Сорбционные и хроматографические процессы. 2024. Т. 24, № 6. С. 885-895.
Sorbtsionnye i khromatograficheskie protsessy. 2024. Vol. 24, No 6. pp. 885-895.

ISSN 1680-0613

Original article

Selecting initial values for iterative fitting of chromatographic peaks with exponentially modified Gaussian function

Mikhail D. Khrisanfov[1,2]✉, Andrey S. Samokhin[1,2]

[1]Lomonosov Moscow State University, Chemistry Department, Moscow, Russian Federation, khrisanfovmike@gmail.com✉

[2]Institute of Physical Chemistry and Electrochemistry RAS (IPCE RAS), Moscow, Russian Federation

Abstract. Various mathematical functions are used to describe shapes of chromatographic peaks. Some of these functions, such as exponentially modified gaussian, polynomially modified gaussian or parabolic variance gaussian are based on the normal distribution, and some are not. These functions have from 4 to 9 parameters that need to be iteratively optimized to fit the model function to the experimental data. Many of these functions are numerically unstable, therefore choosing an optimal initial guess of their parameters becomes crucial for successful fitting. The most commonly employed approaches are based on empirical equations relating the basic peak shape parameters (asymmetry value, width at 10% of peak height) and the parameters of the model function. Additionally, the algorithms for calculating the basic peak shape parameters are not thoroughly described in the literature.

Exponentially modified gaussian (EMG) was used as a model function in this work as it is a *de facto* standard in chromatography. Implementations of EMG in Python programming language libraries were listed. The numerical instability of SciPy implementation was investigated for symmetrical peaks and its probable causes were discussed. It was shown that Kalambet's approach to calculating EMG (based on using several equations depending on the shape of the chromatographic peak) did not show such instability.

Approaches to calculate base peak parameters were discussed. Algorithms to find the apex coordinates, left and right halfwidths and width at selected peak height (10% to 50% of the peak maximum) were described.

It is widely known that the relation between the basic peak shape parameters and the parameters of the EMG function is not linear. The empirical equations that approximate these relations were suggested by Foley and Dorsey in the 1980s. We suggested using interpolation by splines instead. This approach significantly improved accuracy in estimating the model function parameters and allowed broadening the range of usable peak shape values. Splines can be calculated once and knots together with spline coefficients can be saved for future use.

In most of the articles and manuals width and halfwidths of the peak are calculated at 10% of the height. Alternative heights (10% to 50%) to calculate basic peaks parameters were tested. It was concluded that parameters of the EMG function can be calculated without significant difference in accuracy at different heights (from 10 to 30%) for noise-free peaks. For noisy data (S/N=100) 30-35% of the peak height can be considered as an alternative.

Keywords: chromatographic peak shape, exponentially modified gaussian, EMG, gas chromatography, GC, liquid chromatography, HPLC.

Acknowledgments: this research was supported by the Ministry of Science and Higher Education of the Russian Federation (№ 124041900012-4). The authors acknowledge support from M.V.Lomonosov Moscow State University Program of Development for providing access to the Pegasus HT (LECO) mass spectrometer..

For citation: Khrisanfov M.D., Samokhin A.S. Selecting initial values for iterative fitting of chromatographic peaks with exponentially modified gaussian function. *Sorbtsionnye i khromatograficheskie protsessy*. 2024. 24(6): 885-895. https://doi.org/10.17308/sorpchrom.2024.24/12566

ОРИГИНАЛЬНЫЕ СТАТЬИ

Научная статья
УДК 543.544
doi: 10.17308/sorpchrom.2024.24/12566

Выбор начальных значений для итерационной аппроксимации хроматографических пиков экспоненциально модифицированной гауссианой

Михаил Дмитриевич Хрисанфов[1,2✉], Андрей Сергеевич Самохин[1,2]
[1]Московский государственный университет имени М.В. Ломоносова, Химический факультет, Москва, Россия, khrisanfovmike@gmail.com✉
[2]Институт физической химии и электрохимии им. А.Н. Фрумкина РАН, Москва, Россия

Аннотация. Для описания формы хроматографического пика используют различные математические функции, часть из которых основаны на функции нормального распределения: экспоненциально-модифицированная гауссиана, полиномиально-модифицированная гауссиана, гауссиана с параболической дисперсией и другие. Эти функции задаются 4-9 параметрами, которые подбираются итерационно в процессе нелинейной оптимизации. В случае низкой вычислительной устойчивости (характерной для большинства таких функций) выбор начального приближения существенно влияет на сходимость и возможность получения адекватного результата при проведении аппроксимации. Известные подходы к нахождению начальных приближений основаны на использовании эмпирических уравнений, связывающих параметры функций с базовыми параметрами хроматографического пика (коэффициент асимметрии, ширина пика на 10% высоты). При этом описанию алгоритмов расчета этих базовых параметров уделяется недостаточно внимания.

Для моделирования формы пика использовали экспоненциально-модифицированную гауссиану из-за ее широкого применения в хроматографии. В работе рассмотрели реализации этой математической функции, доступные в популярных библиотеках языка Python. Обнаружено, что реализация функции из библиотеки SciPy может обладать численной нестабильностью в случае симметричных пиков. Показано, что подход к расчету экспоненциально-модифицированной гауссианы, предложенный Каламбетом и основанный на использовании нескольких уравнений в зависимости от формы пика, лишен этих недостатков.

Рассмотрены подходы к расчету базовых параметров формы для дискретных пиков. Предложены и описаны алгоритмы для расчета координат точки максимума пика, левой и правой полуширины и ширины пика на заданной высоте (от 10% до 50% от максимума пика).

Известно, что связь базовых параметров пика с параметрами экспоненциально-модифицированной гауссианы является нелинейной. Эмпирические уравнения, описывающие эти связи были предложены Фоли и Дорси в 1980-х. Для описания этих зависимостей нами предложено использовать интерполяцию сплайнами. Показано, что такой подход позволяет точнее описать зависимости и расширить диапазоны допустимых значений параметров. Расчеты сплайна можно провести единожды и сохранить опорные точки сплайна и коэффициенты для дальнейшего использования.

В большинстве работ и методических документах ширина и полуширины хроматографического пика рассчитываются на высоте 10%. Нами рассмотрены альтернативные значения в диапазоне от 10% до 50% высоты пика. Показано, что для пиков без шума расчет ширины и полуширин можно проводить в относительно широком диапазоне высот (от 10 до 30%) без существенного влияния на рассчитанные из них значения параметров экспоненциально-модифицированной гауссианы. Для пиков с отношением сигнал/шум 100 можно использовать 30-35% высоты пика как альтернативный вариант для расчетов.

Ключевые слова: форма хроматографического пика, экспоненциально модифицированная гауссиана, ЭМГ, газовая хроматография, ГХ, жидкостная хроматография, ВЭЖХ.

Благодарности: работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации в рамках госбюджетной темы № 124041900012-4. Работа выполнена с использованием оборудования, приобретенного за счет средств Программы развития Московского университета.

Для цитирования: Хрисанфов М.Д., Самохин А.С. Выбор начальных значений для итерационной аппроксимации хроматографических пиков экспоненциально модифицированной гауссианой // *Сорбционные и хроматографические процессы*. 2024. Т. 24, № 6. С. 885-895. https://doi.org/10.17308/sorpchrom.2024.24/12566

Introduction

Peak shapes in chromatography are a complex result of different physical and physico-chemical processes within the column. Some of these effects are well-studied and can be accounted for and others are more random and specific to a certain column or detector. That is why it is considered impossible to introduce an ideal analytical function to describe peak profiles obtained experimentally [1]. Gaussian function is the simplest one and it has been used to describe chromatographic peaks for decades. Gaussian function is symmetrical which significantly simplifies calculations, but gives only rough approximation because most chromatographic peaks are either tailing or fronting. Exponentially modified gaussian (EMG) function is asymmetrical and therefore it is better suited to real practice. It was first suggested in 1972 [2] and has been widely used in chromatography since. Afterward, many other functions have been suggested. Most of them such as polynomially modified Gaussian (PMG) [3], parabolic variance Gaussian (PVG) [4] and parabolic-Lorentzian-Gaussian (PLG) [5,6] are based on the Gaussian and some like Li [7,8] and Pap-Papai [9] are not. These more modern functions are not as widely used as EMG due to several reasons. Firstly, there are plenty of these functions and it is not easy to choose one when EMG is a *de facto* standard in chromatography. Secondly, there are almost no implementations of these functions available in popular libraries or software. And finally, most of these functions have more parameters which results in additional complexity.

There are plenty of articles describing approximation functions for chromatographic and other peaks. Front and back half widths at a particular height and apex coordinates are usually used to get initial guesses for further nonlinear optimization or to get final fit in earlier works [3,6,10]. However, there is a lack of materials explaining how these basic parameters of a chromatographic peak can be estimated. In the original article [11] Foley and Dorsey suggested estimating parameters of EMG (σ_G τ, and μ) from basic peak shape parameters (A_{10}, B_{10}, W_{10}) using empirical equations. The original approach was based on graphical measurements of basic peak parameters; it looks outdated nowadays. The authors of a more recent paper [12] suggested using linearly modified Gaussian as a model function for characterizing the shape of chromatographic peaks. Empirical equations based on basic peak shape parameters were used to estimate initial parameters similarly to the approach by Foley and Dorsey. In papers on peak shapes in chromatography width and half widths are often calculated at 10% of the peak height; it is often chosen by default because it has been used so widely since the 1980s that it became a *de facto* standard value. Basic algorithms, e.g. calculating width at certain height or finding the coordinates of the apex, are often omitted or referred to as software functions or common routines. While common chromatography processing suites are able to do these things automatically, we found a detailed explanation of the algorithms used only for software from Agilent [13]. In this work we try to give these basic operations the attention they deserve while suggesting some solutions to the problems that arise along the way.

In general, to fit real data with EMG, PMG, PVG, and all other functions mentioned in the previous paragraph a nonlinear optimization is required. In this regard, selection of good initial parameters is important because it ensures and speeds up the convergence of iterative fitting. EMG has only four parameters. Moreover, currently proposed equations for EMG calculation are numerically stable [14]. Therefore, even naive and straightforward approaches work well in most cases even when initial guesses are suboptimal. However, as was mentioned earlier, more parameters in fitting functions (e.g., PMG, PVG and PLG) and worse numerical stability increase the chance of failure during nonlinear optimization. There are two approaches to resolve the convergence issue: improving the initial guess or tightening tolerances (lower and upper bounds) for optimized parameters.

This work is focused on exploring relations between the EMG parameters and basic peak shape parameters and applying these relations to model data. Only a few experimental peaks were fitted to demonstrate the approach in order to avoid discussing more complex related problems such as subtracting the baseline and defining peak boundaries.

Experimental

Software and libraries. Python 3.12 and the following libraries were used to carry out

Table 1. Basic parameters of a chromatographic peak.
Таблица 1. Параметры хроматографического пика

Description	Designation/Symbol
left halfwidth at xx% height	A_{xx} (A_{10} at 10%)
right halfwidth at xx% height	B_{xx} (B_{10} at 10%)
width at xx% height	W_{xx} (W_{10} at 10%)
time coordinate of the apex	t_R or *apex.X*
intensity coordinate of the apex	*apex.Y*
asymmetry factor at xx% height	$f_{asym,xx} = \frac{B_{xx}}{A_{xx}}$
tailing factor at xx% height	$T_{xx} = \frac{A_{xx} + B_{xx}}{2A_{xx}}$
asymmetry ratio at xx% height	$r_{asym,xx} = \frac{B_{xx} - A_{xx}}{A_{xx} + B_{xx}}$

the research: NumPy (array routines, math functions) [15], SciPy (special functions, statistical distributions, curve fitting) [16], Pandas (import/export of csv files, data manipulation routines) [17], Matplotlib (plotting) [18], and Seaborn (plotting) [19]. All source code and interactive Jupyter Notebooks [20] are available on GitHub [21].

Basic peak parameters and EMG parameters. We chose EMG as the model function to calculate all the relations because it is the most studied and widely used function to describe chromatographic peaks. Another important factor was that, while EMG may be not the most accurate compared to some modern functions, it has only four parameters with clearly defined physical meaning [11, 14]:

$$EMG(x) = h * \frac{\sigma_G}{\tau} * \sqrt{\frac{\pi}{2}} * e^{\left(\frac{\sigma_G{}^2}{2\tau^2} - \frac{x-\mu}{\tau}\right)} * erfc\left(\frac{1}{\sqrt{2}} * \left(\frac{\sigma_G}{\tau} - \frac{x-\mu}{\sigma_G}\right)\right) \quad (1)$$

Where the EMG parameters are height (h), scale (σ_G), mean (μ or t_G), exponential decay (τ). These parameters correspond to key aspects of chromatographic peaks: analyte concentration, peak width, retention time and peak asymmetry, respectively. In addition, two derived parameters are also often used: the shape parameter (also known as the fundamental ratio [11]) $K = \frac{\tau}{\sigma}$ and the standard deviation $\sigma_{EMG} = \sqrt{\tau^2 + \sigma_G^2}$.

Basic parameters of a chromatographic peak extracted from raw data are presented in Table 1. There are some alternative suggestions [22,23] to calculate asymmetry value as a ratio of the areas of left and right parts of a chromatographic peak. Nevertheless, we did not use this approach and preferred to calculate the standard asymmetry factor of a peak.

There are several implementations of EMG distribution function in Python libraries (Tensorflow: tfp.distributions.ExponentiallyModifiedGaussian [24], SciPy: sp.stats.exponnorm [16]). SciPy implementation showed some instability for $K<10^{-4}$, see Results and Discussion for more information. We also implemented a version of the computationally stable EMG function by Kalambet et al. [14] and used it in this work.

Relations between EMG parameters and basic peak shape parameters. Full width of the EMG peak was defined as 6*σ_{EMG}, similarly to the 6*σ_G Gaussian full peak width (99.73% of the whole area). In this work there were two types of peaks used: (i) theoretical (continuous) peaks with unlimited number of points, as all EMG parameters for these peaks are known, and therefore *EMG(x)* can be calculated for any *x*, (ii) discrete peaks with 10 and 100 points per full width with and without noise. To model noisy data, normally distributed random values (mean=0, std=S/N) were added to model data. The signal-to-noise ratio (S/N) was set at 100 for this work.

The following basic peak shape parameters were calculated: apex coordinates (*apex.X* and *apex.Y*), left and right halfwidths (A_{10}, B_{10}) and width (W_{10}) at 10% of the peak height, and asymmetry-related values (f_{asym}, T, r_{asym}).

Two parameters were varied in a mesh grid pattern, each of the parameter's intervals was split into 100 points using geometric progression: σ – [0.001;10], K – [0.0001;10]. The relations were initially tested to work for K up to 30 and σ up to 25. Later choosing such extreme ranges was deemed unnecessary because quite more moderate values of σ and K values are usual for gas and liquid chromatography.

Calculation of basic peak parameters for continuous EMG function. Firstly, the coordinates of apex point (*apex.X* and *apex.Y*) were calculated using a default minimization routine from SciPy applied to $y(x) = -1 * EMG(x)$ function. Secondly, the peak was split into left and right halves. For each of the halves the X coordinates corresponding to 10% peak height (*left₁₀.X* and *right₁₀.X*) were calculated by applying root finding routine (scipy.optimize.brentq) to $y(x) = EMG(x) - 0.1 * apex.Y$ function. Halfwidths were calculated the following way: $A_{10} = apex.X - left_{10}.X$, $B_{10} = right_{10}.X - apex.X$, $W_{10} = right_{10}.X - left_{10}.X$. Asymmetry ratio and tailing factor were then calculated using their definitions.

Approaches to calculation of basic peak parameters for the discrete peaks. Due to the discrete nature of experimental data, the simplest approach is to consider only discrete X coordinates corresponding to available nodes. For example, *apex.X* and *apex.Y* can be found just as coordinates of the node having the highest Y coordinate. It is a straightforward approach which is implemented in many software products. Nevertheless, there are some limitations: precision is limited to the size of the quantization step (i.e., distance between nodes) and presence of noise can lead to unexpected results (e.g., a chromatographic peak can have a few local maxima because of high noise level). To estimate peak shape parameters with precision better than quantization step one must perform approximation or interpolation [25].

Calculation of apex coordinates for the discrete peaks. To approximate the top part of a peak (*apex.X* and *apex.Y* coordinates) the following algorithm was used: firstly, a data point with the maximum height y_{max} was selected. Secondly, all data points with height more than $0.85*y_{max}$ were selected. This selection had 3 or more points only for peaks with more than 50 points per full width which is not always the case in gas chromatography. For peaks with less than 3 data points selected, the region was expanded by adding one point to the left and one to the right. Finally, parabola was used to approximate the top of the peak. It was an important step because it allowed for more precise estimation of the *apex.X* (time) coordinate of the misshapen peak. For example, even when the top of the experimental peak was split because of noise or the top of the peak was misshapen due to low number of points, etc.

Calculation of halfwidths for the discrete peaks. The following algorithm was used to calculate halfwidths of a chromatographic peak for each of the sides at a given height (10%**apex.Y* is the default): if 3 or 2 points were present on one side (from the first datapoint to the apex) they were interpolated with parabola or line respectively. If there were more than 3 experimental points on one side of the peak then the points in the range [height – 10%**apex.Y*; height + 10%**apex.Y*] were chosen. The lower bound could not be less than 5%**apex.Y*. For example, for 10% of the peak height the interval would be [5%**apex.Y*; 20%**apex.Y*] and for 30% it would be [20%**apex.Y*; 40%**apex.Y*]. If there were less than 3 points in the selected range then the range was expanded stepwise towards the *apex.X* and away from it until there were 3 points se-

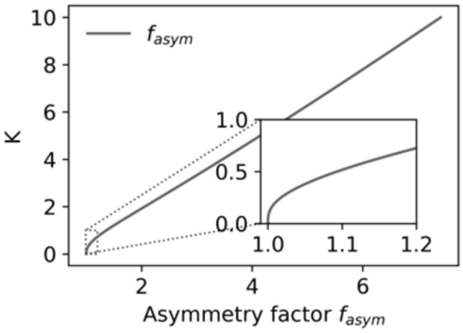

Fig. 1. *K* vs $f_{asym,10}$ relation.
Рис. 1. Зависимость *K* от $f_{asym,10}$

lected. The selected points were approximated with a parabola. Approximation was necessary to precisely calculate the X coordinate at a certain height and also helped to reduce the effects of noise. While more complex functions could be used, parabola was sufficient to describe the front and tail of a peak and was more robust because it is linear with respect to the coefficients.

Interpolation of relations between basic peak parameters and EMG parameters. Continuous peaks were used to calculate some relations between EMG parameters and basic peak shape parameters: f_{asym} vs *K*, σ/W_{10} vs f_{asym}, *(apex.X* - μ)/σ vs f_{asym}. Neither of these relations were fully linear or could be precisely described with low-degree polynomials in a wide range of values. Therefore splines (sp.intepolate.Akima1DInterpolator) were used to get some parameters of EMG function.

GC/MS analysis. GC/MS analysis was performed using an Agilent 7890A gas chromatograph (Agilent) coupled with a Pegasus HT mass spectrometer (LECO). Separation of the model mixture of organic compounds (containing 5-methyl-2-hexanone and cyclohexanone) was carried out on a Varian VF-5ms column (30 m×0.25 mm×0.25 um) in isothermal mode at 40°C.

Results and discussion

Choosing the best EMG implementation. The first step for estimation of EMG parameters from basic peak shape parameters is to study relations between the former and the latter. Continuous EMG peaks calculated for a range of shape parameters (*K*) were used for the task. We found out that SciPy implementation of EMG function is numerically unstable when shape parameter *K* is close to 0 ($K<10^{-4}$) and the peak is mostly Gaussian. Fitting chromatographic peaks is not a common task for users of SciPy library; we did not come across any efforts to improve stability for extreme conditions. EMG profiles calculated using SciPy and Kalambet's implementations are shown in Fig. S1. One can see that in the case of the SciPy library, profiles are not smooth because of numerical instability. The main underlying reason for this instability seemed to be a numerical overflow in an exponential part of the EMG implementation in SciPy. At the same time, our implementation based on Kalambet's approach [14] had only some minor irregularities for $K<10^{-4}$ region but they are quite a bit less severe and did not result in any problems in our calculations. Therefore, it was chosen for this work.

K vs *T*/f_{asym} relation. The shape parameter *K* is used in SciPy to define the shape of the EMG peak. It is the most important parameter of the EMG function, because changing all other parameters can be considered as linear transformations of the original profile: scaling (when height or width is changed) and translation (when peak *apex.X* is changed). Other EMG parameters are easier to calculate when *K* is known. *K* is proportional to both asymmetry factor [26] and tailing factor; however, these relations are not linear (see Fig.1). Both curves are monotonously increasing and have two parts: nonlinear profile (for *K* from 0 to around 1) and mostly linear one (at higher *K* values). At the

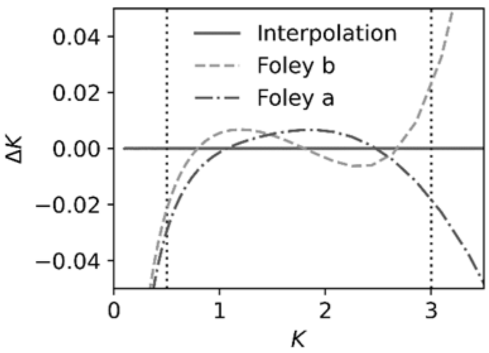

Fig. 2. Errors of K estimated using interpolation with splines and empirical equations proposed by Foley et al. [11] (equations 5a and 5b in the original article, equations 2 and 3 in this article).
Рис. 2. Погрешности оценки K с использованием сплайнов и эмпирических уравнений, предложенных Foley et al. [11] (уравнения 5a и 5b в оригинальной статье, уравнения 2 и 3 в этой статье).

same time, r_{asym} vs K relation has an S-like curve, that is why it is not used in this work.

Comparison with Foley approximation functions. Approximation functions suggested by Foley et al. (5a and 5b) [11] are sufficiently accurate for $f_{asym,10}$ in [1.09; 2.76] (K in [0.49; 2.99]), however, this range is quite narrow and restricted. For example, it does not include symmetrical peaks and highly distorted peaks with $f_{asym,10}$ ≥3which can be sometimes observed when eluting conditions are suboptimal. Foley et al. proposed the following empirical equations:

$$K = \frac{\tau}{\sigma_G} = \sqrt{\frac{W^2{}_{10}}{1.764*f^2{}_{asym,10} - 11.15*f_{asym,10} + 28} - \sigma^2{}_G} \Big/ \frac{W_{10}}{3.27*f_{asym,10}+1.2} \quad (2)$$

$$K = \frac{\tau}{\sigma_G} = \sqrt{\frac{t^2{}_R*(f_{asym,10}+1.25)}{41.7*(t_R/W_{10})^2} - \sigma^2{}_G} \Big/ \frac{W_{10}}{3.27*f_{asym,10}+1.2} \quad (3)$$

It is possible to both increase accuracy of estimation and extend the usable range of f_{asym} inputs by interpolation with splines. As can be seen from Fig. 2, even inside the region proposed by Foley and Dorsey, interpolation with splines shows much smaller error. It requires calculation of relations between parameters of EMG and basic parameters in a wider range of asymmetry values to avoid extrapolation. However, in ideal conditions interpolation with splines can be inherently more precise than approximation with low-degree polynomials due to the high precision of calculated data points.

Akima smooth continuously differentiable cubic spline interpolation was used in this work (scipy.interpolate.Akima1DInterpolator). It contained 200 knots for $f_{asym,10}$ ranging from 1.0 to 7.4 (K from 0.0 to 10.0). The end number for $f_{asym,10}$ was chosen to suit $K=10$ and the range can be further extended up to $f_{asym,10}$=20 (K=30) without any problem. However, such an expansion seems to be counterproductive as peaks with such extreme tailing are rare.

σ/W_{10} vs K relation. The EMG parameter K describes the overall shape of the peak. The width of the peak at 10% of its height (W_{10}) is proportional to the σ parameter of the EMG function. Therefore, a σ/W_{10} vs K relation can be used to estimate the σ EMG parameter when W_{10} and K have been already calculated. The relation can be simplified by omitting conversion from f_{asym} to K and using σ/W_{10} vs f_{asym} relation instead (see Fig.3).

Δμ/σ vs $f_{asym,10}$ relation. Distance between the *apex.X* coordinate and the μ parameter can also be estimated when the σ and f_{asym} parameters are known. This relation is represented by the following equation [11], where t_G is μ and t_R is *apex.X* coordinate and the function $f(B_{10}/A_{10})$ is a parabola:

$$t_G = t_R - \sigma_G * f(B_{10}/A_{10}); \quad (4)$$

$$\Delta\mu = t_R - t_G = \sigma * f(f_{asym}) \quad (5)$$

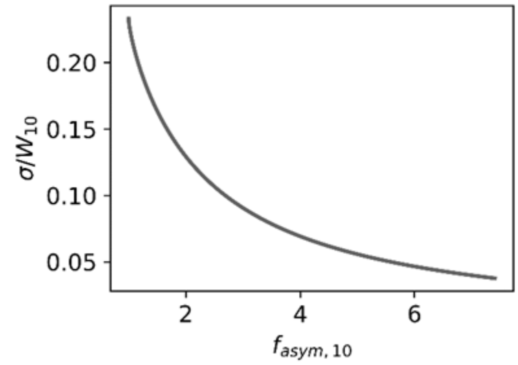

Fig. 3. σ/W_{10} vs $f_{asym,10}$ relation
Рис. 3. Зависимость σ/W_{10} от $f_{asym,10}$

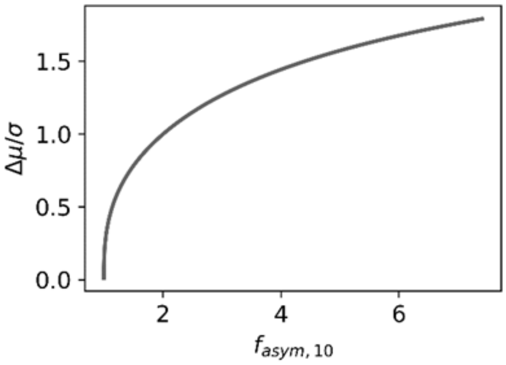

Fig. 4. Δμ/σ vs $f_{asym,10}$ relation
Рис. 4. Зависимость Δμ/σ от $f_{asym,10}$

While the $\sigma_G * f(B_{10}/A_{10})$ part can be transformed either into the function of *K*, it unnecessarily complicates the calculations by adding an additional step. Therefore, the Δμ/σ vs $f_{asym,}$ (see Fig.4) was the relation interpolated by splines (where Δμ=*apex.X* - μ).

Estimating the height. It is known that an apex of the EMG lies on the unmodified Gaussian function [14] (as exponential component equals 1 in this point). Therefore, the *h* parameter of the EMG can be calculated from the *apex.X* coordinate and the equation of the unmodified Gaussian.

Approximation of a model peak. The first step to apply the described approach for peak fitting is to obtain basic parameters of a chromatographic peak. It is supposed that the peak in question is tailing. If the peak is fronting, it can be mirrored with the following equation $x_{new} = 2\mu - x_{old}$.

Final pipeline looks as follows. A_{xx}, B_{xx}, W_{xx}, $f_{asym,xx}$ and peaks apex point coordinates (*apex.X*, *apex.Y*) are calculated. Then precalculated splines are used to estimate all other parameters: *K* is estimated from f_{asym} via a spline, σ is estimated from W_{xx} and $f_{asym,xx}$ via a spline, τ is calculated from *K* and σ by definition, *h* is calculated from unmodified Gaussian and *apex.X*, *apex.Y* coordinates, μ is calculated from *apex.X* coordinate, σ, and $f_{asym,xx}$.

Choosing height to calculate width and halfwidths. While calculating width at 10% of the peak height is the most common approach and some arguments provided in the original paper hold mostly true [11], there is no prohibition to use other values. Going lower is not an option in most cases due to prevalence of noise, which leads to poor fitting results. Choosing a higher point is definitely an option when signal to noise ratio is too low. However, there is a tradeoff: the higher the point to measure the halfwidth – the steeper slope in *K* vs $f_{asym,xx}$ relation and therefore estimation of *K* becomes very sensitive for errors in measuring $f_{asym,xx}$ (Fig. 5a).

Calculation of EMG parameters was tested at heights from 10% to 50% with a step of 5% for 10 *K* values geometrically spaced from 0.1 to 3. Relative difference $\delta y = \frac{\Sigma(abs(\hat{y}-y))}{\Sigma y}$ was calculated, averaged among all K values and used as a metric of goodness of fit, where ŷ are Y coordinates calculated from the initial estimate for *K*, σ, τ, *h* and *y* are values of the original continuous EMG function which was used to calculate discrete chromatographic profiles. It was discovered that this metric was comparable for parameters calculated at 10-30% of height for noise free peaks (Fig. 5b, 5c). The results were worse for 40% and higher values. It could be explained by inherent instability caused by a greater slope of $f_{asym,xx}$ vs *K* relation. For noisy data (S/N = 100), the lowest values of δ*y* were observed for about 30-35% of peak height. Therefore, it can be considered as an alternative point to calculate the basic peak shape parameters and estimate the EMG parameters.

Approximation of experimental GC/MS data. Two compounds were selected to

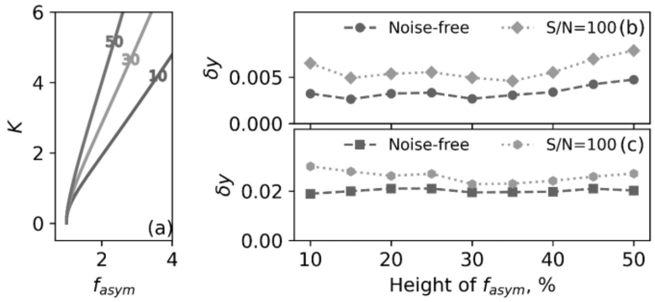

Fig. 5. *K* vs $f_{asym,xx}$ relation (a) for asymmetry factor calculated at different heights (10-50%). Relative difference δy between original EMG profiles and profiles calculated from initial estimates obtained from width and halfwidths corresponding to different heights (b, c). The lower δy, the better initial estimates. Discrete peaks with 100 (b) and 10 (c) points per full width were used. Both noise-free data and peaks with signal-to-noise 100 were considered.

Рис. 5. Зависимость *K* от $f_{asym,xx}$ (a) для фактора асимметрии, рассчитанного на разных значениях высоты пика (10-50%). Относительная разность δy (b, c) между исходной кривой ЭМГ и кривыми, рассчитанными из начальных приближений, полученных из ширины и полуширин на разных значениях высоты пика. Чем меньше δy, тем лучше начальные приближения. Рассматривали дискретные пики с 100 (b) и 10 (c) точками на полную ширину пика, содержащие (отношение сигнал-шум – 100) и не содержащие шум.

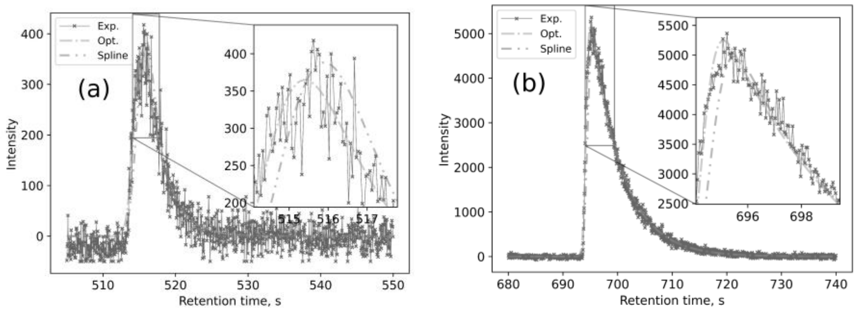

Fig. 6. Extracted ion chromatograms corresponding to the molecular ions of 5-methyl-2-hexanone (a) and cyclohexanone (b). “Exp” represents discrete experimental data points. “Spline” corresponds to the initial values for iterative fitting found using the proposed approach. “Opt” shows the final results obtained by iterative fitting of the experimental data points with the EMG function.

Рис. 6. Хроматограммы по заданным значением m/z, отвечающие молекулярным ионам 5-метил-2-гексанона (a) и циклогексанона (b). Функция ЭМГ, отвечающая начальным значениям параметров, полученных с использованием предложенного подхода, обозначена как "Spline". Функция ЭМГ, полученная после проведения итерационной аппроксимации экспериментальных данных, обозначена как "Opt".

demonstrate the performance of the proposed approach using real GC/MS data. Extracted ion chromatograms corresponding to the molecular ions of 5-methyl-2-hexanone (m/z = 114) and cyclohexanone (m/z = 98) are shown in Fig. 6. In the case of 5-methyl-2-hexanone, the main challenge was the low signal-to-noise ratio, which was only 18. This could create some difficulties in finding the apex and estimating half-widths. The other example (Fig. 6b) displays a highly asymmetrical peak of cyclohexanone ($f_{asym,10}$ ≈ 7 and K ≈ 10). The empirical equations from the article by Foley et al. [11] cannot be applied here as the K value is outside of the supported range (i.e., 0.5 to 3). Despite all these challenges, the initial values for iterative fitting were correctly found using the approach proposed in this work, and the experimental data were correctly described even without performing iterative fitting (Fig. 6).

Conclusions

A spline-based approach for estimating EMG parameters from basic peak shape parameters (A_{xx}, B_{xx}, W_{xx}, *apex.X*, *apex.Y*) was suggested. It is based on finding some relations between EMG parameters and basic peak shape parameters: f_{asym} vs K, σ/W_{xx} vs $f_{asym,xx}$, $(apex.X$ - $\mu)/\sigma$ vs $f_{asym,xx}$. The use of spline interpolation (instead of empirical equations described in the literature) allowed us to expand the range of suitable peak shapes ($f_{asym,10}$ in [1.00; >20]) and increase accuracy. The full algorithm to estimate EMG peak shape from the recorded peak was described in detail, including finding peak *apex.X* and *apex.Y* coordinates along with peak halfwidths at certain height. It is also available as a Jupyter notebook [21]. Additionally, it was shown that for relatively low signal to noise ratios (S/N = 100) it may be beneficial to estimate initial EMG parameters from width and halfwidths calculated at greater peak heights (30-35% of peak height).

Data availablity

Source code for the algorithms and figures for the article are available in our GitHub repository at https://github.com/mkhrisanfov/peak-shape-calculation

Conflict of interest

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

References

1. Di Marco V.B., Bombi G.G. Mathematical functions for the representation of chromatographic peaks. *Journal of Chromatography A*. 2001; 931(1-2): 1-30.

2. Grushka Eli. Characterization of exponentially modified Gaussian peaks in chromatography. *Anal. Chem. American Chemical Society*, 1972; 44(11): 1733-1738.

3. Torres-Lapasió J.R., Baeza-Baeza J.J., García-Alvarez-Coque M.C. A Model for the Description, Simulation, and Deconvolution of Skewed Chromatographic Peaks. *Anal. Chem*. 1997; 69(18): 3822-3831.

4. Baeza-Baeza J.J., García-Alvarez-Coque M.C. Prediction of peak shape as a function of retention in reversed-phase liquid chromatography. *Journal of Chromatography A*. 2004; 1022(1-2): 17-24.

5. Caballero R.D., García-Alvarez-Coque M.C., Baeza-Baeza J.J. Parabolic-Lorentzian modified Gaussian model for describing and deconvolving chromatographic peaks. *Journal of Chromatography A*. 2002; 954(1-2): 59-76.

6. Baeza-Baeza J.J., Ortiz-Bolsico C., García-Álvarez-Coque M.C. New approaches based on modified Gaussian models for the prediction of chromatographic peaks. *Analytica Chimica Acta*. 2013; 758: 36-44.

7. Li J. Development and Evaluation of Flexible Empirical Peak Functions for Processing Chromatographic Peaks. *Anal. Chem*. 1997; 69(21): 4452-4462.

8. Li J. Comparison of the capability of peak functions in describing real chromatographic peaks. *Journal of Chromatography A*. 2002; 952(1-2): 63-70.

9. Pap T.L., Pápai Zs. Application of a new mathematical function for describing chromatographic peaks. *Journal of Chromatography A*. 2001; 930(1): 53-60.

10. Pápai Z., L. Pap T. Determination of chromatographic peak parameters by non-linear curve fitting using statistical moments. *Analyst. Royal Society of Chemistry*, 2002; 127(4): 494-498.

11. Foley J.P., Dorsey J.G. Equations for calculation of chromatographic figures of merit for ideal and skewed peaks. *Anal. Chem*. 1983; 55(4): 730-737.

12. Baeza-Baeza J.J., García-Alvarez-Coque M.C. Characterization of chromatographic peaks using the linearly modified Gaussian model. Comparison with the bi-

Gaussian and the Foley and Dorsey approaches. *Journal of Chromatography A.* 2017; 1515: 129-137.

13. Evaluating System Suitability - CE, GC, LC and A/D ChemStation - Revisions: A.03.0x-->A.08.0x.

14. Kalambet Y, Kozmin Y., Mikhailova K., Nagaev I., Tikhonov P. Reconstruction of chromatographic peaks using the exponentially modified Gaussian function. *J. Chemometrics.* 2011; 25(7): 352-356.

15. Harris C.R., Harris C.R., Millman K.J., Van Der Walt S.J., Gommers R., Virtanen P., Cournapeau D., Wieser E., Taylor J., Berg S., Smith N.J., Kern R. Array programming with NumPy. *Nature. Nature Publishing Group,* 2020.; 585(7825): 357-362.

16. Virtanen P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods. Nature Publishing Group,* 2020; 17(3): 261-272.

17. The pandas development team. pandas-dev/pandas: Pandas. Zenodo, 2024.

18. Hunter J.D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering.* 2007; 9(3): 90-95.

19. Waskom M.L. seaborn: statistical data visualization. *Journal of Open Source Software.* 2021; 6(60): 3021.

20. Granger B.E., Pérez F. Jupyter: Thinking and Storytelling With Code and Data. *Computing in Science & Engineering.* 2021; 23(2): 7-14.

21. mkhrisanfov/peak-shape-calculation [Electronic resource]. URL: https://github.com/mkhrisanfov/peak-shape-calculation (accessed: 14.05.2024).

22. Zenkevich I.G., Makarov A.A., Pavlovskii A.A. New approaches to the calculation and interpretation of asymmetry factors of chromatographic peaks. *J Anal Chem.* 2017; 72(7): 710-718.

23. Mallard W.G., Reed J. AMDIS – USER GUIDE. National Institute of Standards and Technologies, 2019.

24. TensorFlow Developers. TensorFlow. Zenodo, 2024.

25. Samokhin A.S., Kalambet Yu.A. Opredelenie parametrov funkcii razvertki kvadrupol"nogo mass-spektrometra: 2. *Analitika i kontrol'.* 2018; 22(2): 168-176. (In Russ.)

26. Barber W.E., Carr P.W. Graphical method for obtaining retention time and number of theoretical plates from tailed chromatographic peaks. *Anal. Chem.* 1981; 53(12): 1939-1942.

Информация об авторах / Information about the authors

М.Д. Хрисанфов – аспирант кафедры аналитической химии химического факультета Московского государственного университета им. М.В. Ломоносова, м.н.с лаборатории “умных” методов химического анализа Института физической химии и электрохимии им. А.Н. Фрумкина РАН, Москва, Россия

А.С. Самохин – к.х.н, м.н.с. лаборатории масс-спектрометрии химического факультета МГУ им. М.В. Ломоносова, м.н.с лаборатории “умных” методов химического анализа Института физической химии и электрохимии им. А.Н. Фрумкина РАН, Москва, Россия

M.D. Khrisanfov – PhD student at analytical chemistry division, chemistry department, Lomonosov Moscow State University, junior researcher at laboratory of “smart” methods of chemical analysis, Institute of Physical chemistry and electrochemistry, Moscow, Russian Federation, e-mail: khrisanfovmike@gmail.com

A.S. Samokhin – PhD, junior researcher at laboratory of mass-spectrometry, chemistry department, Lomonosov Moscow State University, junior researcher at laboratory of “smart” methods of chemical analysis, Institute of Physical chemistry and electrochemistry, Moscow, Russian Federation

Статья поступила в редакцию 20.08.2024; одобрена после рецензирования 26.09.2024; принята к публикации 16.10.2024.
The article was submitted 20.08.2024; approved after reviewing 26.09.2024; accepted for publication 16.10.2024.