Generating realistic data through modeling and parametric probability for the numerical evaluation of data processing algorithms in two-dimensional chromatography

Milani, N.B.L.; García-Cicourel, A.R.; Blomberg, J.; Edam, R.; Samanipour, S.; Bos, T.S.; Pirok, B.W.J.

DOI
10.1016/j.aca.2024.342724

Publication date
2024

Document Version
Final published version

Published in
Analytica Chimica Acta

License
CC BY-NC

Link to publication

Citation for published version (APA):
Milani, N. B. L., García-Cicourel, A. R., Blomberg, J., Edam, R., Samanipour, S., Bos, T. S., & Pirok, B. W. J. (2024). Generating realistic data through modeling and parametric probability for the numerical evaluation of data processing algorithms in two-dimensional chromatography. *Analytica Chimica Acta*, *1312*, Article 342724. https://doi.org/10.1016/j.aca.2024.342724

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Generating realistic data through modeling and parametric probability for the numerical evaluation of data processing algorithms in two-dimensional chromatography

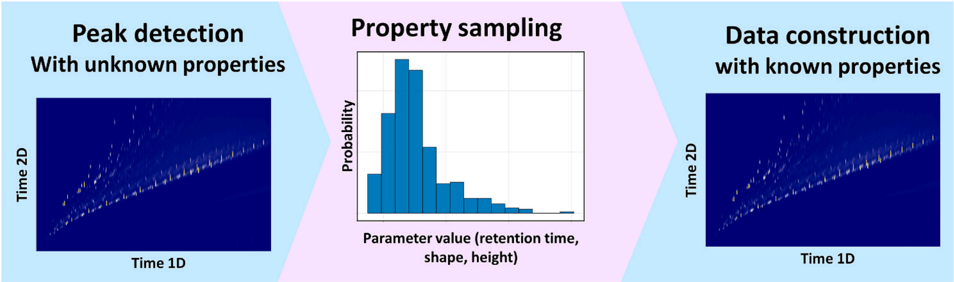
Nino B.L. Milani [a,b,*], Alan Rodrigo García-Cicourel [c], Jan Blomberg [c], Rob Edam [c], Saer Samanipour [a,b], Tijmen S. Bos [a,b], Bob W.J. Pirok [a,b,**]

[a] Van't Hoff Institute for Molecular Science (HIMS), University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, the Netherlands
[b] Centre for Analytical Sciences Amsterdam (CASA), the Netherlands
[c] Shell Global Solutions International B.V., Grasweg 31, 1031 HW, Amsterdam, the Netherlands

HIGHLIGHTS

- Benchmark data is needed for objective evaluation of data-processing algorithms.
- A Skewed Lorenz-Normal distribution is applied to describe chromatographic peaks.
- A tool was developed to generate highly realistic chromatographic data.
- The simulation of realistic data is demonstrated on LC × LC and GC × GC signals.
- This tool may facilitate further the development of data analysis workflows.

GRAPHICAL ABSTRACT



ARTICLE INFO

Handling Editor: Prof. L. Buydens

Keywords:
Two-dimensional chromatography
Peak detection
Peak modeling
Data synthesis

ABSTRACT

Background: Comprehensive two-dimensional chromatography generates complex data sets, and numerous baseline correction and noise removal algorithms have been proposed in the past decade to address this challenge. However, evaluating their performance objectively is currently not possible due to a lack of objective data.
Result: To tackle this issue, we introduce a versatile platform that models and reconstructs single-trace two-dimensional chromatography data, preserving peak parameters. This approach balances real experimental data with synthetic data for precise comparisons. We achieve this by employing a Skewed Lorentz-Normal model to represent each peak and creating probability distributions for relevant parameter sampling. The model's performance has been showcased through its application to two-dimensional gas chromatography data where it has created a data set with 458 peaks with an RMSE of 0.0048 or lower and minimal residuals compared to the original data. Additionally, the same process has been shown in liquid chromatography data.
Significance: Data analysis is an integral component of any analytical method. The development of new data processing strategies is of paramount importance to tackle the complex signals generated by state-of-the-art separation technology. Through the use of probability distributions, quantitative assessment of algorithm

* Corresponding author. Van't Hoff Institute for Molecular Science (HIMS), University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, the Netherlands.
** Corresponding author. Van't Hoff Institute for Molecular Science (HIMS), University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, the Netherlands.
E-mail addresses: n.b.l.milani@uva.nl (N.B.L. Milani), B.W.J.Pirok@uva.nl (B.W.J. Pirok).

https://doi.org/10.1016/j.aca.2024.342724
Received 13 December 2023; Received in revised form 22 April 2024; Accepted 14 May 2024
Available online 15 May 2024

performance of new algorithms is now possible. Therefore, creating new opportunities for faster, more accurate, and simpler data analysis development.

Abbreviations

RMSE	root mean square error
POI	Peak of Interest
SLN	Skewed Lorentz-Normal

1. Introduction

Comprehensive two-dimensional (2D) gas- and liquid chromatography are two powerful techniques for the analysis of complex samples [1]. The strength of comprehensive 2D chromatographic techniques arises from the fact that each first-dimension (^{1}D) fraction is subjected to a second-dimension (^{2}D) separation. If the two separation dimensions are sufficiently different (*i.e.* orthogonal), then the total peak capacity approaches that of the product of the peak capacities of the individual dimensions [2]. This dramatically improves the separation power over one-dimensional techniques and has redefined chromatography for complex mixtures. Of course, there are significant differences between both techniques. Benefiting from clean mobile phases, *i.e.* low background, and high diffusion coefficients, comprehensive 2D gas chromatography (GC × GC) is considered easier to use and more robust than comprehensive 2D liquid chromatography (LC × LC). GC × GC is more established than LC × LC and has received commercial support for a long time [3]. On the other hand, LC × LC profits from the extensive amount of selectivity combinations of liquid-phase separations, its versatility, and applicability, although modulation and optimization are consequently more complex [4–6].

Despite these significant differences, both techniques do have one thing in common. Signal processing becomes dramatically more difficult due to the large discrepancy in data points between the first and second dimension, as well as the added complexity of grouping peaks in multiple modulations to form a single two-dimensional peak. Furthermore, all difficulties associated with one-dimensional data processing (*e.g.* baseline drift) occur simultaneously along an additional plane in two-dimensional data [7]. Here lies the important challenge for chemometrics: the development of strategies that require low effort and consistently extract meaningful information from these complex datasets. In principle, pre-processing techniques devised for either GC × GC or LC × LC can be beneficial for the other, provided they are applied to similar types of data.

This challenge arises from the structure of the data, which is shown in Fig. 1. In the case of a single-channel detector such as the FID or single-wavelength UV, comprehensive 2D chromatography yields a long one-dimensional vector of data (Fig. 1A) with the individual ^{2}D modulations recorded in sequence by the detector. This vector is then folded into a 2D matrix of data (Fig. 1B). The time domains spanned by the ^{1}D and ^{2}D are widely different. Indeed, because the ^{2}D samples ^{1}D fractions, the ^{1}D analysis time can span up to hours, whereas this is often minutes or seconds for the ^{2}D modulation time. In contrast, the number of data points in the ^{1}D is equal to the number of modulations, and consequently, the ^{1}D is sparse in data (Fig. 1D). The opposite is true for the ^{2}D where sampling frequencies easily reach several hundred data points per second (Fig. 1E). Consequently, it is difficult to describe ^{1}D peak distributions and even to determine which ensemble of ^{2}D peaks represents a ^{1}D chromatographic peak (Fig. 1B, inset). While the spectrum contains information pertaining to the chemical identity of the eluting analytes at that time point, which certainly can facilitate peak detection, the dataset, and data preprocessing become much more computationally heavy.

In this context, chemometricians have developed methods along the data processing chain. Often the first step is the removal of low and high-frequency (*e.g.* spikes and baseline drift) background signals and the subsequent extraction of the – generally mid-frequency – chromatographic peaks. For example, random high-frequency noise can be removed by simply taking the average of a number of data points. The number of data points that need to be averaged, *i.e.* the window, needs to be sufficiently large to remove the noise, yet not so large that it will deteriorate the signal of interest. Additionally, a weight factor can be

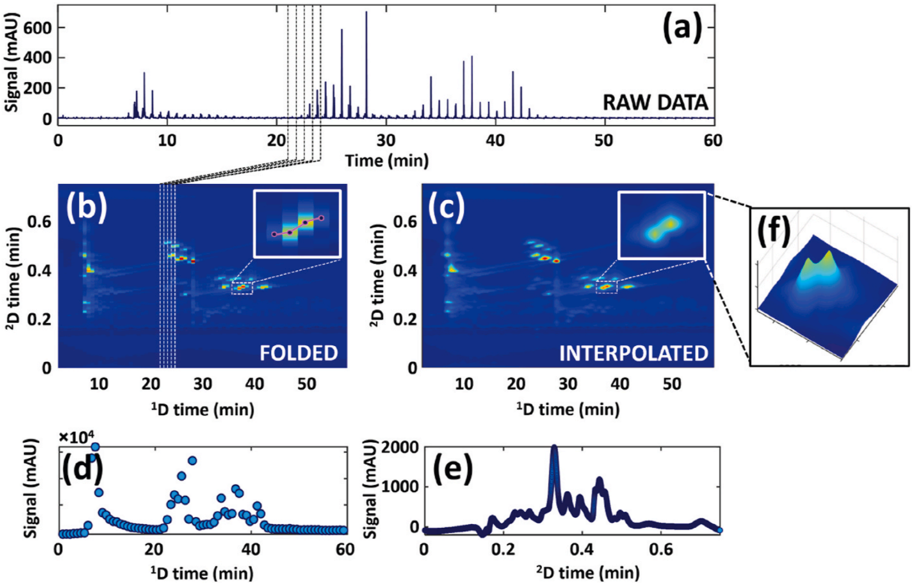


Fig. 1. (A) Illustration of a raw LC × LC chromatogram. Dashed lines delineate 2D modulations. (B) Folded 2D representation of the raw data depicted in panel (A). (C) Interpolated rendition of the data presented in panel (B). (D) 1D chromatogram obtained by summing all 2D datapoints. (E) 2D chromatogram obtained by summing all 1D datapoints. (F) Minor shifts in retention time may lead to the detection of two peaks. Adapted from Ref. [8] with permission.

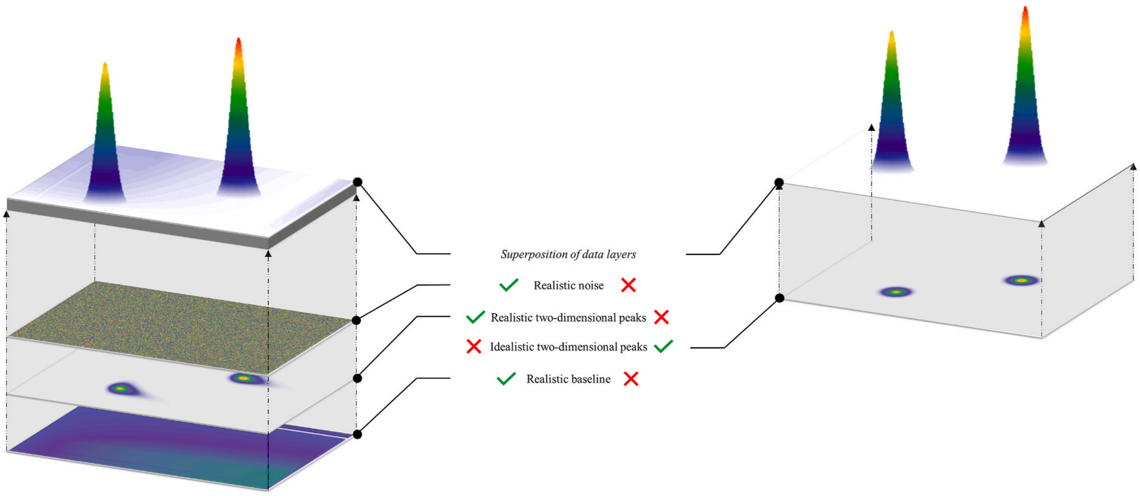

Superposition of data layers
✓ Realistic noise ✗
✓ Realistic two-dimensional peaks ✗
✗ Idealistic two-dimensional peaks ✓
✓ Realistic baseline ✗

Fig. 2. Left, realistic simulated data is presented, featuring elements such as tailing, shifting, baseline fluctuations, and noise. Right, simulated ideal data is depicted, characterized by pristine Gaussian peaks without any noise or baseline artifacts.

introduced to alleviate this phenomenon by placing more emphasis on the center of the window than the edges. The best-known example of such an algorithm is the Savitzky-Golay smoother [9]. Baseline removal is less trivial because of the low frequency nature, and often high run-to-run variation of this signal component. Over time, a large number of baseline removal algorithms have been proposed [10]. For example, the asymmetric least squares algorithm makes clever use of the, generally true assumption, that chromatographic peaks are only going in the positive direction of the y-axis. Therefore, when fitting the baseline using a least squares approach, much more emphasis is placed on downward trends as there are no peaks in this direction [11]. In addition to various least square-based baseline correction methods, algorithms based on iterative polynomial fitting [12,13] Fourier filtering [14–16], and even neural network-based methods [17,18] have been brought forward.

In addition to noise and baselines, a unique challenge encountered in 2D chromatographic data analysis is the misalignment between modulations. Various algorithms have been developed to address this issue, including correlation-optimized warping (COW) [19], dynamic time warping (DTW), parametric time warping (PTW) [20], correlation-optimized shifting (COSHIFT) [21,22], and Parallel Factor Analysis (PARAFAC2) [23]. These methods aim to correct misalignment, thereby enhancing peak integration accuracy and facilitating data interpretation.

The next step is peak detection and clustering in which the chromatographic elution bands are obtained from the data. An integral aspect of achieving quantitative peak detection is deconvolution. Two classical approaches involve employing signal derivatives [24] or curve-fitting methods [25]. The former is capable of detecting peaks regardless of their number but is susceptible to noise interference. Conversely, the latter is less affected by noise but often necessitates an estimation of the number of peaks within a given region. A more contemporary method of deconvolution involves utilizing wavelet transformation [26]. An advantage of this wavelet-based approach is its efficacy in noise and baseline signal removal.

The two main approaches often used for 2D chromatographic data are the so-called Two-Step [27,28] and Watershed algorithms [29]. While deconvolution often improves quantification accuracy, the Watershed algorithm typically does not include this step, whereas the Two-Step algorithm employs a derivative-based approach. The performance of the two-step and watershed algorithms has frequently been compared by their authors. Truyols and Janssen observed a favorable performance of their two-step algorithm relative to the watershed algorithm for peak detection in a Diesel sample [30]. The group of Reichenbach later disputed this with their comparison of the two algorithms using synthetic data [31].

While such comparisons provide valuable insights into the performance of data processing algorithms, objective numerical evaluation is arguably limited and highly case-specific. Consequently, the robustness of these algorithms against different signal properties is not well understood which complicates further development and – more importantly - implementation. To understand the latter it is useful to note that the last decade has seen the development of as many as 13 [7,15,17,18,32–40] different background correction algorithms. It was earlier shown that the performance of these algorithms depends dramatically on the signal properties [10]. Efforts have been made to automate pre-processing in order to reduce reliance on case-specific solutions [41,42]. However, these endeavors are not consistently applied to the specific types of data under consideration. To facilitate the extension of these efforts to diverse datasets, readily available benchmark data should be accessible. As a result, for users of chromatography without the required expertise and time, it remains difficult to select the correct algorithm that matches their case-specific signal properties with objective numerical evaluation.

The lack of objective numerical evaluation is not related to strategic development choices by the authors of algorithms nor a lack of willingness. Instead, it arises from a lack of datasets suitable for objective assessment and is a symptom of a scientific two-fold dilemma. To understand this, it is first important to note that any data processing method will always introduce an error, as none is – and can be – perfect. For a peak on a noisy background signal, a background correction algorithm should ideally exclusively remove the background and leave the area of the peak intact. However, in practice, some of the true areas will typically be affected. If the ground truth is known, error analysis could be performed by comparing the ground truth with the processed signal.

The arising dilemma is shown in Fig. 2. On one hand, scientists employ real experimental datasets to maximize the usefulness of the evaluation of new algorithms. Pre-processing techniques need to be able to deal with peak overlap, noise, various types of baseline, peak asymmetry, and misalignment. The experimental data indeed contains realistic degrees of peak overlap, noise, peak asymmetry, etc., yet requires a (non-perfect) algorithm to obtain the peak characteristics.

It is thus impossible to determine the ground truth peak characteristics for such datasets and current approaches allow relative comparisons at best. The second strategy considers the use of fully simulated data. While such signals do allow absolute determination of the peak characteristics, they typically represent idealistic situations and lack signal distortions and properties pertaining to experimental data.

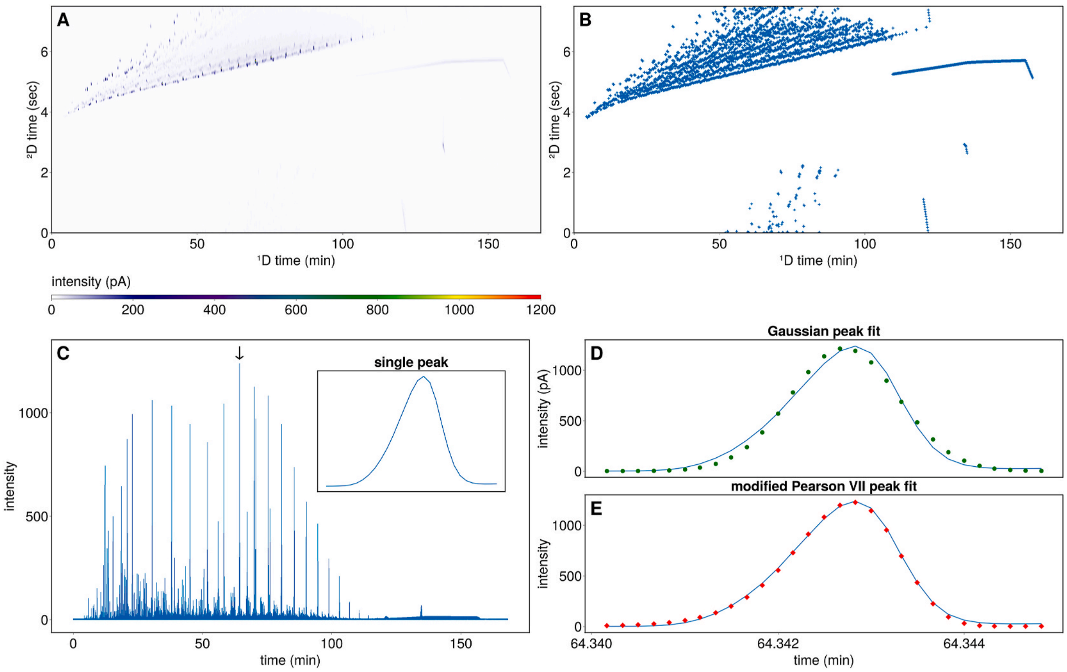


Fig. 3. A) Folded heat-map representation of a GC × GC-FID separation of a Diesel sample. B) Detected peaks on a GC × GC-FID chromatogram. Note that the markers depicted may represent multiple occurrences of the same chromatographic peak across different modulations. C) Raw one-dimensional signal of the separation shown in panel A, with the inset depicting a selected chromatographic peak within a modulation. D) Example of a Gaussian fit of the peak shown in Panel C where the blue trace is the raw data and the green dots are the fit. E) Example of a Skewed Lorentz-Normal (SLN) fit of the peak shown in Panel C where the blue trace is again the raw data and the red diamonds are the fit. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Consequently, the performance as a function of signal properties cannot be obtained. For instance, De La Mata et al. have used simulated data to assess the effect of integration parameters on quantification [43]. In this work, multiple ^{1}D Gaussians were used to create a ^{2}D peak. Parastar et al. have used exponentially modified Gaussian peaks for testing retention time alignment [44]. While these allow for more complex peak shapes than a regular Gaussian, it is still just one example of a peak with limited complexity around it (*e.g.* omission of baseline, coelution, etc.).

Weggler et al. have made significant efforts to create a compromise between the advantages of simulations and experimental data, by creating a very well-characterized open-source data set of standard compound measurements and a set of chocolate aroma profiles measured on GC × GC [45]. While these datasets hold significant value for individuals involved in the development and testing of data analysis algorithms, it’s important to note that they represent specific cases. While well-characterized real datasets serve as the gold standard benchmark, creating such meticulously characterized datasets from standards can be excessively time-consuming, expensive, or unfeasible if standards are unavailable.

Enhancing the utilization and advancement of data processing methods would greatly benefit from datasets that are not contingent upon standards. These datasets should encompass both the absolute values of simulated data and the variability in signal properties observed in experimental data.

In this work, a platform was developed that allows the modeling and generation of simulated data that contains experimental features. By supplying any two-dimensional data set (*e.g.* GC × GC, LC × LC, etc.), the algorithm will capture as much information as possible and generate a new version of this dataset with all corresponding peak data. The concept is based on distilling recorded chromatograms down to a series of probability distributions, that can be used to construct highly realistic chromatograms. This strategy allows us to achieve a compromise between the realism of experimental data and the numerical ground truth provided by simulated data.

2. Experimental

2.1. GC and LC materials and methods

This study utilized data from GC × GC and LC × LC experiments. GC × GC data was obtained from diesel samples measured using a GC × GC-FID with a modulation time of 7.5 s and a frequency of 100Hz. The methodology used to record the GC × GC dataset was identical to the one described by van Mispelaar et al. [46]LC × LC data was obtained from the study by Pirok et al. on a dye mixture [47].

2.2. Computation and algorithm

Computations were carried out on a Dell XPS 9500 laptop running an Intel Core i9 10885H with 64 GB of DDR4 dual channel memory at 2933 MHz. The algorithm was written in the open-source language Julia 1.6 [48] using the Visual Studio code intergraded development environment. In addition, the following Julia packages were used: NetCDF v0.11.7; LsqFit v0.13.0; Interpolations v0.14.7; Statistics (part of Julia standard library); DataFrames v1.4.4; CSV v0.10.9; ProgressBars v1.4.1; SpecialFunctions v2.1.7; Plots v1.38.3; StatsBase v0.33.21; JLD2 v0.4.30; Peaks v0.4.3; MAT v0.10.3; Distributions v0.25.80; Random (part of Julia standard library). Figs. 3–7 were made using the Makie.jl V0.19.2 plotting frond-end and GLMakie.jl V0.8.2 backend [49]. The functions that the algorithm comprises can be found at https://github.com/cast-amsterdam/one_simulation. A flowchart to visualize the steps of the algorithm can be found in Supplementary Information section S1.

3. Results & discussion

3.1. Probability distributions to describe realistic data

One of the key issues with simulated data is that it typically provides an incomplete description of realistic cases in terms of the number of peaks, peak shape, elution patterns, and peak overlap. At the same time, realistic experimental data is often too case-specific to objectively evaluate an algorithm. Moreover, the use of experimental data only allows for relative numerical evaluation of algorithms rather than absolute. For this reason, our unique approach employs the concept of probability distributions to represent experimental data. To achieve this objective, it was imperative to construct our algorithm in a manner that ensures the dependable interpretation of real-world data prior to the computation of these probability distributions.

To do so, a Gaussian- (eq (1)) and Skewed Lorentz-Normal (eq (2)) model were used in a curve-fitting approach. Naturally, fitting one curve to the entire chromatogram is not possible. Therefore, the first step in describing an experimental dataset was finding and determining the location of peaks on the raw signal. This was accomplished using the Peaks.jl derivative-based local maxima finder. Fig. 3A shows an example of such retention times found on a GC × GC signal. At this point, it is important to note that the raw signal of any single trace comprehensive 2D chromatogram is a long one-dimensional signal as shown in Fig. 3C. Indeed, through the folding of the signal based on the modulation time, the data is converted into a heatmap (Fig. 3B). We deliberately consult all raw signals in their 1D state, to avoid any effect of under-sampling of the ^{1}D relative to the ^{2}D, as the number of modulations per peak is insufficient to accurately model peak moments. Consequently, the markers depicted in Fig. 3A may represent multiple occurrences of the same chromatographic peak across different modulations. This will later allow us to also represent the sampling of ^{1}D as a characteristic probability distribution.

Now that a list of the raw one-dimensional signal retention times has been acquired, the algorithm can be provided with all target locations for peak modeling. For the peak modeling, we designed the algorithm to fit Gaussian (Equation (1)) or the Skewed Lorentz-Normal (Equation (2)) function of which examples are shown in 2-D and 2-E, respectively.

$$\frac{h}{\sigma\sqrt{2\pi}} \times e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \text{(Eq. 1)}$$

$$\frac{h}{\pi\gamma} \times \left(1+\frac{(x-\mu)^2}{M\times\ (\gamma+E\times(x-\mu))^2}\right)^{-M} \qquad \text{(Eq. 2)}$$

From Equations (1) and (2) it is apparent that for the Gaussian case, there are three parameters, the mean, μ, standard deviation, σ, and the height factor, h, whereas for Skewed Lorentz-Normal there are 5. The Skewed Lorentz-Normal contains a height factor h_i, scale parameter, γ, and mean, μ, but also the M and E term, allowing for more variability in shape. The shape factor M controls the overall peak shape, akin to kurtosis. This factor can alter the shape from resembling a Cauchy ($M = 1$ and $E = 0$) and a Gaussian ($M = \infty$) function. In Gaussian form γ equals $\sigma\sqrt{2}$. Whereas the E is the asymmetry, with negative values being fronting peaks, positive values being tailing peaks and 0 being perfectly symmetrical. Please note that the height factor h_i, is only partially responsible for peak height. The true height of the peak is denoted by $\frac{h}{\sigma\sqrt{2\pi}}$, and $\frac{h}{\pi\gamma}$, for the Gaussian and Skewed Lorentz-Normal respectively. Therefore, the use of the term “height” in this article refers to the “true height” rather than the mathematical height “h”.

3.1.1. *Defining the correct thresholds*

The main aspect that stands in the way of fully autonomous algorithms is usually the definition of thresholds. The advantage of our algorithm is that it largely self-adjusts these values based on the data that is being presented to it. Still, a few thresholds need to be set in order for the algorithm to function correctly.

The local maxima finder has two thresholds. A signal-to-background, and a peak inclusion threshold. The signal-to-background threshold is meant to separate the signal from the background. This threshold is set quite low on purpose so as not to miss any peaks. In the case of GC × GC data, this threshold was set based on manual tuning to 5 pA as the baseline was at 3 pA. For the LC chromatograms, the same values and reasoning were used (i.e. a 5 A U. threshold for a 3 A U. baseline). The peak inclusion threshold is stricter and pertains to selecting peaks that have a high likelihood of being fitted correctly. Where the first threshold filters peaks detected on the linear representation of the data before they are grouped. The second threshold determines if the individual modulation peaks are high enough to be part of a 2D peak. Starting at the highest apex in a list of apices that result from the first threshold filtering, the algorithm will start to group apices until an entire peak is described over all its modulations. It will continue to do so until a point is reached where no remaining apices are higher than the second threshold. Determining this threshold is non-trivial as “over-fitting” can occur and therefore skew the parameters in the probability distributions. In order to prevent over-fitting a series of tests were done with this second threshold varying from 5 pA to 45 pA. 25 pA was selected as a middle ground between peak number (*i.e.* no prevalent peaks being absent) and fit quality. For the LC × LC data, the same process was followed, and concluded that 25 A U. was the most advantageous threshold. Figures supporting this decision, as well as a more in-depth discussion, can be found in the supplementary information S2–S3 (Figs. S2–S25) Another important threshold is that of the modulation grouping. This is performed by looking for an apex in a ^{2}D time window in the modulations adjacent to the current peak of interest (POI). For GC × GC this window was defined as ± 0.05 s, for LC × LC this was defined as ± 0.25 s. This threshold was set based on visual inspection of the data, by assessing the shift in second dimension retention time of the modulations of some well-defined peaks.

Apart from these thresholds mentioned above, there are also some thresholds that do not need to be adjusted on a case-by-case basis. For instance, the iterative least squares regression required initial parameters and their upper and lower bounds to be specified, which significantly affected the quality of the fit. The proposed approach employed the values obtained from the initial local maxima finder to obtain initial fit parameters. Of course, chromatographic peaks in experimental chromatograms can be expected to overlap with neighboring peaks. The presence of these neighbors can significantly affect the fit of the POI and the required time. When a neighbor is no longer baseline-separated, it is imperative to include its presence in the fit. Therefore, the algorithm counts the number of peak apices that fall within a 6 γ distance of the apex of the POI. A resolution of 1.5 is generally regarded as well separated, since this is based on the width at the base of the peak a multiplication by 4 converts it roughly to the γ equivalent. This is a conservative number that assumes co-elution to be significant even at 3 γ. Next, a Skewed Lorentz-Normal function is fitted for each detected peak, and the parameters for the POI are stored. It should be noted that this can be simplified by always fitting the neighbors regardless of their position, although a significant Inter-peak distance would increase the change of artifacts and might impact performance.

Unfortunately, even the most advanced models are unlikely to capture every experimental nuance in a chromatographic signal. Nevertheless, because the probability distributions are obtained using fitted peak models, the quality of fitting is paramount. Therefore, peaks that could not be satisfactorily modeled were omitted from the reconstructed dataset. For example, from the vast number of peaks present in a chromatogram as shown in Fig. 3B, only a few hundred were suitable for use in the probability distributions. This process is similar to the threshold during initial peak detection described above. Specifically, if the relative difference (difference between signal and fit, divided by the signal), or the adjusted R^2 exceeds the user-defined threshold, the peak was deemed unfit for entry into the probability distributions. In this study, the relative difference was set to 1, for height, width, and position, but this can be changed on a case-by-case basis.

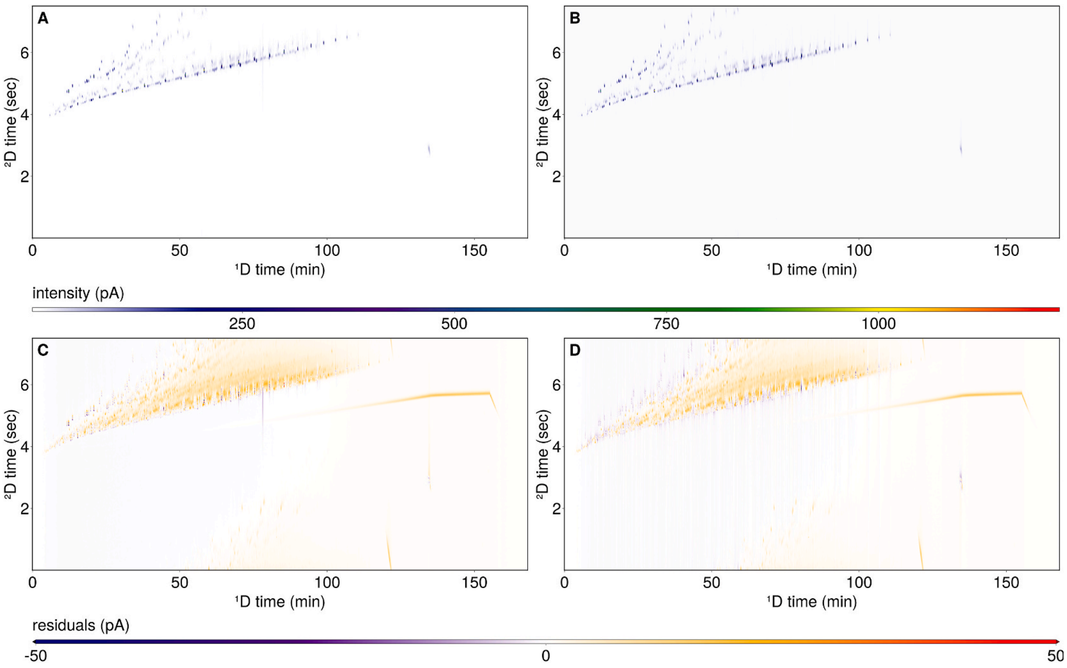


Fig. 4. Comparison of Gaussian fit (A) and Skewed Lorentz-Normal (B). The residuals for the Gaussian and Skewed Lorentz-Normal are displayed in C and D respectively, with red/orange indicating a positive deviation, blue/purple indicating a negative deviation, and white indicating residuals near 0. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

3.1.2. Model performance evaluation

Fig. 4 shows a comparison between Gaussian peak models (Fig. 4A and B) and Skewed Lorentz-Normal models (Fig. 4C and D). Fig. 4A and C are the constructed data and Fig. 4B and D are the residuals, both for the Gaussian and Skewed Lorentz-Normal peak models respectively. The residuals are obtained by subtracting the simulated data from the raw data. White areas indicate residuals that are (very close to) 0, indicating the same information in the raw and simulated data. Red/orange areas indicate positive residuals, indicating information being present in the raw data, that is not present in the simulated data. The purple/blue areas indicate regions that have a higher signal in the simulated data than the raw data. The Skewed Lorentz-Normal performs better than the Gaussian model. In addition to the residuals, the Root Mean Square Error (RMSE), for each of the peaks has been calculated and all of the 458 number fits had an RMSE below 0.0048 in the Gaussian case and 0.0045 in the Skewed Lorentz-Normal case. The full distribution of RMSEs can be found in Supplementary Fig. S4. Both peak models show a big orange cloud in the residuals, indicating small peaks that are not captured by the model, but are present in the raw data. On top of that the Gaussian model has more red peaks that are not captured well, leading to the conclusion that the Skewed Lorentz-Normal performs better. This orange cloud nicely illustrates the current shortcoming, not only of our algorithm but peak detection as a whole. This area consists of a high number of small unresolved peaks. This big wavey signal is very hard to describe accurately, even with curve fitting as the exact number of peaks is unknown.

Although the Skewed Lorentz-Normal model exhibits superior performance compared to the Gaussian model in all illustrated cases, it is worth noting that the Gaussian model possesses a distinct advantage in terms of computational speed. The Gaussian peak model demonstrates significantly faster fitting, ranging from 10 to 50 times faster than the Skewed Lorentz-Normal model. In scenarios where peak symmetry is relatively high, and computational efficiency is of paramount importance, the Gaussian peak model may be considered a more appropriate choice.

In supplementary information S5, we further discuss the application of the algorithm on a benchmark data set, displaying comparable linearity between the calibration curve of simulated data and the calibration curves resulting from commercially available software packages as well as providing a brief overview of the functions used for this in supplementary information S6.

3.1.3. Establishing probability distributions

To create the probability distributions, the fit results of all peaks could now be binned and scaled in order to obtain the histograms for visualization. Fig. 5 shows the probability distributions generated for a GC × GC-FID chromatogram when using a Skewed Lorentz-Normal model to represent the peaks. Several observations can already be made. For example, the distribution for peak height exhibits a strong tail at larger values indicating the typical presence of a relatively larger occurrence of smaller peaks in separations. Conversely, the distribution for peak width is in line with GC theory that peaks become broader at a higher second dimension time. The M and E parameter fits are more difficult to place into context due to these higher-order peak moments being caused by more obscure chromatographic phenomena.

The parameter M denotes a modification influencing the shape of chromatographic peaks. Due to the interdependency between parameters E and M, essential for peak shape, improper constraints may hinder fitting accuracy. To address this, M was restricted to a range of 1–1000, and E to -0.3 to $+0.3$, preventing implausible peak fits while favoring mathematically ideal solutions. Predominantly, peaks exhibit a narrow range of shapes, consistent with expectations from samples processed using similar techniques. Notably, an accumulation of extreme M values is observed, particularly among smaller, challenging-to-fit peaks, suggesting deviation of higher-order peak moments under difficult fitting conditions.

In supplementary information S2, various peak height thresholds are examined, revealing increased M variability with lower thresholds, indicative of challenging fits. A threshold of 25 strikes a balance between peak quantity and fit quality. E represents asymmetry, with 0 denoting symmetry, negative values indicating tailing peaks, and positive values signifying fronting peaks, aligning with expected behavior in the dataset. Similarly, an accumulation of extreme E values is observed with excessively low thresholds, mirroring the behavior of

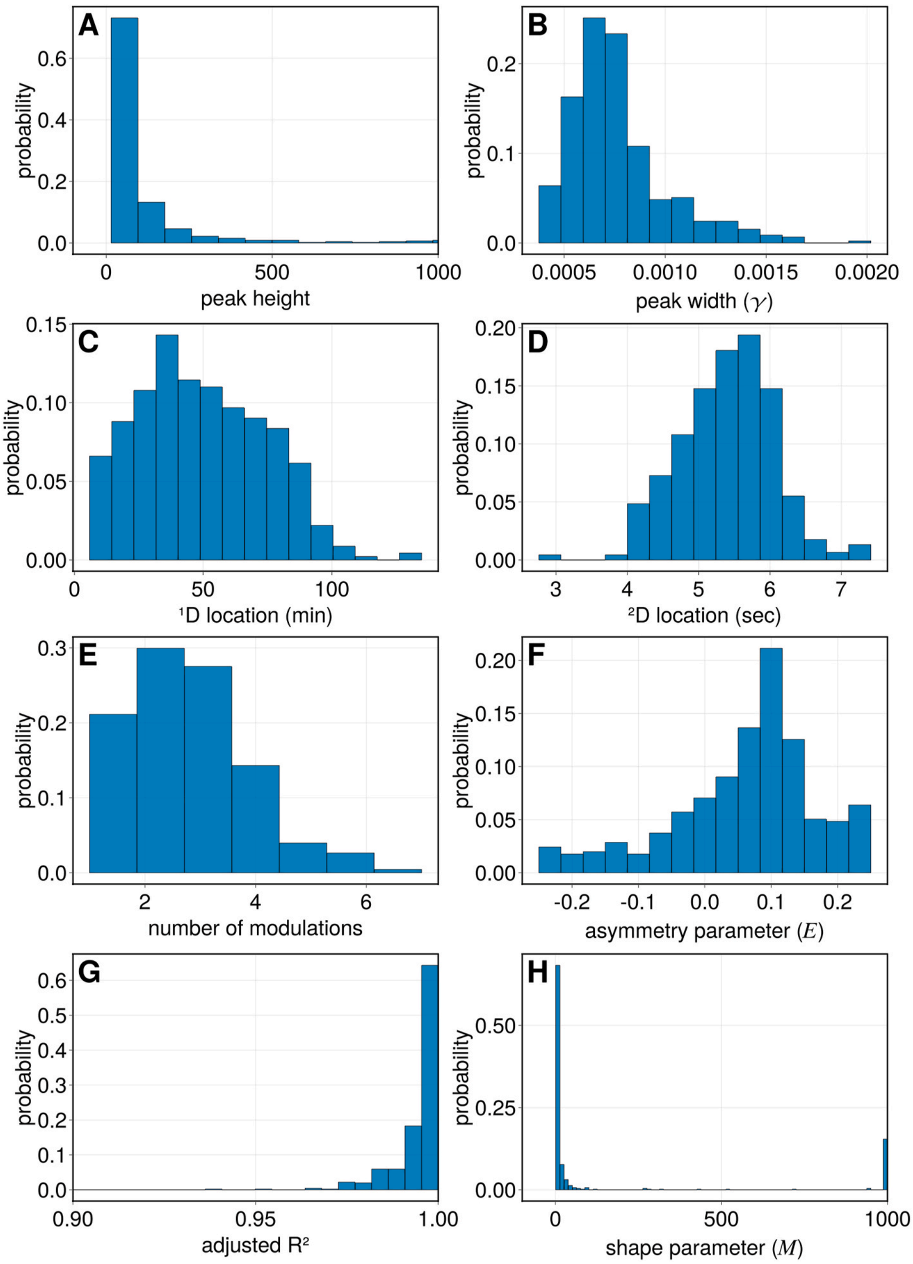


Fig. 5. The probability distributions for peak features: peak height (A); peak width (B); ^{1}D location (C), ^{2}D location (D); number of modulations (E); Asymmetry or E-factor (F); adjusted R^2 of the fit (G); and peak shape or M-parameter (H).

M.

The determination of modulations relies on a basic clustering algorithm, which identifies peak presence in successive modulations within a ^{2}D time window, considering both positive and negative user-defined margins. Additionally, it adheres to an unimodal criterion akin to Peters et al. [28]. Although more sophisticated clustering algorithms exist [50, 51], a deliberate decision was made to adopt the simplest approach at each stage initially, to avoid early bias in the simulations. Adjusted R^2, depicted in Fig. 5G, measures the correlation between fitted peaks and raw data, ranging from 0 to 1. Despite the dataset's complexity, R^2 values generally indicate the validity of inputs into the probability distribution.

Since each peak provides a single entry in each probability distribution, a dynamic version of Fig. 5 can be made. In such a plot the user can define a sub-range for one of the parameters and look at the effect this has on the other parameters. For example, when isolating the higher ^{2}D dimension times the distribution of peak width will shift towards higher values as well. This provides many opportunities to inspect the data from an otherwise unavailable perspective.

3.2. Construction of simulated data

By leveraging the algorithm to capture the intricacies of peak properties and reconstructing this information, a simulated chromatogram can be generated, as illustrated in Fig. 4B. This results in a chromatogram closely resembling the raw data, with each peak feature quantified accurately.

Alternatively, rather than reconstructing a chromatogram in simulated form as shown above, the addition of probability distributions allows for the ability to create unique chromatograms that are statistically representable for the modeled data. This is achieved by sampling from the probability distributions and using these parameters to construct a chromatogram that is entirely quantifiable. The separation space is divided into several tiles. The size of these tiles needs to be

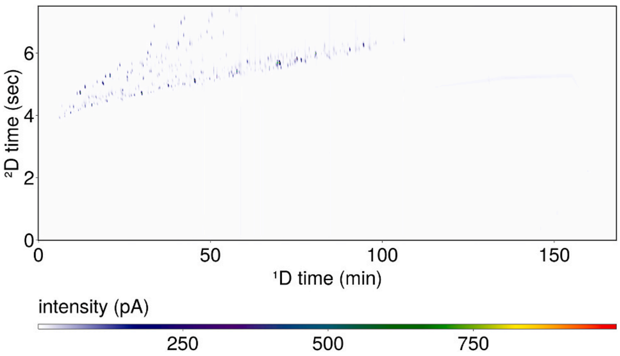


Fig. 6. Simulated GCxGC chromatogram with input parameters based on localized probability.

chosen carefully. If the size of the tiles is too large, the characteristic patterns of the chromatogram will not show up, however, if the tiles are too small the selection of parameters within the tile will be too small. By maintaining a user-defined level of correlation between the probability distributions, one can sample from each of the histograms for the peak model parameters and get a peak that is statistically probable for that location. For instance, in GC × GC the peak width increases with higher second-dimension retention time. By only including values from peaks with a similar location this same trend will inherently still be present in the simulated chromatogram. To achieve this, a starting parameter and dependency margin (i.e. the fuzzy correlation) are chosen. In this example, location is used as the starting parameter. The algorithm selects one of the locations within a tile according to the probability distribution and subsequently collects all the corresponding fit parameters of the peaks with a location within the margin of this initially selected peak. This creates a list of viable parameters for that initial selection. A random combination of peak parameters is taken from this list (*i.e.* not necessarily originating from the same original peak fit), however, due to the fuzzy correlation between the distributions, it is entirely statistically valid that such a peak would be in such a location of the chromatogram. By repeating this proses as many times as desired a simulated chromatogram can be created that is in its combination of peak moments highly representative of the recorded version yet completely original. Following the simulation of peaks, the signal obtained from a blank measurement is overlaid onto the peak signal, ensuring alignment of the baseline and noise characteristics with the original dataset. In Fig. 6 an example of this is shown. To prevent peaks from piling on top of each other a slide random shift in both the ^{1}D and ^{2}D time has been applied. The distributions are not sampled continuously as that would not allow for the needed correlation between the different parameters. Rather, the sampling takes place on the exact values that are put in with on interpolation applied. Therefore, the number of peaks going into the probability distribution of course greatly influences the complexity and density of the resulting chromatogram.

3.3. *Application to LC × LC-UV data*

GC data predominantly exhibit sharp peaks characterized by minimal tailing. In contrast, LC peaks frequently exhibit broader profiles and a higher degree of asymmetry, thereby increasing the likelihood of coelution. Fig. 7 displays the application of our algorithm on a LC × LC-UV chromatogram of a dye mixture. The raw data of this chromatogram can be seen in the supplementary Figure S 27 as well as the threshold determination in supplementary section S3 (Figs. S14–S25). Fig. 7A is an exact rebuild using a Gaussian model whereas 7B is an exact rebuild using the Skewed Lorentz-Normal peak model. With this kind of dataset, the advantage of the two additional terms in the Skewed Lorentz-Normal becomes clear, as it is able to describe the tailing characteristics of the peaks much better than the Gaussian example. Fig. 7C and D, display the residuals of this fit for the Gaussian model and Skewed Lorentz-Normal respectively. These residuals are obtained by subtracting the fitted data from the raw chromatogram. The residuals for the Skewed Lorentz-Normal (Fig. 7D) consist almost entirely of baseline attributes, which the algorithm purposely avoids. The residuals of the Gaussian model (Fig. 7C) consist of these same baseline attributes as well as some peaks that cannot be described well by a Gaussian peak model.

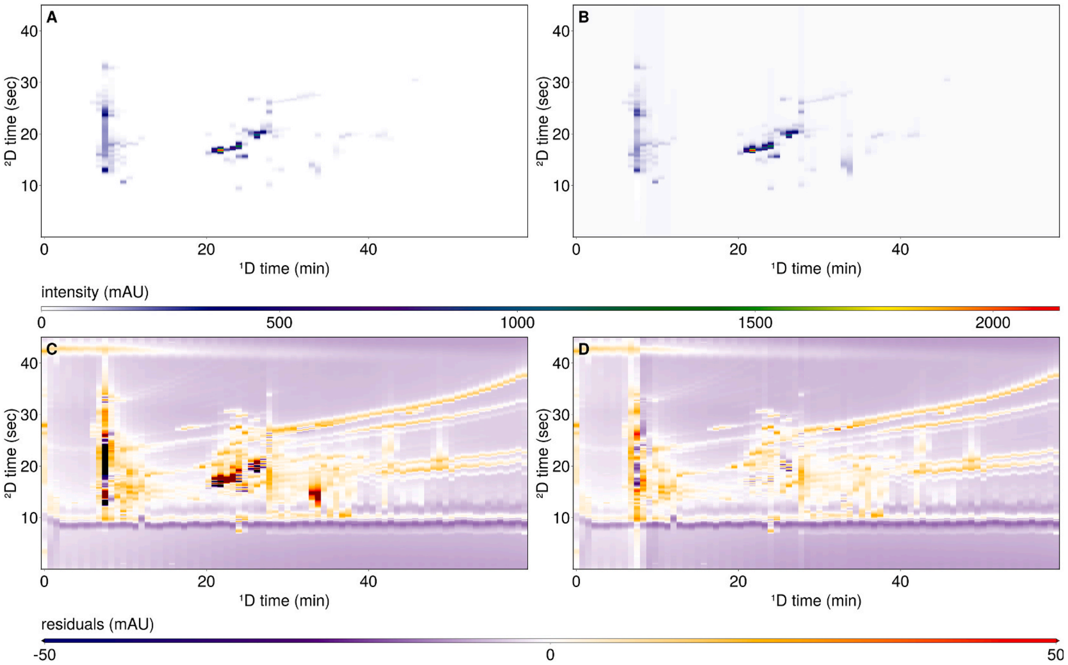


Fig. 7. LC × LC-UV chromatogram of a synthetic dye mixture, with A) reconstruction of the LC × LC chromatogram using peaks fitted with the Gaussian model, B) residuals of the Gaussian fit. C) reconstruction of the LC × LC chromatogram using peaks fitted with the Skewed Lorentz-Normal model. D) residuals of the Skewed Lorentz-Normal VII.

4. Conclusion

A new algorithm to construct highly realistic datasets based on real input chromatograms was developed. The algorithm locates each peak and then models it using a Skewed Lorentz-Normal function to capture detailed peak properties. All of these parameters are used to reconstruct the data set together with noise and a baseline obtained from a solvent blank. This creates a dataset that is highly similar to the original, yet completely numerically defined, allowing it to be used for objective numerical evaluation. Additionally, all of the parameters obtained through fitting are converted into probability distributions that allow us to create new datasets, that have peaks with similar properties to the original yet completely original and characterized. This new approach was necessary because there to date has been no way to conduct an objective numerical evaluation of data (pre-)processing methods. This new workflow generates probability distributions that allow users to construct highly realistic representations of their case-specific datasets. A compromise was made between the absolute values of simulated data and the higher degree of realism and wider scope of parameters associated with experimental data needed for objective numerical evaluation of current and future data analysis algorithms. This new approach allows for the assessment of data and data analysis in a number of ways that were previously not possible.

CRediT authorship contribution statement

Nino B.L. Milani: Writing – original draft, Visualization, Software, Methodology, Investigation, Conceptualization. **Alan Rodrigo García-Cicourel:** Writing – review & editing, Data curation. **Jan Blomberg:** Writing – review & editing, Resources. **Rob Edam:** Resources, Data curation. **Saer Samanipour:** Writing – review & editing. **Tijmen S. Bos:** Writing – review & editing, Supervision, Software. **Bob W.J. Pirok:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

A link to the GitHub is included in the manuscript.

Acknowledgements

This research is part of the PARADISE project (ENPPS.TA.019.001) and received funding from the Dutch Research Council (NWO) in the framework of the Science PPP Fund for the top sectors and from the Ministry of Economic Affairs of the Netherlands in the framework of the "PPS Toeslagregeling".

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2024.342724.

References

[1] X. Shi, S. Wang, Q. Yang, X. Lu, G. Xu, Comprehensive two-dimensional chromatography for analyzing complex samples: recent new advances, Anal. Methods 6 (2014) 7112–7123, https://doi.org/10.1039/C4AY01055H.

[2] X. Li, D.R. Stoll, P.W. Carr, Equation for peak capacity estimation in two-dimensional liquid chromatography, Anal. Chem. 81 (2009) 845–850, https://doi.org/10.1021/AC801772U/ASSET/IMAGES/AC-2008-01772U_M011.GIF.

[3] N.B.L. Milani, E. van Gilst, B.W.J. Pirok, P.J. Schoenmakers, Comprehensive two-dimensional gas chromatography— a discussion on recent innovations, J. Separ. Sci. 46 (2023), https://doi.org/10.1002/jssc.202300304.

[4] G. Groeneveld, B.W.J. Pirok, P.J. Schoenmakers, Perspectives on the future of multi-dimensional platforms, Faraday Discuss 218 (2019) 72–100, https://doi.org/10.1039/C8FD00233A.

[5] B.W.J. Pirok, A.F.G. Gargano, P.J. Schoenmakers, Optimizing separations in online comprehensive two-dimensional liquid chromatography, J. Separ. Sci. 41 (2018) 68–98, https://doi.org/10.1002/jssc.201700863.

[6] R.S. van den Hurk, M. Pursch, D.R. Stoll, B.W.J. Pirok, Recent trends in two-dimensional liquid chromatography, TrAC, Trends Anal. Chem. 166 (2023), https://doi.org/10.1016/j.trac.2023.117166.

[7] T.S. Bos, W.C. Knol, S.R.A. Molenaar, L.E. Niezen, P.J. Schoenmakers, G. W. Somsen, B.W.J. Pirok, Recent applications of chemometrics in one- and two-dimensional chromatography, J. Separ. Sci. 43 (2020) 1678–1727, https://doi.org/10.1002/jssc.202000011.

[8] B.W.J. Pirok, J.A. Westerhuis, Challenges in Obtaining Relevant Information from One- and Two-Dimensional LC Experiments, LCGC North America, 2020, pp. 8–14, https://doi.org/10.56530/LCGC.NA.JK4782S5.

[9] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639, https://doi.org/10.1021/AC60214A047/ASSET/AC60214A047.FP.PNG_V03.

[10] L.E. Niezen, P.J. Schoenmakers, B.W.J. Pirok, Critical comparison of background correction algorithms used in chromatography, Anal. Chim. Acta 1201 (2022) 339605, https://doi.org/10.1016/J.ACA.2022.339605.

[11] P.H.C. Eilers, H.F.M. Boelens, Baseline Correction with Asymmetric Least Squares Smoothing, vol. 1, Leiden University Medical Centre Report, 2005, p. 5.

[12] C.A. Lieber, A. Mahadevan-Jansen, Automated Method for Subtraction of Fluorescence from Biological Raman Spectra, 2003, pp. 1363–1367, https://doi.org/10.1366/000370203322554518, 10.1366/000370203322554518 57.

[13] F. Gan, G. Ruan, J. Mo, Baseline correction by improved iterative polynomial fitting with automatic threshold, Chemometr. Intell. Lab. Syst. 82 (2006) 59–65, https://doi.org/10.1016/J.CHEMOLAB.2005.08.009.

[14] F. Vogt, Data filtering in instrumental analyses with applications to optical spectroscopy and chemical imaging, J. Chem. Educ. 88 (2011) 1672–1683, https://doi.org/10.1021/ED100984C/SUPPL_FILE/ED100984C_SI_002.PDF.

[15] M.F. Wahab, T.C. O'Haver, Wavelet transforms in separation science for denoising and peak overlap detection, J. Separ. Sci. 43 (2020) 1998–2010, https://doi.org/10.1002/JSSC.202000013.

[16] S. Cappadona, F. Levander, M. Jansson, P. James, S. Cerutti, L. Pattini, Wavelet-based method for noise characterization and rejection in high-performance liquid chromatography coupled to mass spectrometry, Anal. Chem. 80 (2008) 4960–4968, https://doi.org/10.1021/AC800166W/ASSET/IMAGES/LARGE/AC-2008-00166W_0003. JPEG.

[17] A. Kensert, G. Collaerts, K. Efthymiadis, P. Van Broeck, G. Desmet, D. Cabooter, Deep convolutional autoencoder for the simultaneous removal of baseline noise and baseline drift in chromatograms, J. Chromatogr. A 1646 (2021) 462093, https://doi.org/10.1016/J.CHROMA.2021.462093.

[18] A. Mani-Varnosfaderani, A. Kanginejad, K. Gilany, A. Valadkhani, Estimating complicated baselines in analytical signals using the iterative training of Bayesian regularized artificial neural networks, Anal. Chim. Acta 940 (2016) 56–64, https://doi.org/10.1016/J.ACA.2016.08.046.

[19] D. Zhang, X. Huang, F.E. Regnier, M. Zhang, Two-dimensional correlation optimized warping algorithm for aligning GC×GC−MS data, Anal. Chem. 80 (2008) 2664–2671, https://doi.org/10.1021/ac7024317.

[20] P.H.C. Eilers, Parametric time warping, Anal. Chem. 76 (2004) 404–411, https://doi.org/10.1021/ac034800e.

[21] H. Parastar, M. Jalali-Heravi, R. Tauler, Comprehensive two-dimensional gas chromatography (GC×GC) retention time shift correction and modeling using bilinear peak alignment, correlation optimized shifting and multivariate curve resolution, Chemometr. Intell. Lab. Syst. 117 (2012) 80–91, https://doi.org/10.1016/j.chemolab.2012.02.003.

[22] G. Tomasi, F. Savorani, S.B. Engelsen, icoshift: an effective tool for the alignment of chromatographic data, J. Chromatogr. A 1218 (2011) 7832–7840, https://doi.org/10.1016/j.chroma.2011.08.086.

[23] R. Bro, C.A. Andersson, H.A.L. Kiers, PARAFAC2—Part II. Modeling chromatographic data with retention time shifts, J. Chemom. 13 (1999) 295–309, https://doi.org/10.1002/(SICI)1099-128X(199905/08)13:3/4<295::AID-CEM547>3.0.CO;2-Y.

[24] G. Vivó-Truyols, J.R. Torres-Lapasió, A.M. van Nederkassel, Y. Vander Heyden, D. L. Massart, Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals: Part II: peak model and deconvolution algorithms, J. Chromatogr. A 1096 (2005) 146–155, https://doi.org/10.1016/j.chroma.2005.03.072.

[25] K.J. Goodman, J.T. Brenna, Curve fitting for restoration of accuracy for overlapping peaks in gas chromatography/combustion isotope ratio mass spectrometry, Anal. Chem. 66 (1994) 1294–1301.

[26] H.-Y. Fu, J.-W. Guo, Y.-J. Yu, H.-D. Li, H.-P. Cui, P.-P. Liu, B. Wang, S. Wang, P. Lu, A simple multi-scale Gaussian smoothing-based strategy for automatic chromatographic peak extraction, J. Chromatogr. A 1452 (2016) 1–9, https://doi.org/10.1016/j.chroma.2016.05.018.

[27] G. Vivó-Truyols, J.R. Torres-Lapasió, A.M. Van Nederkassel, Y. Vander Heyden, D. L. Massart, Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals: Part I: peak detection, J. Chromatogr. A 1096 (2005) 133–145, https://doi.org/10.1016/j.chroma.2005.03.092.

[28] S. Peters, G. Vivó-Truyols, P.J. Marriott, P.J. Schoenmakers, Development of an algorithm for peak detection in comprehensive two-dimensional chromatography, J. Chromatogr. A 1156 (2007) 14–24, https://doi.org/10.1016/j.chroma.2006.10.066.

[29] Q. Song, A. Savant, S.E. Reichenbach, E.B. Ledford, <title>Digital image processing for a new type of chemical separation system</title&gt, in: Applications of Digital Image Processing XXII, SPIE, 1999, pp. 2–12, https://doi.org/10.1117/12.365811.

[30] G. Vivó-Truyols, H.G. Janssen, Probability of failure of the watershed algorithm for peak detection in comprehensive two-dimensional chromatography, J. Chromatogr. A 1217 (2010) 1375–1385, https://doi.org/10.1016/j.chroma.2009.12.063.

[31] I. Latha, S.E. Reichenbach, Q. Tao, Comparative analysis of peak-detection techniques for comprehensive two-dimensional chromatography, J. Chromatogr. A 1218 (2011) 6792–6798, https://doi.org/10.1016/j.chroma.2011.07.052.

[32] Y. Liu, X. Zhou, Y. Yu, A concise iterative method using the Bezier technique for baseline construction, Analyst 140 (2015) 7984–7996, https://doi.org/10.1039/C5AN01184A.

[33] M. Navarro-Reig, C. Bedia, R. Tauler, J. Jaumot, Chemometric strategies for peak detection and profiling from multidimensional chromatography, Proteomics 18 (2018), https://doi.org/10.1002/pmic.201700327.

[34] Z. Li, D.J. Zhan, J.J. Wang, J. Huang, Q.S. Xu, Z.M. Zhang, Y.B. Zheng, Y.Z. Liang, H. Wang, Morphological weighted penalized least squares for background correction, Analyst 138 (2013) 4483–4492, https://doi.org/10.1039/C3AN00743J.

[35] X. Ning, I.W. Selesnick, L. Duval, Chromatogram baseline estimation and denoising using sparsity (BEADS), Chemometr. Intell. Lab. Syst. 139 (2014) 156–167, https://doi.org/10.1016/J.CHEMOLAB.2014.09.014.

[36] H.Y. Fu, H.D. Li, Y.J. Yu, B. Wang, P. Lu, H.P. Cui, P.P. Liu, Y. Bin She, Simple automatic strategy for background drift correction in chromatographic data analysis, J. Chromatogr. A 1449 (2016) 89–99, https://doi.org/10.1016/J.CHROMA.2016.04.054.

[37] S.-J. Baek, A. Park, Y.-J. Ahn, J. Choo, Baseline correction using asymmetrically reweighted penalized least squares smoothing, Analyst 140 (2014) 250–257, https://doi.org/10.1039/C4AN01061B.

[38] I. Selesnick, Sparsity-assisted signal smoothing (revisited), ICASSP, in: IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2017, pp. 4546–4550, https://doi.org/10.1109/ICASSP.2017.7953017.

[39] V. Mazet, C. Carteret, D. Brie, J. Idier, B. Humbert, Background removal from spectra by designing and minimising a non-quadratic cost function, Chemometr. Intell. Lab. Syst. 76 (2005) 121–133, https://doi.org/10.1016/J.CHEMOLAB.2004.10.003.

[40] J.A. Navarro-Huerta, J.R. Torres-Lapasió, S. López-Ureña, M.C. García-Alvarez-Coque, Assisted baseline subtraction in complex chromatograms using the BEADS algorithm, J. Chromatogr. A 1507 (2017) 1–10, https://doi.org/10.1016/J.CHROMA.2017.05.057.

[41] J. Engel, J. Gerretzen, E. Szymańska, J.J. Jansen, G. Downey, L. Blanchet, L.M.C. Buydens, Breaking with trends in pre-processing? TrAC, Trends Anal. Chem. 50 (2013) 96–106, https://doi.org/10.1016/j.trac.2013.04.015.

[42] X. Bian, K. Wang, E. Tan, P. Diwu, F. Zhang, Y. Guo, A selective ensemble preprocessing strategy for near-infrared spectral quantitative analysis of complex samples, Chemometr. Intell. Lab. Syst. 197 (2020) 103916, https://doi.org/10.1016/j.chemolab.2019.103916.

[43] A.P. De La Mata, J.J. Harynuk, Limits of detection and quantification in comprehensive multidimensional separations. 1. a theoretical look, Anal. Chem. 84 (2012) 6646–6653, https://doi.org/10.1021/ac3010204.

[44] H. Parastar, M. Jalali-Heravi, R. Tauler, Comprehensive two-dimensional gas chromatography (GC×GC) retention time shift correction and modeling using bilinear peak alignment, correlation optimized shifting and multivariate curve resolution, Chemometr. Intell. Lab. Syst. 117 (2012) 80–91, https://doi.org/10.1016/j.chemolab.2012.02.003.

[45] B.A. Weggler, L.M. Dubois, N. Gawlitta, T. Gröger, J. Moncur, L. Mondello, S. Reichenbach, P. Tranchida, Z. Zhao, R. Zimmermann, M. Zoccali, J.F. Focant, A unique data analysis framework and open source benchmark data set for the analysis of comprehensive two-dimensional gas chromatography software, J. Chromatogr. A (2021) 1635, https://doi.org/10.1016/j.chroma.2020.461721.

[46] V.G. Van Mispelaar, A.K. Smilde, O.E. De Noord, J. Blomberg, P.J. Schoenmakers, Classification of highly similar crude oils using data sets from comprehensive two-dimensional gas chromatography and multivariate techniques, J. Chromatogr. A 1096 (2005) 156–164, https://doi.org/10.1016/j.chroma.2005.09.063.

[47] B.W.J. Pirok, J. Knip, M.R. van Bommel, P.J. Schoenmakers, Characterization of synthetic dyes by comprehensive two-dimensional liquid chromatography combining ion-exchange chromatography and fast ion-pair reversed-phase chromatography, J. Chromatogr. A 1436 (2016) 141–146, https://doi.org/10.1016/J.CHROMA.2016.01.070.

[48] J. Bezanson, S. Karpinski, V.B. Shah, A. Edelman, Julia: A Fast Dynamic Language for Technical Computing, 2012. http://arxiv.org/abs/1209.5145.

[49] S. Danisch, J. Krumbiegel, Makie.jl: flexible high-performance data visualization for Julia, J. Open Source Softw. 6 (2021) 3349, https://doi.org/10.21105/joss.03349.

[50] J. Bao, W. Wang, T. Yang, G. Wu, An incremental clustering method based on the boundary profile, PLoS One 13 (2018), https://doi.org/10.1371/journal.pone.0196108.

[51] T.R. Noviandy, A. Maulana, N.R. Sasmita, R. SuhendraMuslem, G.M. Idroes, M. Paristiowati, Z. Helwani, E. Yandri, S. RahimahMuhammad, Irvanizam, R. Idroes, The implementation of K-Means clustering in kovats retention index on gas chromatography, IOP Conf. Ser. Mater. Sci. Eng. 1087 (2021) 012051, https://doi.org/10.1088/1757-899x/1087/1/012051.