AutoGCA - a software tool for the automated analysis of gas chromatograms

Henning Schröder[e,a], Tomass Andersons[a], Alexander Brächer[c], Roland Peschke[c], Martina Beese[a,b], Christoph Kubis[b], Mathias Sawall[a], Robert Franke[c,d], Klaus Neymeyr[a,b]

[a] *Universität Rostock, Institut für Mathematik, Ulmenstrasse 69, 18057 Rostock, Germany*
[b] *Leibniz-Institut für Katalyse, Albert-Einstein-Strasse 29a, 18059 Rostock, Germany*
[c] *Evonik Industries AG, Paul-Baumann Straße 1, 45772 Marl, Germany*
[d] *Lehrstuhl für Theoretische Chemie, Ruhr-Universität Bochum, 44780 Bochum, Germany*
[e] *PLANET AI GmbH, Warnowufer 60, 18057 Rostock, Germany*

Abstract

Gas chromatography (GC) is an important tool in analytical chemistry and large amounts of data are routinely produced. Despite the advances in the available software, the analysis of a dataset can still be time-consuming task, in part due to the manual or semi-automated marking of peaks. An automation of this tedious task is proposed in this work.

Keywords: Baseline, gas chromatography, peak detection, peak fitting

1. Introduction

With the widespread availability and use of different chromatographic techniques comes the need to interpret large amounts of experimental data. To extract valuable information from a chromatogram, the area and position of the contributing peaks must be determined. We call this the Peak Extraction Task (PET). It can be solved easily if the peaks are clearly separated. Therefore, we focus on chromatograms with a high number of at least partially overlapping peaks that might be affected by baseline distortion. Due to their difficulty, these cases are typically analyzed manually or semi-automatically, e.g., with peak suggestions that must be fine-tuned with a high expenditure of time. Although the full automation of PET has been extensively researched [1, 2, 3, 4, 5], it often has poor applicability for highly overlapping or baseline distorted peaks, or there is a large number of parameters to be manually tuned.

This paper presents the concepts of AutoGCA (Automatic Gas Chromatogram Analysis), a fully automated algorithm for the PET. The novel aspects are the use of a fully asymmetric peak model, an automated noise level estimation and baseline approximation. Despite the name, the concepts are applicable or at least transferable to general chromatographic applications. All source code is available online on GitHub `https://github.com/henning1419/gcd` and is written in MATLAB. The significance and application of this work arose in cooperation with Evonik Industries AG.

The motivation for automating tedious chromatogram evaluations is to save time of researchers and laboratory staff, see [2] and more recently [1] as examples. Some methods in our work have been inspired by the work on peak detection in multi-overlapping chromatographic signals by Vivó-Truyols et al. [4]. Some review papers also consider a multitude of methods that are relevant for solving the PET. One such work is [6], which lists approaches for background correction, peak detection, and peak properties, among others.

1.1. The experimental dataset

In this paper, AutoGCA is explained and evaluated with the help of an GC dataset of an oxo oil LS13 from Evonik Oxeno GmbH & Co. KG, which is a paraffin-olefin-mixture. This oxo oil, being a distillation section from the production of C13 oxo alcohols is manufactured in continuous production processes at Evonik's largest production location in Marl, Germany.

This complex mixture of predominantly branched C12 alkanes and alkenes are used as intermediates for the production of solvents and as blending components for diesel fuels. A major challenge in the analysis of the GC data of this product is the detection and separation of the large amount of isomers which is well in the three-digit range.

This dataset is representative for some of the particular situations that were encountered in the development of AutoGCA, for example, the dataset contains overlapping peaks including peaks on shoulders of other peaks and there is a strong baseline drift in the last part of the dataset.

The manual evaluation was done with Agilent OpenLab CDS.

1.2. List of symbols

g	chromatographic signal,
t	retention time grid,
$s, \tilde{s}$	original signal segment and smoothed signal segment,
α	peak width ratio parameter for baseline correction,
k	index of the center of the largest peak in a segment,
i_r, i_l	indices (right and left) for the peak width calculation,
b	baseline,
I	interval between two adjacent supporting points of the baseline,
g_B	baseline-corrected chromatogram,
w	windows for the noise level approximation,
η	approximated noise level,
$g(x;\mu,\sigma)$	Gaussian with the center μ and standard deviation σ,
$l(x;\mu,\sigma)$	a nearly linear curve model,
$p(x;\mu,h,\sigma_L,\sigma_R,\alpha_L,\alpha_R)$	peak model with center μ, height h, standard deviation σ_L (left) and σ_R (right) and ratio of the nearly linear curve α_L (left) and α_R (right),
$r^{(i)}$	the residual of a cluster segment after i fitted peaks.

2. The overview of the algorithm

For a given chromatogram, AutoGCA extracts the positions, areas, and other complementary features of the underlying peaks. The general concept is outlined in Fig. 1 and explained in this section.

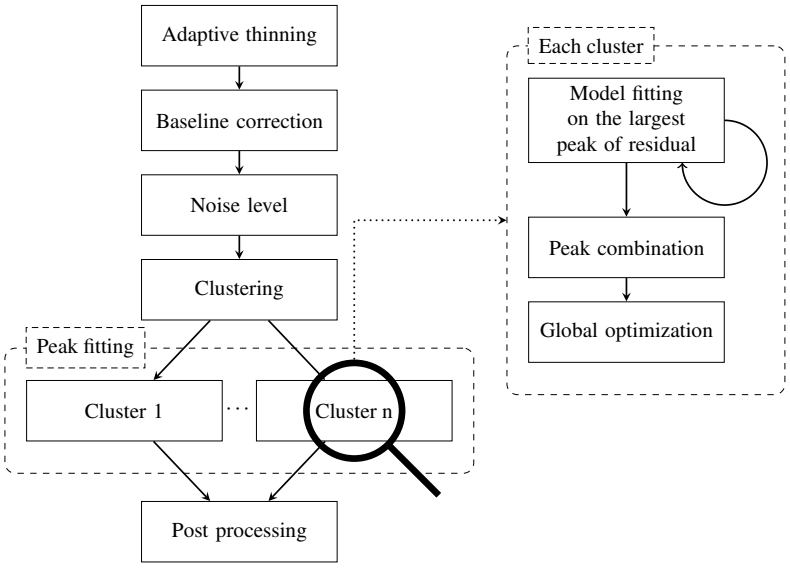

Figure 1: A flowchart describing the AutoGCA algorithm.

Remark 2.1. *It should be noted that in some cases precise parameter specifications are given. These have been determined heuristically by analyzing various GC profiles and are fine-tuned to the problems that motivated AutoGCA. Since the source code is available online, adjustments can be made to adapt the method to different scenarios.*

2.1. Normalization along the retention time axis

The AutoGCA algorithm requires only the chromatographic signal $g \in \mathbb{R}^n$ and the corresponding retention time grid $t \in \mathbb{R}^n$ as inputs. In a first step a simple linear interpolation is used to downscale g to 400 data points per minute of retention time. This value is chosen to ensure that even narrow peaks of interest are preserved by a sufficient number of data points, but also reduces the computational time of all subsequent steps. In addition, subsequent algorithmic steps can be based on this normalized input and parameter values can be fine-tuned, see Remark 2.1. For simplicity, the corrected signal is also denoted g.

2.2. Baseline correction

A fundamental problem in the analysis of chromatograms is the drift of the baseline, see Fig.2. This can also be interpreted as low frequency noise, which must be subtracted. It differs from classical noise, which is characterized by high-frequency stochastic signals, see [7]. Baseline correction is a well-studied field with many established methods. Considerations of baseline drift correction for GC datasets date back to the work of Wilson and McInnes [8] in 1965. Since then, several approaches have been successfully implemented and compared, e.g. [7, 9]. Various reviews have also been written on the subject, see for example [8, 10] or more recently [11]. Baseline correction is often part of more general reviews, such as [6].

Automated GC profile analysis requires a robust correction method that ensures accurate results by preventing distortion of peak areas. This is particularly challenging in areas with many overlapping peaks, because automated methods often tend to overestimate baselines, thereby cropping some areas of the peaks. We propose a method that determines a smooth curve that, firstly, lies predominantly below the signal g and secondly, is as close to the signal as possible, except in areas with peaks or peak clusters. Our algorithm is based on the assumption that the baseline of two adjacent narrow signal segments containing only a few peaks, can be determined by connecting the minimal points within each signal segment. The challenge is to reliably determine such segments.

The algorithm starts with the entire signal g and recursively divides it into segments. Each segmentation step involves smoothing the signal, detecting the largest peak, and checking whether further division is necessary. For each signal segment s with a number of $\mathrm{len}(s)$ data points, a termination criterion is evaluated first: If $\mathrm{len}(s) \cdot \alpha < 22$ with $\alpha = 0.43$ (see Remark 2.1) holds, further segmentation is stopped. If this criterion is not met, the width of the largest peak is determined. The signal segment s is smoothed with a Savitzky-Golay filter (window width 40, degree 3) to obtain $\tilde{s}$. The maximum of $\tilde{s}$ indicates the center of the largest peak at index k. To determine the peak width:

- for the right (decreasing) flank, increment i from k until the condition $\tilde{s}_i < \tilde{s}_{i+1}$ (increasing trend between two adjacent points) has happened 80 times, resulting in i_r.
- for the left flank, decrement i from k analogously to find i_l.

If $(i_r - i_l)/\mathrm{len}(s) \geq \alpha$ holds, the current segment is characterized as a peak. Otherwise, the peak is considered as narrow and it is assumed that the segment can be subdivided into two segments, see Fig. 3 (top). The division index is determined by, firstly, subtracting a linear interpolation of $\tilde{s}$ based on its start and end points, resulting in $\bar{s}$, and secondly, finding the minimum value within the centered 33% range of $\bar{s}$, see Fig. 3 (center). This process is repeated recursively until segments are too small or are characterized as a peak.

The minima in each segment as well as the first and last point define a list of supporting points, see Fig. 3 (bottom). They are used to calculate a piecewise cubic Hermite interpolating polynomial b. To ensure that the baseline is predominantly below the signal, the ratio

$$\frac{\int_I \min(s-b, 0)\,\mathrm{d}t}{\int_I s - b\,\mathrm{d}t}$$

is calculated for each interval I between two adjacent supporting points. If it is greater than 0.1, then this interval is subdivided as described above and the Hermite splines are recalculated. Finally, the baseline is subtracted from the signal and the corrected chromatogram, denoted by g_B, is used for further calculations.

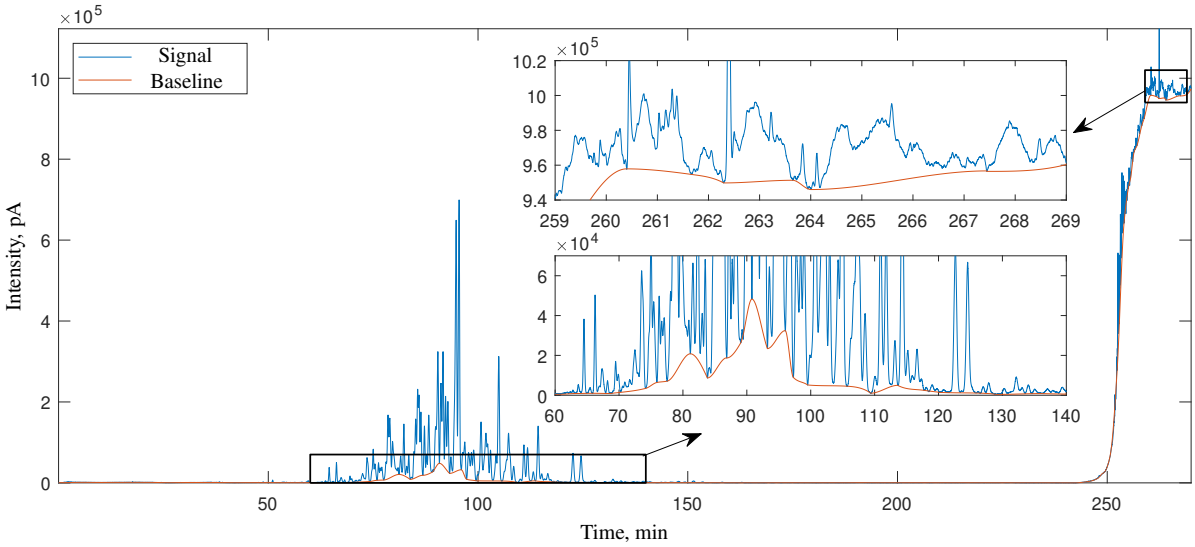

Figure 2: The dataset from Sec. 1.1 and the calculated baseline in red.

2.3. Noise level approximation

Unlike the baseline, the high-frequency noise in the signal does not usually produce very distorted results. However, it can still be difficult to distinguish between the noise and very small peaks. This has been recognized in the literature and approaches based on classical measures such as the median absolute deviation are presented in [2]. There they are used in a similar fashion - to validate peaks.

In AutoGCA, however, we assume that there exists a sufficiently long segment in g_B that does not contain any peaks. This is true in most applications. In any possible window w of g_B with the length of 501 data points (see Remark 2.1) the distance between the maximum and the minimum values of w is calculated. The approximated noise level η is then defined as 1.3 times the minimum over all these distances. The prefactor results from the observation that the subsequent steps perform better if the noise is slightly overestimated. This noise intensity approximation is used not only to ignore peaks below the noise level, but also in several other parts of AutoGCA to achieve robustness in numerical methods or to define termination criteria.

2.4. Clustering

The use of a clustering step is justified similarly to [4], because it allows parallel execution of peak fitting, the most time-consuming part of the algorithm. Initially, g_B is divided into a number of 300 equidistant segments. The mean value over each segment is subtracted from the data so that the mean of the segment is zero. Then it is checked whether the segment contains values outside the interval $[-\eta, \eta]$ with the noise level approximation η of Sec. 2.3. If so, this segment and the two preceding and two following segments are assumed to contain peaks. After application to all segments, all remaining non-peak segments are neglected in the following steps. Furthermore, this defines independent clusters of subsequent segments of g_B, in a sense that peak fitting can be performed in parallel, which can drastically reduce the computation time.

2.5. Peak model

The modeling of a peak in various chromatographic applications and more generally in the natural sciences is a long-standing problem. The underlying peak model is crucial for our approach to automate peak extraction. There are classical peak models such as Gaussian, Lorentzian, Voigt or the Exponentially Modified Gaussian (EMT) among others, see e.g. [12, 13]. However, these models are severely limited when certain effects such as tailing or fronting occur. Thus, our goal is to pragmatically approximate the actual peak and not necessarily to find a single peak model

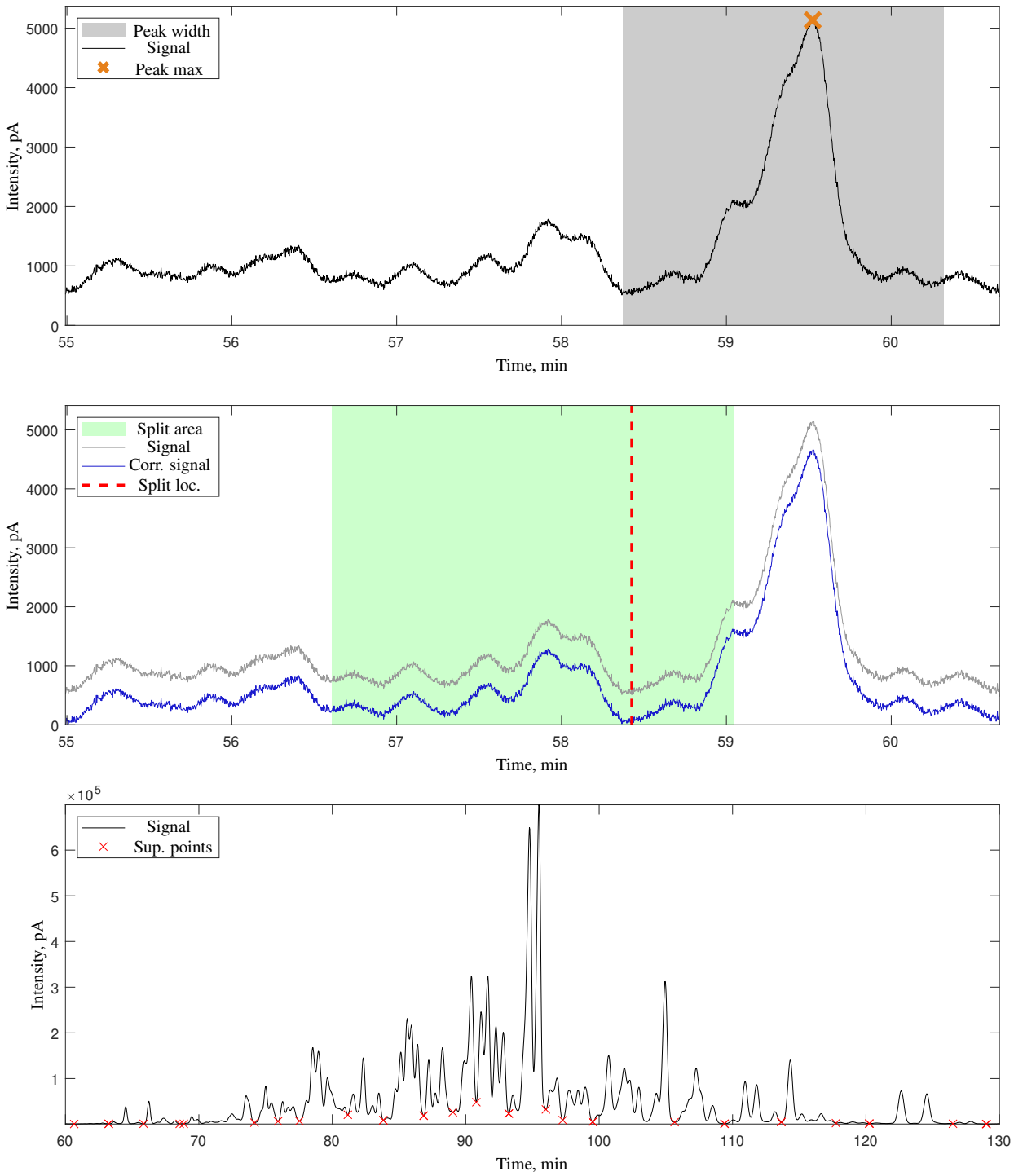

Figure 3: Visualization of the splitting stages to determine the baseline support points. Top: The localization of the signal maximum (orange cross) and the subsequent approximation of the peak width (gray area) are shown. The approximated peak is considered narrow, because the ratio of its width to the total section length is small enough. This induces the splitting of the segment. Center: The determination of the final position of the segment split is shown. The gray line marks the original signal, while the blue line is corrected by subtracting a linear interpolant. The green area marks the centered 33% of the segment where splitting is allowed. Finally, the red dashed line marks the minimum of the blue signal within the green area and thus the final splitting location. Bottom: Plot of the resulting non-equidistant baseline support points of the portion of the signal with large peaks.

that is physically correct. The two main ideas for the reconstruction of complex peak shapes are to use a simple model that accounts for possible peak asymmetry, and to use superpositions of several of these simple models.

A simple and asymmetric peak model: The peak is divided at the center into its left and right parts, which are then considered separately in terms of their parameters. Only the peak height is shared between the two sides. Each flank is a convex combination of a Gaussian $g(x;\mu,\sigma)$ and a nearly linear curve $l(x;\mu,\sigma)$. The latter is obtained by smoothing the function

$$f(x;\mu,\sigma) = \begin{cases} 1 & |x-\mu| \le \frac{1}{8}\sigma \\ 0 & |x-\mu| \ge \frac{17}{8}\sigma \\ \frac{17}{16} - \operatorname{sgn}(x)\frac{x}{2\sigma} & \text{else} \end{cases}$$

with a moving average with a window size of $0.05 \cdot \sigma$. Thus, the final peak model p with center μ is given by

$$p(x;\mu,h,\sigma_L,\sigma_R,\alpha_L,\alpha_R) = \begin{cases} h((1-\alpha_L)g(x;\mu,\sigma_L) + \alpha_L l(x;\mu,\sigma_L)) & x \le \mu \\ h((1-\alpha_R)g(x;\mu,\sigma_R) + \alpha_R l(x;\mu,\sigma_R)) & x > \mu \end{cases},$$

see also Fig. 4. The linear part especially addresses the problem of *fronting* and *tailing*. This peak model only overcomes some of the limitations of such simple models, but still cannot approximate general peak shapes.

Superpositions: For experimental data, a peak is modeled as the sum of several subpeaks, each of which conforms to the proposed peak model, see Fig. 4. This allows a wide variety of peaks to be modeled while maintaining the relative simplicity of the underlying optimization problem, as each of the subpeaks can be fitted individually and then added to form the final peak.

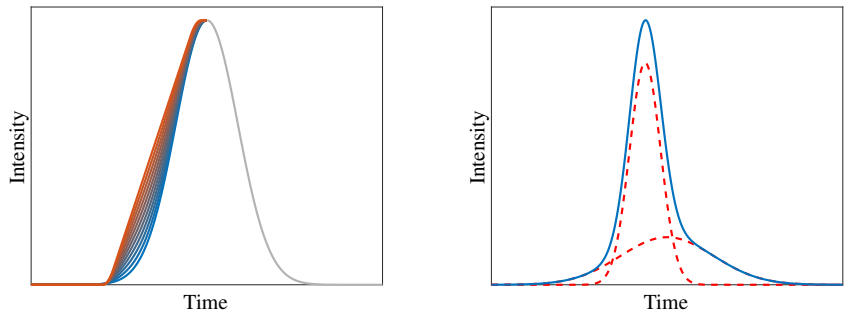

Figure 4: Left: A model of the left flank of a peak model, showing the almost linear curve (red), Gaussian (blue) and their convex combinations. Right: Two subpeaks approximating a more complex peak.

2.6. Peak detection, fitting and combination

Peak detection is a multidisciplinary problem and the approaches are transferable to different applications, e.g. a comparison of methods in mass spectrometry [14] includes methods for smoothing, baseline correction and peak finding criterion that are also useful in chromatography. In particular, derivatives are often used to detect peaks [15, 3] in a variety of applications. Some practical approaches to automating peak detection are given in, e.g., [2, 4, 5, 16].

We have adopted a more classical residual approach. Starting with the complete cluster segment $r^{(0)}$ of g_B, initial parameter estimates are determined for the largest peak. These parameters are then refined by optimization. The resulting peak profile p is subtracted from the cluster segment, updating the residual profile to $r^{(i+1)} = r^{(i)} - p$. The process is repeated until the maximum of the residual profile is less than 3η, the relative model fit error $\|r^{(i+1)}\|_2^2 / \|r^{(0)}\|_2^2$ is below 0.001 or other constraints are met, e.g. the maximum number of peaks or the maximum number of failed optimization attempts (see Remark 2.1).

For the optimization, we use a weighted combined objective function based on a tight peak range and broad peak range to address multiple peaks that are close to each other. Fitting only a single peak would typically lead to over- and underestimation at the locations of neighboring peaks. In the tight range, the model should fit the current $r^{(i)}$ exactly. In the broad range, it is sufficient if the model is smaller than $r^{(i)}$. We solve the optimization problem with the nonlinear least squares solver `lsqnonlin` in MATLAB, based on the trust-region-reflective algorithm [17].

Finally, we apply a rule-based approach to decide whether the fitted peaks should be combined into subsets describing more chemically meaningful peaks. A final simultaneous fitting of all peaks often results in slightly better quality, but also adds a significant amount of computational time.

3. Results: AutoGCA versus manual GC analysis

The results of AutoGCA are illustrated here using the example dataset described in Sec. 1.1. The total computation time of the algorithm was 57.1 seconds on a desktop computer with a 13th generation Intel Core i9 processor. Most of this time was used for peak fitting, which took 55.5 seconds. Six parallel MATLAB workers from the Parallel Computing Toolbox were used to reduce the computation time. The total computation time without parallelization was 92.2 seconds. The result of AutoGCA is both an integral table and a more detailed description of the peaks, see Fig. 5. The final result explains 97.7% of the area under the curve of the dataset.

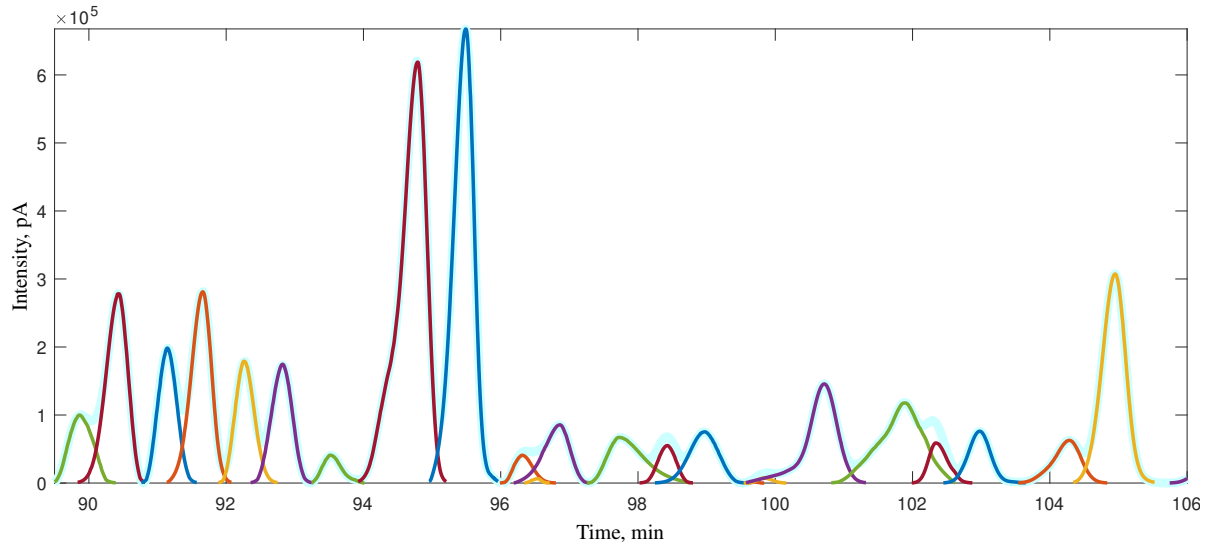

Figure 5: Some of the peaks resulting from an AutoGCA analysis of the dataset in Sec. 1.1 shown in different colors. The wide cyan line in the background shows the baseline corrected raw data.

For comparison, this dataset was evaluated by two experienced chemical analysts, designated User A and User B. The manual evaluation of the dataset took approximately 10 minutes for each analyst, and only the time interval [48, 170] minutes of the data set was evaluated. The peaks in Table 1 refer only to peaks in this time interval and only the peak locations in the resulting integration tables were compared. The matching of peaks based on their location between the different results of AutoGCA, User A and User B was performed with a retention time tolerance of 0.07 minutes. Only the two closest peaks between two evaluations were matched and no peak area information was used.

Interestingly, the results of the two manual evaluations do not agree very well. This can be explained by the challenging nature of the dataset. For example, the peak at 94.8 minutes (light blue in Fig. 5) was recognized as a single peak by AutoGCA and User B, but User A split it into two separate peaks. The results in Table 1 show that AutoGCA is in reasonably good agreement with the users. The results for small peaks have the most discrepancies with the manual evaluation.

In Fig. 6 the evaluations are compared in more detail by showing the matched peaks in black and peaks unique to only one of the evaluations in a different color. It can be seen that some discrepancies can be explained by different splitting of peaks. Sometimes the peak centers were too far apart to be considered the same peak. Also, AutoGCA uses a different baseline than the one used in the manual evaluation, which explains some differences in the peak areas.

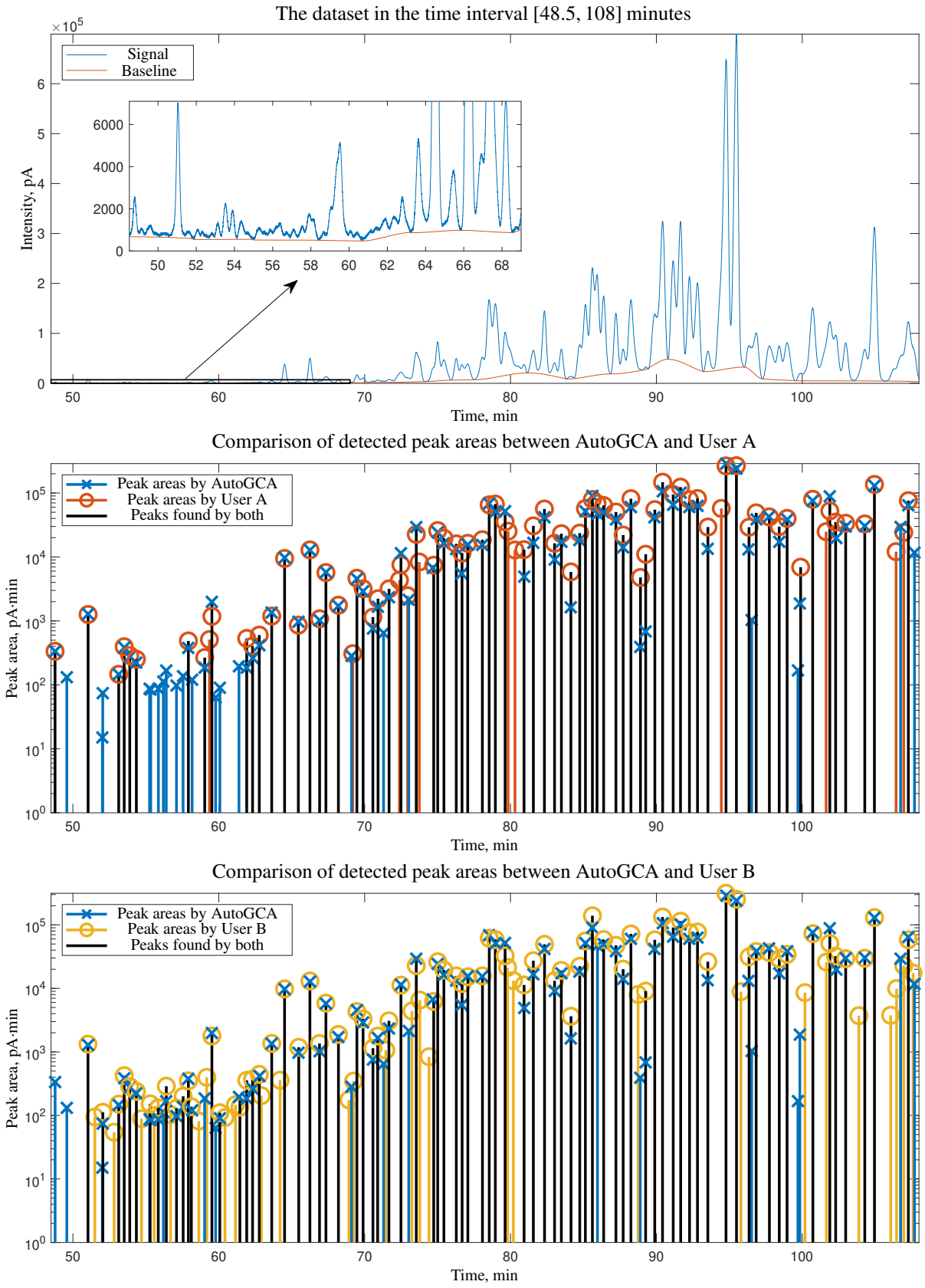

Figure 6: A comparison of the results: The dataset (top) as well as the detected peaks and their areas.

Rate of detection for peaks

Peak prediction by	Peak reference by	Peak size			
		All	Small	Medium	Large
AutoGCA	User A	81.4%	62.3%	86.0%	95.0%
AutoGCA	User B	85.0%	72.7%	88.4%	90.0%
User A	AutoGCA	89.1%	88.9%	74.0%	79.2%
User B	AutoGCA	75.3%	66.7%	71.7%	85.7%
User A	User B	90.6%	87.0%	90.0%	87.5%
User B	User A	73.4%	56.0%	84.9%	100.0%
AutoGCA	Both by User A and by User B	77.1%	57.1%	86.0%	90.0%
Both by User A and by User B	AutoGCA	93.1%	93.6%	82.2%	85.7%
AutoGCA	All peaks by users	89.3%	77.9%	88.4%	95.0%

Table 1: The rate of peaks from the prediction evaluation that were also found in the reference evaluation. Only the position of the peak center was taken into account. "Both by User A and by User B" refers only to those peaks which both users found in agreement. "All peaks by users" refers to peaks found by User A, User B or both. All peaks were heuristically divided into three groups according to their area, with medium peaks ranging from 4000 pA·min to 45000 pA·min.

4. Conclusion

The analysis of gas chromatogram peaks by personnel in industrial processes is often a time-consuming task. Especially when this analysis is a recurring task, an automated procedure seems to be welcome. This report covers several areas of study and presents novel methods for automating the analysis of GC data. In particular, automated peak integration of a challenging GC dataset can be achieved with similar precision to a time-consuming manual evaluation. Furthermore, an automated analysis always produces the same results, but a manual evaluation can vary from person to person. While automation of each of the required steps has been researched individually, a combined and practical automated approach is lacking in traditional software solutions. The proposed AutoGCA method has been successfully applied in industrial process analysis at Evonik Oxeno GmbH & Co. KG. We hope that this report demonstrates the potential time savings that can be achieved by automating routine procedures in analytical chemistry and serves to motivate the development of an in-depth integrated industrial solution based on the ideas of AutoGCA.

References

[1] F.H.G. Zucatelli and R.K. Nogueira. Tea component analysis: From chromatography raw data to a fully automated report. *Chromatographia*, 85:193–211, 2022. doi: 10.1007/s10337-021-04118-8.

[2] S.J. Dixon, R.G. Brereton, H.A. Soini, M.V. Novotny, and D.J. Penn. An automated method for peak detection and matching in large gas chromatography-mass spectrometry data sets. *J. Chemom.*, 20:325–340, 2006. doi: 10.1002/cem.1005.

[3] J.L. Excoffier and G. Guiochon. Automatic peak detection in chromatography. *Chromatographia*, 15:543–545, 1982. doi: 10.1007/BF02280372.

[4] G. Vivó-Truyols, J.R. Torres-Lapasió, A.M. van Nederkassel, Y. Vander Heyden, and D.L. Massart. Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals: Part I: Peak detection. *J. Chromatogr. A*, 1096(1-2):133–45, 2005. doi: 10.1016/j.chroma.2005.03.092.

[5] K.H. Jarman, D.S. Daly, K.K. Anderson, and K.L. Wahl. A new approach to automated peak detection. *Chemom. Intell. Lab. Syst.*, 69(1-2): 61–76, 2003. doi: 10.1016/S0169-7439(03)00113-8.

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

[7] X. Ning, I.W. Selesnick, and L. Duval. Chromatogram baseline estimation and denoising using sparsity (BEADS). *Chemom. Intell. Lab. Syst.*, 139:156–167, 2014. doi: 10.1016/j.chemolab.2014.09.014.

[8] J.D. Wilson and C.A.J. McInnes. The elimination of errors due to baseline drift in the measurement of peak areas in gas chromatography. *J. Chromatogr. A*, 19:486–494, 1965. doi: 10.1016/S0021-9673(01)99489-0.

[9] J. Schneidewind and H. Olickel. Improving data analysis in chemistry and biology through versatile baseline correction. *Chem. Methods*, 1 (2):89–100, 2021. doi: 10.1002/cmtd.202000027.

[10] Ł. Komsta. Comparison of several methods of chromatographic baseline removal with a new approach based on quantile regression. *Chromatographia*, 73:721–731, 2011. doi: 10.1007/s10337-011-1962-1.

[11] L.E. Niezen, P.J. Schoenmakers, and B.W.J. Pirok. Critical comparison of background correction algorithms used in chromatography. *Anal. Chim. Acta*, 1201:339605, 2022. doi: 10.1016/j.aca.2022.339605.

[12] J.P. Foley. Equations for chromatographic peak modeling and calculation of peak area. *Anal. Chem.*, 59(15):1984–1987, 1987. doi: 10.1021/ac00142a019.

[13] J.J. Baeza-Baeza and M.C. García-Alvarez-Coque. Characterization of chromatographic peaks using the linearly modified gaussian model. Comparison with the bi-Gaussian and the Foley and Dorsey approaches. *J. Chromatogr. A*, 1515:129–137, 2017. doi: 10.1016/j.chroma.2017.07.087.

[14] C. Yang, Z. He, and W. Yu. Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. *BMC Bioinf.*, 10:4, 2009. doi: 10.1186/1471-2105-10-4.

[15] F. Holler, D.H. Burns, and J.B. Callis. Direct use of second derivatives in curve-fitting procedures. *Appl. Spectrosc.*, 43(5):877–882, 1989. doi: 10.1366/0003702894202292.

[16] M. Lopatka, G. Vivó-Truyols, and M.J. Sjerps. Probabilistic peak detection for first-order chromatographic data. *Anal. Chim. Acta*, 817:9–16, 2014. doi: 10.1016/j.aca.2014.02.015.

[17] T. Coleman and Y. Li. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optim.*, 6(2):418–445, 1996. doi: 10.1137/0806023.