaks. If the SAX-strings of low-coverage peaks have a distance of 0 to any of the reference strings, the corresponding peaks remain in the data set and are used for gap filling. Otherwise, the table row in the alignment stays the same. In our study, we define low-coverage peaks as peaks that occur in 5 % or less of the samples.

Parameter optimization and evaluation

The evaluation and comparison between the described methods of filling gaps in the alignment have been done by using a number of different quality measures. Since the classes “peak detected” and “no peak detected” are imbalanced for this data set, it is important to not just evaluate the results by considering the predictive accuracy³³ (i.e. the number of correctly classified EICs divided by the total number of peaks) but rather a measure that is robust in regards to imbalanced data sets. Our preferred choice in this case is Matthews Correlation Coefficient (MCC),³⁴ since it is a well-tested measure for such types of data sets.³⁵ It is described as

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}}$$

with TP being the number of true positives, TN being the number of true negatives, FP being the number of false positives and FN being the number of false negatives. The MCC is a holistic and unbiased measure, in the sense that all values from the confusion matrix (TP, TN, FP, FN) are included at equal importance. While it is reductive to describe a classification in a single measure, the MCC is still favorable for its unbiased (in regards to type I and II errors) evaluation of any given classifier. Unlike most other quality measures, the MCC ranges from -1 to 1, since it is a correlation measure. While the predictive accuracy gives insight into the overall performance of the algorithm, sensitivity and specificity³³ indicate whether the algorithm generally produces more type I or type II errors. While both types of errors are equally valued in our case, a look at these secondary measures can gauge whether the algorithms in question produce balanced results. The MCC is used as a general measure for the optimization of the parametrization. The parameters of the SAX (a, w), the EIC extraction time (t) and the percentile for re-evaluation of low-coverage peaks (p) have not previously been optimized for the classification of EICs. To optimize these specific parameters, we used a grid search over different parameter magnitudes. There are several

reasons for using a grid search: First, a grid search over magnitudes very distinctly avoids overfitting to our specific data set. Second, the methods' success is not strongly depending on the exact parametrization of a and w , as long as the parameters are within a reasonable margin for the specific use case.¹⁷ Third: The discreteness and unpredictable interactions of these parameters mean that gradient descent methods for parameter optimization are very unlikely to work.

The specific values of which all combinations were tested are:

$$a \in \{4, 6, 8\}, w \in \{3, 5, 8, 11\}, t \in \{3.5, 7, 10.5, 14, 17.5, 21\}, p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$$

For the evaluation, the data set of 5,000 aligned peaks has been split into two parts: The cross-validation set, containing 3,000 aligned peaks, and the holdout set with 2,000 aligned peaks. For parameter optimization, a 10-fold cross-validation was used, meaning that the cross-validation set was split into 10 parts with 300 peaks and validated each time with one part left out respectively as the test data set. The optimization criterion for the algorithm was maximizing MCC . After training, the optimal parameters were applied to each test data set and the optimal parameters and evaluation metrics were calculated. The optimized parameter sets were then used to process the holdout data set as an independent validator of these parameters. The parameter t was also optimized for the simple gap filling, as the EIC width has a large impact on whether this gap filling process detects a peak or not. The optimization was done by simply testing all predefined values for t and optimizing for the maximization of MCC . These optimization processes were done twice: Once with only the gap data, meaning that the picked peaks were left out entirely from the validation and once with the full data set. Inconclusive EICs were excluded from both data sets, since they can not be used as reference values for the classification. Seven total combinations of methods and data sets were evaluated: One for no gap filling, three for the gap filling methods with evaluation parameters calculated from using the full data set and three for the gap filling

methods with the evaluation parameters calculated from using the data set containing only the original peak gaps. The holdout data set exists as a final means of evaluation and is the closest result to a worst-case evaluation of the algorithm on this MS-method. By leaving out this data set right from the beginning and only evaluating it once at the end without iterating on the results acquired from this final evaluation, we gain an unbiased estimation of the performance of our algorithm with the parameters we chose. While a repeated cross-validation usually results in an estimation closer to an average real-world case, the usage of an extra holdout data set can show possible biases that may be present in the training data set for the cross-validation.

Cross-validation results

Table 1: Average results for four evaluation measures of the cross-validation for seven different methods and/or sets of peaks: (1) Utilizing the full data set, but without any gap filling; (2) The results of the simple gap filling with the full data set; (3) The results of the Peak Finder algorithm with the whole data set; (4) The results of the SAX gap filling with the whole data set; (5) The results of the simple data set with evaluating only the gaps specifically; (6) The results of the Peak Finder algorithm with evaluating only the gaps specifically; (7) The results of the SAX gap filling with evaluating only the gaps specifically; The highest values in each table are bold-faced (multiple bold-faced numbers occur if the difference is only marginal)

Method	Accuracy	MCC	Sensitivity	Specificity
(1) Full data set (no gaps filled)	0.868	0.572	0.51	0.967
(2) Full data set (filled – Simple)	0.443	0.279	0.994	0.29
(3) Full data set (filled – Peak Finder)	0.598	0.318	0.857	0.527
(4) Full data set (filled – SAX)	0.871	0.654	0.81	0.888
(5) Only gaps (filled – Simple)	0.385	0.214	0.988	0.3
(6) Only gaps (filled – Peak Finder)	0.565	0.166	0.707	0.545
(7) Only gaps (filled – SAX)	0.871	0.51	0.698	0.895

The cross-validation results (Table 1) show that the SAX gap filling has the highest average *MCC* across all test data sets with approximately 0.654 with an average accuracy of 0.871. The “Peak Finder” gap filling results in the second highest average *MCC* of 0.318 and an accuracy of 0.598. The simple gap filling’s average accuracy is 0.4425 with an average

MCC of 0.279. The evaluation for only the gap data shows similar results with *MCC* and accuracy being equal or lower than in their respective full data sets.

A more detailed analysis of the *MCC* of the cross-validation for each run (Fig. A2, Tab. A1) shows that the results for the *MCC* were in the same general range across all test folds, meaning that each method yields consistent results, independent of the specific peak set.

Optimal parametrization

The detailed cross-validation results of the SAX gap filling (Table A1) show that there were only four parameter sets that were deemed optimal by the training process, which proves that the SAX gap filling yields consistently good results if given the right parameters. The metrics for the four parameter sets show that there is only a small difference between them in terms of classification strength. It is however notable that the parameter combination

$$a = 8, w = 8, t = 14, p = 0.9$$

is occurring most often across all evaluation folds of the SAX-based gap filling and is therefore the preferable combination to use for this type of data. It shows desirable results when evaluating the full data set as well as only the gap data. The other parameterization results are not very different in terms of their parameters, as there is a trade-off between a and w . Generally, if a is higher for an optimal parameterization, then w is lower and vice versa. It should also be noted that optimal EIC extraction retention time windows seem to be either ± 7 or ± 14 seconds, with the higher time being correlated with a higher meaning that the general segment length of the SAX (in terms of seconds per segment) stays approximately constant. The peak re-evaluation threshold p is in most cases 0.9, showing that a conservative approach to evaluation of low-coverage peaks is better for overall performance. The previously mentioned parameter set $a = 8, w = 8, t = 14, p = 0.9$ was used for the holdout data set evaluation.

Holdout data set evaluation

Table 2: Table showing the average results for four evaluation measures of the holdout data set evaluation for seven different methods and/or sets of peaks: (1) Utilizing the full data set but without any gap filling; (2) The results of the simple gap filling with the full data set; (3) The results of the Peak Finder algorithm with the whole data set; (4) The results of the SAX gap filling with the whole data set; (5) The results of the simple data set with evaluating only the gaps specifically; (6) The results of the Peak Finder algorithm with evaluating only the gaps specifically; (7) The results of the SAX gap filling with evaluating only the gaps specifically; The highest values in each table is bold-faced (multiple bold-faced numbers occur if the difference is only marginal)

Method	Accuracy	MCC	Sensitivity	Specificity
(1) Full data set (no gaps filled)	0.868	0.572	0.51	0.967
(2) Full data set (filled – Simple)	0.443	0.279	0.994	0.29
(3) Full data set (filled – Peak Finder)	0.598	0.318	0.857	0.527
(4) Full data set (filled – SAX)	0.871	0.654	0.81	0.888
(5) Only gaps (filled – Simple)	0.385	0.214	0.988	0.3
(6) Only gaps (filled – Peak Finder)	0.565	0.166	0.707	0.545
(7) Only gaps (filled – SAX)	0.871	0.51	0.698	0.895

The holdout data set evaluation (Tab. 2) shows similar results as compared to the cross-validation. The *MCC* of the SAX method is again higher than for any other method and the accuracy is comparable to that of the data set where no gaps are filled. The power of the SAX gap filling is slightly less than in the cross-validation, but this is normal behavior for the evaluation of a holdout data set. The results of all other methods are very similar to the results from the cross-validation, which is also expected due to the fact that these methods are largely nonparametric and should therefore stay consistent independent of the data set.

Applicability

While this algorithm has only been tested on environmental data, it is very likely also applicable on LC-MS data in other fields, like metabolomics, proteomics and lipidomics. The success of this algorithm is mainly dependant on substances having consistent peak shapes and a good initial peak detection. As long as these requirements are fulfilled, the

algorithm should work with similar results on other LC-MS data. Whether the algorithm works on GC-MS data needs to be tested, as there are several factors that favor or disfavor the success of this algorithm on such data. On the one hand, GC-MS is usually far less noisy and has therefore very consistent peak shapes compared to LC-MS data, on the other hand, lower resolution might be problematic for identifying the distinct shapes.

In its current form, the algorithm only needs a peak alignment table containing intensity values for each peak in each sample and mass spectrometric data in .mzML format corresponding to each sample column in the table.

Conclusion

The results show that the SAX gap filling enhances the general informativeness of the EIC classification process while sacrificing only a small margin of specificity, compared to not applying any gap filling method. Although the general accuracy of the SAX gap filling is equal to that of no gap filling, the much higher sensitivity shows that more peaks can be correctly detected in mass spectrometry with little caveats. With no gap filling method applied, only about half of all actual peaks were correctly detected. Choosing the approach without gap filling yields a very high specificity, meaning that the number of false negatives is very low compared to other methods. By using no gap filling, only about 4 % of the EICs are wrongly classified as containing a peak when they do not contain it, but the high number of undetected peaks is a strong caveat for leaving out gap filling. The simple gap filling has the highest sensitivity, but at the cost of a very low specificity. This is due to the fact that it detects a peak in every EIC where a signal is present, and therefore the specificity value for this method strongly correlated to the number of extracted EICs without any signal. While the Peak Finder algorithm managed to detect about 85 out of 100 peaks, it also led to the misclassification of about half of all not-peak representing EICs as peaks in the test data. The SAX gap filling achieves almost the same sensitivity without such a strong trade-off.

With our method, more than 80 % of all peaks can be detected while also misclassifying only around 15 % of EICs which do not contain a peak as containing it.

Acknowledgement

This study was supported by the SOLUTIONS project funded by the European Union Seventh Framework Programme (FP7-ENV-2013-two-stage Collaborative project) under grant agreement number 603437. We are grateful to Liza-Marie Beckers for the provision of the raw mass spectral data. The QExactive Plus LC-HRMS used is part of the major infrastructure initiative CITEPro (Chemicals in the Terrestrial Environment Profiler) funded by the Helmholtz Association.

Safety

Not applicable for this study.

Notes

The authors declare no competing financial interest.

Source code implementing this algorithm is available at https://github.com/ermueller/SAX_Gap_Filling

ORCID

Erik Müller: <https://orcid.org/0000-0003-3288-0439>

Carolin Elisabeth Huber: <https://orcid.org/0000-0002-9355-8948>

Werner Brack: <https://orcid.org/0000-0001-9269-6524>

Martin Krauss: <https://orcid.org/0000-0002-0362-4244>

Tobias Schulze: <https://orcid.org/0000-0002-9744-8914>

Author Contributions

Conceptualization: E.M., M.K. and T.S. Methodology: E.M., C.E.H., M.K. and T.S. Software: E.M. Validation: E.M., C.E.H. and T.S. Writing-original draft preparation: E.M. and T.S. Writing-review and editing: C.E.H, M.K., W.B. and T.S. Visualization: E.M. Supervision: T.S., M.K. and W.B. Funding acquisition: W.B.

Supporting Information Available

The following files are available free of charge.

- "1_SAX_Gap_Filling_SI.odt": A word document showing a table and a boxplot containing more in-depth data of the cross validation. It also contains additional information about the sampling and data processing used for the test data set.
- The full evaluation data set²² and related files are available for free download on Zenodo³⁶
- All raw mass spectral data is accessible for free download on MetaboLights³⁷ (under curation)

References

- (1) Hollender, J.; Schymanski, E. L.; Singer, H. P.; Ferguson, P. L. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? *Environmental Science & Technology* **2017**, *51*, 11505–11512.
- (2) Brack, W.; Hollender, J.; de Alda, M. L.; Müller, C.; Schulze, T.; Schymanski, E.; Slobodnik, J.; Krauss, M. High-Resolution Mass Spectrometry to Complement Monitoring and Track Emerging Chemicals and Pollution Trends in European Water Resources. *Environmental Sciences Europe* **2019**, *31*, 62.

- (3) Tautenhahn, R.; Böttcher, C.; Neumann, S. Highly Sensitive Feature Detection for High Resolution LC/MS. *BMC Bioinformatics* **2008**, *9*, 504.
- (4) Lange, E.; Tautenhahn, R.; Neumann, S.; Gröpl, C. Critical Assessment of Alignment Procedures for LC-MS Proteomics and Metabolomics Measurements. *BMC Bioinformatics* **2008**, *9*, 375.
- (5) Smith, C. A.; Want, E. J.; O’Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry* **2006**, *78*, 779–787.
- (6) Benton, H. P.; Want, E. J.; Ebbels, T. M. D. Correction of mass calibration gaps in liquid chromatography-mass spectrometry metabolomics data. *Bioinformatics* **2010**, *26*, 2488.
- (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) Röst, H. L. et al. OpenMS: A Flexible Open-Source Software Platform for Mass Spectrometry Data Analysis. *Nature Methods* **2016**, *13*, 741–748.
- (9) Loos, M. J. Mining of High-Resolution Mass Spectrometry Data to Monitor Organic Pollutant Dynamics in Aquatic Systems. Doctoral Thesis, ETH Zurich, 2015.
- (10) Di Guida, R.; Engel, J.; Allwood, J. W.; Weber, R. J. M.; Jones, M. R.; Sommer, U.; Viant, M. R.; Dunn, W. B. Non-Targeted UHPLC-MS Metabolomic Data Processing Methods: A Comparative Investigation of Normalisation, Missing Value Imputation, Transformation and Scaling. *Metabolomics* **2016**, *12*.
- (11) Chong, J.; Wishart, D. S.; Xia, J. Using MetaboAnalyst 4.0 for Comprehensive and

- Integrative Metabolomics Data Analysis. *Current Protocols in Bioinformatics* **2019**, *68*, e86.
- (12) Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R. B. Missing Value Estimation Methods for DNA Microarrays. *Bioinformatics* **2001**, *17*, 520–525.
- (13) Gromski, P. S.; Xu, Y.; Kotze, H. L.; Correa, E.; Ellis, D. I.; Armitage, E. G.; Turner, M. L.; Goodacre, R. Influence of Missing Values Substitutes on Multivariate Analysis of Metabolomics Data. *Metabolites* **2014**, *4*, 433–452.
- (14) Armitage, E. G.; Godzien, J.; Alonso-Herranz, V.; López-González, Á.; Barbas, C. Missing Value Imputation Strategies for Metabolomics Data. *Electrophoresis* **2015**, *36*, 3050–3060.
- (15) Stekhoven, D. J.; Bühlmann, P. MissForest—Non-Parametric Missing Value Imputation for Mixed-Type Data. *Bioinformatics* **2012**, *28*, 112–118.
- (16) Lin, J.; Keogh, E.; Lonardi, S.; Chiu, B. A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. New York, NY, USA, 2003; pp 2–11.
- (17) Lin, J.; Keogh, E.; Wei, L.; Lonardi, S. Experiencing SAX: A Novel Symbolic Representation of Time Series. *Data Mining and Knowledge Discovery* **2007**, *15*, 107–144.
- (18) Takahashi, H.; Morimoto, T.; Ogasawara, N.; Kanaya, S. AMDORAP: Non-Targeted Metabolic Profiling Based on High-Resolution LC-MS. *BMC Bioinformatics* **2011**, *12*, 259.
- (19) Shen, C.; Sun, Z.; Chen, D.; Su, X.; Jiang, J.; Li, G.; Lin, B.; Yan, J. Developing

- Urinary Metabolomic Signatures as Early Bladder Cancer Diagnostic Markers. *OMICS: A Journal of Integrative Biology* **2015**, *19*, 1–11.
- (20) Lemonakis, N.; Poudyal, H.; Halabalaki, M.; Brown, L.; Tsarbopoulos, A.; Skaltsounis, A.-L.; Gikas, E. The LC–MS-Based Metabolomics of Hydroxytyrosol Administration in Rats Reveals Amelioration of the Metabolic Syndrome. *Journal of Chromatography B* **2017**, *1041-1042*, 45–59.
- (21) Olivon, F.; Grelier, G.; Roussi, F.; Litaudon, M.; Touboul, D. MZmine 2 Data-Preprocessing To Enhance Molecular Networking Reliability. *Analytical Chemistry* **2017**, *89*, 7836–7840.
- (22) Müller, E.; Huber, C.; Beckers, L.-M.; Brack, W.; Krauss, M.; Schulze, T. A Data Set of 255,000 Randomly Selected and Manually Classified Extracted Ion Chromatograms for Evaluation of Peak Detection Methods. *Metabolites* **2020**, *10*.
- (23) Beckers, L.-M.; Brack, W.; Dann, J. P.; Krauss, M.; Müller, E.; Schulze, T. Unraveling longitudinal pollution patterns of organic micropollutants in a river by non-target screening and cluster analysis. *Science of The Total Environment* **2020**, *727*, 138388.
- (24) Goldin, D.; Kanellakis, P. C. On Similarity Queries for Time-Series Data: Constraint Specification and Implementation. 1995; pp 137–153.
- (25) R Core Team, *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019.
- (26) Walker, A. *Openxlsx: Read, Write and Edit XLSX Files*; 2019.
- (27) Mirai Solutions GmbH, *XLConnect: Excel Connector for R*; 2018.
- (28) Wickham, H.; Bryan, J. *Readxl: Read Excel Files*; 2019.
- (29) Dowle, M.; Srinivasan, A. *Data.Table: Extension of ‘data.Frame’*; 2019.

- (30) Kuhn, M. Building Predictive Models in R Using the Caret Package. *Journal of Statistical Software, Articles* **2008**, *28*, 1–26.
- (31) Csárdi, G.; FitzJohn, R. *Progress: Terminal Progress Bars*; 2019.
- (32) Pierce, B. A. *Genetics: A Conceptual Approach*; Macmillan, 2008; Vol. 1.
- (33) Tharwat, A. Classification Assessment Methods. *Applied Computing and Informatics* **2018**,
- (34) Matthews, B. W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* **1975**, *405*, 442–451.
- (35) Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal Classifier for Imbalanced Data Using Matthews Correlation Coefficient Metric. *PLOS ONE* **2017**, *12*, e0177678.
- (36) Müller, E.; Huber, C.-E.; Beckers, L.-M.; Brack, W.; Krauss, M.; Schulze, T. A Data Set of 255,000 Randomly Selected and Manually Classified Extracted Ion Chromatograms for Evaluation of Peak Detection Methods. 2020; <https://doi.org/10.5281/zenodo.3756211>.
- (37) Müller, E.; Huber, C.-E.; Beckers, L.-M.; Brack, W.; Krauss, M.; Schulze, T. A Data Set of 255,000 Randomly Selected and Manually Classified Extracted Ion Chromatograms for Evaluation of Peak Detection Methods. 2020; <https://www.ebi.ac.uk/metabolights/MTBLS1455>.

Graphical TOC Entry

