Title: A unique data analysis framework and open source benchmark data set for the analysis of comprehensive two-dimensional gas chromatography software.

Authors: Benedikt A. Weggler[1,#,*], Lena M. Dubois[1*], Nadine Gawlitta[2], Thomas Gröger[2], John Moncur[3], Luigi Mondello[5], Steven Reichenbach[4], Peter Tranchida[5], Zhijun Zhao[6], Ralf Zimmermann[2], Mariosimone Zoccali[7], Jean-François Focant[1]

#corresponding author, benedikt.weggler@googlemail.com

*shared first authorship

Affiliations:

1) University of Liège, MolSys - Organic and Biological Analytical Chemistry Group, Quartier Agora, Place du Six Août 11, 4000 Liège, Belgium
2) Joint Mass Spectrometry Centre, Helmholtz Zentrum München and the University of Rostock, Ingolstädter Landstr. 1, Neuherberg, 85764, Germany
3) SpectralWorks Limited, The Heath Business and Technical Park, Runcorn, Cheshire, WA7 4QX, United Kingdom
4) University of Nebraska-Lincoln, Lincoln, NE, USA
5) Dipartimento di Scienze Chimiche, Biologiche, Farmaceutiche ed Ambientali, University of Messina, Polo Annunziata, 98168 Messina, Italy
6) J&X Technologies, Shanghai, China
7) Department of Mathematical and Computer Science, Physical Sciences and Earth Sciences, University of Messina, Messina, Italy

ABSTRACT:

Comprehensive two-dimensional gas chromatography (GC×GC) is amongst the most powerful separation technologies currently existing. Since its advent in early 1990, it has become an established method which is readily available. However, one of its most challenging aspects, especially in hyphenation with mass spectrometry is the high amount of chemical information it provides for each measurement. The GC×GC community agrees that there, the highest demand for action is found. In response, the number of software packages allowing for in-depth data processing of GC×GC data has risen over the last couple of years. These packages provide sophisticated tools and algorithms allowing for more streamlined data evaluation. However, these tools/algorithms and their respective specific functionalities differ drastically within the available software packages and might result in various levels of findings if not appropriately implemented by the end users.

This study focuses on two main objectives. First, to propose a data analysis framework and second to propose an open-source dataset for benchmarking software options and their specificities. Thus, allowing for an unanimous and comprehensive evaluation of GC×GC software. Thereby, the benchmark data includes a set of standard compound measurements and a set of chocolate aroma profiles. On this foundation, eight readily available GC×GC software packages were anonymously investigated for fundamental and advanced functionalities such as retention and detection device derived parameters, revealing differences in the determination of e.g. retention times and mass spectra.

Keywords: Data Processing; Open Source Data; Chocolate, Fragrances and Allergens

1 INTRODUCTION:

There are various factors to consider when determining the best solution for analytical challenges. Besides the fundamental question (targeted or untargeted analysis) the complexity of the sample matrix holds a significant influence in determining the suitable analytical technique [1,2]. In brief, high sample complexity requires high specificity and sensitivity of the utilized technique. Therefore, separation techniques with high chromaptographic resolution have become increasingly popular over the last decade. These techniques allow for enhanced separation and therefore for the evaluation of the total burden of (specific/target) analytes in a single chromatographic analysis, as well as the identification of unexpected and unknown compounds especially when hyphenated to mass spectrometry (MS).

In particular, comprehensive two-dimensional gas chromatography (GC×GC) has become popular in the field of separation science. Due to its versatility, it includes applications related to forensic, life-/medical, environmental and/or petro-sciences [3–11]. Consequently, the number of reports and applications utilizing GC×GC has increased as illustrated in Figure S1 in the supporting information (SI). Increased separation capability, however, does not necessarily solve the general challenge in chromatography, namely coelution, or facilitate the extraction of meaningful chemical information. In fact, it demands fast detector acquisition techniques resulting in information rich data sets with higher order complexity and file size [12] especially when coupled to sophisticated detection techniques. Evaluation of these datasets is considered a major challenge in GC×GC and the community agrees that growth, development and a certain degree of automation is needed [4,6,13–17] in this particular area. It is thereby little surprising that the variety and availability of dedicated GC×GC software packages (SPs) rose within the last years.

During the analytical workflow, the general questions (targeted, untargeted, quantitative or qualitative analysis) dictate the parameters to be extracted from such datasets and the demands of functionality of the applied SP. Particular consideration is thereby placed on retention and/or detection-device derived values (e.g. retention time/s, concentration values, calibration correlations, matching factors etc.), which in case of GC×GC undergo statistical pre-treatment and chemometrical analysis. Additionally, each research team has its individual data analysis workflow adding to the complexity of the overall evaluation procedure.

The high chromatographic resolution and the advances in detection technology make GC×GC well suited for highly powered data-driven evaluation procedures enabling e.g. signal deconvolution and picture/peak list based chemometrical analysis. Noteworthy is thereby the potential of GC×GC data allowing differentiation between chemical sample types or classes for which prior alignment procedures is necessary. Accordingly, the determination of standard parameters derived from the separation (retention times and retention indices) or from the detection device (mass spectra or quantitative information), need to be as accurate and reliable as possible [18]. Such demands, the increased complexity and amount of data do not only require reliable extraction and interpretation of chemical variation [18], they also lead to increased dependence of analytical chemists on software tools such as peak alignment features based on retention or (if available) mass spectral parameters, compensating for retention time variation, e.g. due to injection over a long period of time [19–21]

Thus, systematic assessment of these SPs is required as illustrated by Koh et al. [22]and Niu et al. [23]. In these studies, specific datasets were created for performance assessment of one-dimensional (1D) GC alignment tools. Although these studies limit themselves to 1D-GC

alignment, they illustrate the importance on a reliable data analysis framework and benchmark data for systematic evaluation of different software packages using a reliable analytical data workflow.

In this study, the authors introduce a data analysis framework to address a characterized and readily available multi-purpose open-source data set (benchmark data) allowing for an objective performance assessment of eight readily available software packages able to handle GC×GC data. The standardized evaluation and comparative analysis of fundamental and advanced functionality demonstrate transparent, systematic and standardized benchmarking practices that can be easily adapted to individual demands. In addition, the well-defined and characterized benchmark datasets could prove highly valuable for didactic purposes such as e.g. introducing novices to general data processing strategies for higher dimensional data.

In summary, the present study focuses on two main objectives: at first, the introduction of a systematic data analysis frame work and, at second the establishment of an open-source dataset for benchmarking purposes of GC×GC software.

2 MATERIAL AND METHODS

The presented benchmarking dataset comprises two parts: firstly, a set of standard compounds (fragrances and allergens) measured at different concentration levels (standard set); secondly, a set of nine different chocolate aroma profiles (chocolate set). Thereby, the dataset contains enough features to be relevant and representative whilst the complexity is low enough allowing for reusability and reproducibility. The datasets are readily available and can be downloaded and referenced at the Harvard Dataverse repository (accessible using the DOIs listed in reference [24,25]). The data is available as analytical data interchange format in compliance with the ASTM E1947 standard (netCDF) [26]. Additionally, information such as composition of the chocolates is documented and available as well. Figure 1 shows GC×GC measurements contained in the standard and chocolate data set whereas Figure S2 in the supporting information shows a more detailed breakdown of the datasets structure.

2.1 STANDARD DATA SET

Fragrance and Allergen Standard (Restek, Bellfonte, USA) was prepared and diluted using methyl tert-butyl ether to concentration levels of 2, 1, 0.4 and 0.2 ppb in 20mL Headspace vials. Subsequently followed by the addition of 1 µL (20 µg/mL) 1-fluoronaphtalene (Restek, Bellefonte, USA) and 0.3 µL (50 µg/mL) of retention standard mixture (Restek, Bellfonte, USA). In total, the mix contained 36 fragrance and allergen compounds. Refer to table TS1 in the supporting information for more details. Each level was measured in triplicate and in block randomized fashion.

2.2 CHOCOLATE DATA SET

Sample preparation of the chocolate samples is described elsewhere [27]. In brief, dark chocolate bars with nine different types of flavor were purchased at a local chocolate factory. 37.5 g of each filling type was cryogenically homogenized followed by subsequent division into 5 g aliquots. Accordingly, these aliquots were placed in 15 mL Falcon tubes and mixed with a total of 5 mL milliQ water and 2.5 mL HCl (2M). After vigorously mixing (2 min), the sample aliquots were allowed to sit in an 80 °C water bath for 15 minutes. Once again, the samples were mixed (2 min) followed by centrifugation (5 min/3000 rpm), which leads to a separation into 4 distinct phases (solid, aqueous, solidified fat and oil). 3 mL of the aqueous and 200 µl of the respective oily phase were placed upon 2.9 g of sodium bicarbonate and 1.2 mL of NaOH (2M) in a 20 mL headspace vial subsequently followed by the addition of 1 µl (20 µg/mL) 1-fluoronaphtalene (Restek, Bellefonte, USA) and 0.3 µl (50 µg/mL) of retention standard (Restek, Bellfonte, USA). Each chocolate type was measured in quadruplicate in block randomized fashion.

2.3 SPME-GC×GC-ToF/MS ANALYSIS

A commercially available 10 mm polydimethylsiloxane/divinylbenzene (PDMS/DVB) (SUPELCO, Darmstadt, Germany) fiber was conditioned prior to the analysis at 250 °C for 15 min. Samples and standards were submitted to 5 min incubation followed by a 40 min extraction procedure at 60 °C sample temperature and 250 rpm agitation speed. Desorption was performed for 2 min in splitless mode at a desorption temperature set to 250 °C which corresponds to a 5 °C lower set point than the maximum recommended coating temperature. Solid phase microextraction (SPME) was automated using an HTA autosampler (HTA, Brescia, Italy).

For this study, a JEOL AccuToF GC+ mass spectrometer (JEOL, Brussels, Belgium) coupled to an Agilent 7890 GC (Agilent, Santa Clara, USA) was used. The GC×GC analysis was carried out using a

30 m Rxi5-ms (0.25 mm/0.25 µm) capillary column (Restek, Bellefonte, PA, USA) as first dimension and a 2 m Rxi17 (0.1 mm/0.1 µm) capillary column (Restek, Bellefonte, PA, USA) in the second dimension. Helium was used as carrier gas with a constant flow rate of 1ml/min. Modulation of the first dimensions effluent was obtained using a solid-state modulator (SSM 1800 – J&X Technologies, Shanghai, China) equipped with a 1.05m GsBP-1 (0.25 mm/0.1 µm) trapping column (GS-Tek, Newark,US).

After 2 min at the oven's starting temperature (70 °C) the oven was ramped at 10 °C/min to its final temperature (310 °C) and held for 10 min. The modulator's temperature settings followed the GC oven with the appropriate modifications according to the manufacturer's instructions for optimal performance (temperature offset for entry and exit zones as well as trap settings). In particular, the exit temperature has been increased preventing the formation of cold spots. Refer to the supporting information (Figure S3-4) for a detailed graphical representation of oven and modulator temperature settings. The modulation time was set to 4 s.

Transfer line and source temperature were set to 250 °C. Ionization was carried out in electron ionization mode at 70 eV with a detector acquisition frequency of 50 Hz (maximal achievable acquisition frequency) over the mass range m/z: 35-500 amU.

2.4 PRE-PROCESSING AND DATA ANALYSIS FRAMEWORK

The centroid data was converted into the net ANDI MS (*.cdf) data file format. In *.cdf files, time and mass/charge information are linearly stored. For the 2D information to be accessible, the linear information has to undergo transformation. For this purpose, the acquisition frequency of the instrument is multiplied with the targeted second dimension time, allowing to recombine the appropriate number of spectra as second dimension recording. Based on the number of spectra

and the pre-set modulation time, the $^{2}t_r$ is calculated (refer Figure S7-S8 for additional information on data reconstruction and peak placement).

In Figure 2 the conceptual and general data analysis framework used in this study is shown, illustrating the different data treatment steps, necessary actions and the parameters extracted for comparison purposes. The internal pre-treatment actions are thereby carried out within each SP and if possible, kept to the equivalent settings to ensure comparability of the obtained data. In detail, the signal-to-noise threshold (or equivalent parameter) was set to S/N: 30 with baseline correction shortly above the calculated noise. After removing of peaks resulting from column bleed, the resulted methods yielded 75-120 peaks for the standard dataset (depending on the concentration level) and 150-300 peaks for the chocolate samples. The additional unintentional peaks in the standard measurement result from manufacturer side impurities in the substances and injection artifacts. To ensure comparability within the obtained results, the pre-processing methods were adjusted, under the guidance of expert users and software developers, to yield approximately the same number of compounds per analysis and maximum quality for the observed detector response.

The authors consider parameters such as first dimension retention time ($^{1}t_r$), second dimension retention time ($^{2}t_r$), retention index (RI), correlation coefficient (R^2) of obtained calibration curves and mass spectral quality of crucial importance and therefore mandatory for evaluation by dedicated GC×GC software (standard parameters). These parameters were obtained for the 1 ppb level of the standard measurement and averaged across the acquired triplicates. Additionally, the deconvoluted mass spectra were exported and averaged for comparison (refer section 2.6 for details).

191 Based on these standard parameters eight different software packages (SPs) were closely
192 examined: Analyzer ProXD v. 1.8 (Spectral Works Ltd, UK), Canvas v.1.6 (J&X- Technologies,
193 China), ChromaTof v.4.72 (LECO St. Joseph, USA), ChromSpace v. 1.5.1 (SepSolv, UK),
194 ChromSquare v.2.3. (Shimadzu, Japan), GCImage v. 2.5 HR (GCImage LLC, USA), GasPedal
195 (Decodon, Germany) and OpenChrom (Lablicate, Germany). The purpose of this article is to
196 highlight the necessity and value of a data analysis framework and a readily available benchmark
197 dataset, not to promote the use of a specific SP, therefore the results for the SPs are anonymized.
198 Table 1 displays the standard duty capabilities of the investigated SPs. Comparing the above-
199 mentioned metrics points out differences within the applied SPs. Thus allowing improvements or
200 the adaption of specific best practices in terms of data processing, generally aiming to increase
201 comparability across different platforms. For benchmark purposes it is therefore crucial that
202 expert users, of the tested SPs, carry out the processing or give advice for the settings of the
203 investigated algorithms [28,29].

204 The authors are aware that they are not necessarily experts in the operation of all the investigated
205 SPs, which is why, during this study, they kept contact with developers and experts for the
206 respective SPs to ensure the best performance of each SP.

207 Alignment of the chromatographic data in the chocolate set was carried out in the using software
208 packages with this functionality built in. Aligned peak lists were exported, sorted and cleaned
209 (data wrangling). The area values were standardized to the internal standard (1-
210 flouronaphthalene), mean centered and normalized using z-score normalization for each
211 compound (variable) followed by a global square root data transformation. Analysis of variance
212 (ANOVA), with a significance value of $\alpha = 0.05$, was carried out to filter for only significant

compounds. These compounds were then used in a principal component analysis (PCA) to visualize the effect in discriminating the individual samples and potential as marker substances. In-built statistics that lead to PCA were performed as allowed by the individual SPs. These PCAs were then compared to the results obtained with the routine described above.

2.5 CALIBRATION CURVE AND LEAVE ONE OUT EXPERIMENT

Based on the different concentration levels, a calibration curve and the corresponding correlation coefficient between the response values and concentrations were calculated for each compound within each SP capable of doing so. Additionally, calibration curves omitting the 0.4 ppb level were produced subsequently followed by the determination of the 0.4 ppb levels concentration (leave one out experiment). These re-predictions were then used as measure for in-software coherence.

2.6 MASS SPECTRAL COMPARISON

Mass spectral information was extracted and exported from each of the 36 standard compounds in triplicate. Due to the differences in library matching algorithms, comparability of the results needed to be ensured. Therefore, the deconvoluted mass spectra were exported, averaged across the triplicates and matched against the NIST library (v2.2, 2014) using the NIST matching algorithm. This step was performed externally to ensure fair comparability of the matching factors. Additionally, unprocessed raw spectra were extracted from the original data file with an in-house developed tool (refer figure S15 in the supporting information for additional details). The software processed spectra were then compared to the raw spectra highlighting the effects of base processing actions on mass spectra.

Software	^1t_r	^2t_r	RI	Calibration	Library	Alignment
A	✓	✓	✓	✓	✓	✓
B	✓	✓	✓	✓	✓	✗
C	✓	✓	✓	✗	✓	✓
D	✓	✓	✓	✓	✓	✓
E	✓	✓	✗	✗	✗#	✓
F	✓	✓	✗	✗	✓	✓
G	✓	✓	✓	✓	✓	✗*
H	✓	✓	✓	✓	✓	✓

*Table 1: Standard duty capabilities of the individual SPs. Checkmark represent that the feature is implemented in the software whilst the cross indicates the feature is not yet implemented. # Software requires pre-processed peak lists and spectra, which is contains an already performed library search for each compound. * Software does not allow alignment; however, pre-determined and locally restricted regions can be compared in a batch-wise manner.*

2.7 CHEMOMETRIC DATA ANALYSIS

The chocolate dataset was used to investigate the performance of features that solely relate to GC×GC data such as alignment and in-built statistics. Aligned peak tables, containing normalized area information, were exported for each capable SP and wrangled in the shape outlined in Figure S9 in the supporting information. Values for peaks not present in certain chocolate types were

246 set to read not-any-number (NaN). The area values for each analyte and measurement were
247 standardized to min-max values and their residuals investigated for normal distribution using the
248 Kolmogorov-Smirnoff test. Since all obtained lists show positive skewness and positive excess in
249 kurtosis, the data was transformed with a square root transformation, resulting in normal
250 distributed data fitted for further analysis.

251 Analysis of Variance (ANOVA) was applied for data reduction purposes filter for analytes with a
252 significance value of α=0.05. This reduced dataset then underwent principal component analysis
253 in MATLAB allowing for cluster- and residual analysis highlighting differences in the chemical
254 composition of the individual chocolates flavor profiles.

255

3 RESULTS AND DISCUSSION

To demonstrate the added value of the introduced data analysis framework and benchmark dataset, the performance of the eight different software packages is compared below. As mentioned before, the authors consider time and detection-dependent parameters as standard requirements for GC×GC software and focus on them in close detail, highlighting the impact of the individual SPs on the results and the data.

3.1 FIRST- AND SECOND DIMENSION RETENTION TIMES

Ideally, the retention information for each compound would be identical and independent from the used SP. However, a comparison of the processed $^{1}t_{r}$ information revealed deviations among the individual SPs. These variations are illustrated as distribution for d-limonene, linalool, lilial and benzylcinnamate (Substances # 1, 5, 24 and 36) in Figure 3A. These compounds were chosen to represent the dataset since they cover a wide span of chromatographic space and represent a large variety of chemical compounds. The distribution of all compounds is listed in the supporting information (Table TS2 and Figures S10-12). Considering different data transformation, peak detection and placement algorithms (refer to Figure S8, supporting information) as reason for such variations, software-dependent tendencies are expected. To investigate for such tendencies, the obtained values for each compound (specifically mean values across the replicates) were sorted in ascending order, allowing to rank the SPs according to their yielded value (SP yielding the highest value is ranked 8 the SP with the lowest value is ranked 1). Plotting this rank with a color code for each SP in order of compound elution allows to determine whether a certain SP is prone to yield either higher or lower values (illustrated in Figure 3B). It is evident that SPs C and D mainly provide the extreme values within the $^{1}t_{r}$.

In terms of comparability, a change in software from the SP yielding the highest to the SP yielding the lowest value would thereby represent the "worst-case" with a maximal difference. Plotting this difference against the compounds elution order (Figure 3C) allows to read for trends based on (in this case) volatility. Lower volatile compounds appear to be more prone for larger maximum differences. Thereby, the mean value calculated from the differences would be the worst-case expectancy value for a change in software. For the applied eight SPs this expectancy value calculates to 4.2 s (figure 3 C), which roughly corresponds to the applied modulation time of 4 s.

This discrepancy indicate general differences in raw data transformation and peak placement between software C and D. The fact that software C purely processes based on 1D raw data and software D relies exclusively on a 2D transformed raw data set, supports this conclusion. It should also be valid in case of general differences within the data transformation algorithms, which is investigated by the creation of two "next-to-worse" scenarios by leaving out either the SP prone to yield the highest or lowest values (either SP C or SP D) for the calculation of the expectancy value. This resulted in a change of the expectancy value to 3.7/3.8 s, which again resembles the applied modulation period of 4 s and thus indicating differences in the data transformation algorithms.

Figure 4 depicts the results for $^{2}t_{r}$ information. Evaluation for tendencies were performed as described above. Again, software dependent tendencies are observed with SP C yielding the highest value of $^{2}t_{r}$ whilst software E yielded the lowest value. The worst-case expectancy (Figure 4C) evaluates as deviation of 0.15 s and the "next-to-worse" case scenario evaluates to 0.1 s deviation. In the GC×GC community 2nd dimension peak widths are generally considered as a

criterion for the quality of the 2nd dimension's separation. For cryogenic modulation good peak widths are generally considered between 0.1 and 0.2 s. Considering this ideal peak widths and the average differences in ^2t_r of 0.15 allows for the conclusion of deviations of up to 1-2 peak widths as result of a simple change of software. Underlying missing values for the compounds (#18, 19, 20, 21 and 22) is the phenomenon of wrap-around. Depending on the SP, these compounds' ^2t_r s are reported close to the modulation time, e.g. 3.992 s, or wrapped e.g. 0.015 s and therefore result of course in tremendous differences.

The comparison of ^1t_r information demonstrates that the choice of software affects the ^1t_r results due to differences in raw data transformation and peak detection algorithms. Thereby the maximal differences estimate to be close to the applied modulation period. However, within one SP the ^1t_r determination is consistent regarding each compound. The analysis of ^2t_r revealed differences among the individual SPs with maximal deviations up to 2 peak widths, thereby supporting the conclusion drawn from the results for ^1t_r.

3.2 RETENTION INDICES

Allowing for instrument independent normalization, the application of linear temperature-programmed retention indices is widely accepted in GC laboratories. Despite several attempts [30–32] introducing a similar concept for the second dimension in comprehensive GC, an applicable and robust system has not been established within the community yet. For this reason, the dataset permits only “Van den Dool” indices to be calculated. The overall robustness of the RI within each software package was calculated as the standard deviation of the calculated RI within the replicate injections. Thereby this measure, with few exceptions, ranked below 1 retention index unit (RIU) indicating that coherence within each individual software package is granted.

Figure 5 illustrates distribution, ranking and worst-case scenario of the RIs. Software D shows tendencies to yield the lowest RI values, whereas no software seems to yield exclusively the highest value. Thereby, worst-case scenario calculates to a difference of 4 RIU. Additionally, an increasing tendency of the maximal deviation of the RI can be observed with increasing retention time as illustrated by Figure 5C. The gap thereby is caused by the second isomer of lyral for which the majority of the SPs could not determine a plausible RI. The reason for this is unknown to the authors calculating the RI manually yields plausible results however, five out of six SPs capable to calculate RIs, calculate a RI for this particulate compound with numbers in the range of 1e6.

The variations most likely originate from the aforementioned differences in peak placement or the inbuilt RI calculation itself, since the total chromatographic runtime is explicitly stored in the ANDI MS (*.cdf) file-format.

As a derived value, RIs allow for intersystem comparability. However, calculation algorithms of the SPs yield results with averaged 5 RIU differences with increasing tendencies toward higher boiling substances. Nevertheless, the calculation within each SP appeared coherent with deviations of ≤ 1 RIU. Comparability must be ensured, especially when RIs are used to identify unknowns. As demonstrated by using the benchmark data set, the individual SPs have a significant influence on the RI's investigations either through peak placement or calculation. Thereby the benchmark dataset poses a useful possibility for identifying and harmonizing the algorithms across different SPs reducing inter-laboratory variability.

3.3 MASS SPECTRAL COMPARISON

Table 2 lists the library matching factors (MF) achieved by the individual SPs for d-limonene, linalool, lilial, benzylcinnamate and safrole. Additionally, the maximal difference across the SPs for each compound is listed. Overall, the values of the MFs are close together and yield acceptable results despite slight deviations. However, these differences suggest that MS spectra deviate from one software to another, indicating alteration induced by each SP. Figure 6 clearly illustrates this behavior by the comparison of spectral sections of safrole, extracted from SPs A, D and the raw data. In Figure 6A the m/z range 160-165 is compared. This area represents the region encompassing the molecular ion at m/z: 162. Software A and D produce similar results although the raw data yields a higher relative intensity of the non-base peak ions. Considering the fragmentation of safrole, the abstraction of CH_3O and C_2H_3 yield prominent m/z 131 and 135 signals (Figure 6B). Figure 6C elaborates this region more closely. Again, slight differences within the ion intensities are observable. However, the ion ratios within the individual spectra changes. For example, the ratio between m/z 131 and 135 appears to increase by 4% in software D. Instead of ion ratios sometimes the pure difference between the ion in question is considered (intensity(131)-intensity(135)). For this particular example the absolute value of the difference increases by 28% in software D.

Throughout the entire spectrum the differences between significant and relevant ion clusters (50-54, 76-80, 101-105) calculates between 4 and 6%. The added value of the benchmark data set presented is particularly evident with regard to the mass spectra. In GC×GC-MS, the mass spectra are the basis for quantitative, qualitative and exploratory analysis.

	A	B	C	D	E	F	G	H	max diff. [%]
d-limonene	905	907	899	895	905	911	903	903	0.6
linalool	947	902	937	862	947	902	912	817	13.6
safrole	950	974	969	940	950	971	976	976	3.6
lilial	946	955	940	891	945	953	954	955	5.6
benzyl cinnamate	967	974	968	944	966	977	977	977	3.3

Table 2: Matching factors for d-limonene, linalool, lilial, safrole and benzylcinnamate achieved by comparing the extracted mass spectra to the NIST (v.2.2, 2014) library. The maximal difference shows the difference in matching factor for the worst-case scenario (e.g. d-limonene changing from D to F).

372 **3.4 CALIBRATION CURVES AND "LEAVE ONE OUT"**

373 The above stated differences between the spectra might cause differences for quantitative
374 analysis between the software packages reflected in calibration curves. Table 3 lists the
375 correlation coefficients (R^2) for calibration curves generated from manually extracted values and
376 automatically generated calibration curves (bold). In general, both approaches yield similar
377 results with exception for benzylcinnamate. This particular compound is picked as multiple peaks
378 and therefore the overall peak area is split into several parts. However, with a R^2 value of greater
379 than 0.95, the majority of the curves yield acceptable results for quantitation purposes. Table 4
380 contains R^2 and predictive values for d-limonene, linalool, lilial and benzylcinnamate.

381 For some of the compounds the 2-ppb point is the detector is saturated (e.g. lilial and
382 benzylcinnamate), this allows to investigate the SPs capability to handle such occurrences.
383 Whereas only software G offers to carry out a linearity check, none of the other SPs allow for
384 compensation or warning mechanisms.

385 Table 4 lists the predictive values for the 0.4 ppb point for the "leave one out" experiments.
386 Although the prediction appears very close, software B underestimates the concentration whilst
387 software D overestimates the concentration slightly.

388

	A	B	C*	D	E*	F*	G	H
d-limonene	0.95 | **0.95**	0.97 | **0.90**	0.95	0.89 | **0.64**	0.95	0.96	0.93 | **0.86**	0.90 | **0.99**
linalool	0.96 | **0.96**	0.98 | **0.61**	0.94	0.80 | **0.62**	0.96	0.93	0.93 | **0.86**	0.98 | **0.89**
lilial	0.86 | **0.96**	0.79 | **0.99**	0.88	0.80 | **0.88**	0.86	0.83	0.86 | **0.85**	0.83 | **0.92**
benzyl-cinnamate	0.55 | **0.91**	0.47 | **0.94**	0.99	0.61 | **0.76**	0.55	0.52	0.59 | **0.95**	0.67 | **0.98**

389 *Table 3: R^2 values calculated for (manually extracted values) | **(each software's algorithm)**.*

390 *Thereby the peak areas were selected "as-is" without manual reintegration. Concentration levels*

391 *0.2-2 ppb for d-limonene and linalool and 0.2-1 ppb for lilial and benzylcinnamte. Software marked*

392 *with * does not provide an in-build feature for calibration curves.*

	A	B	C	D	E	F	G	H
d-limonene	0.43	0.35	0.43	0.55	0.43	0.43	0.42	0.50
linalool	0.43	0.33	0.47	0.40	0.43	0.35	0.44	0.30
lilial	0.42	0.38	0.55	0.52	0.42	0.43	0.43	0.36
benzylcinnamate	0.42	0.39	0.45	0.48	0.42	0.43	0.29	0.36

393 *Table 4: Predictive values for the 0.4 ppb level based on the leave one out experiments.*

394

The comparison of mass spectral information demonstrated the alterations induced by each SP e.g. the intensities of prominent ions and ion ratios differed up to 6%. These deviations then result in different matching factors when performing library comparison. Differences in algorithms such as deconvolution and background subtraction are the most likely explanation of the observed mass spectral differences.

The investigation of the quantitative analysis indicate consistency within the SPs. However, comparison of the re-prediction among the SPs reveal deviations related to the calculated response e.g. software D yields higher ion intensities than software A for d-limonene and also overestimates in the re-prediction of the 0.4 ppb level. Regardless of that, the presented benchmark dataset ensures that the underlying data is identical.

3.5 Alignment

Software Packages A, C, D, E, F and H are capable of chromatographic alignment using either peak list or pixel based. Although the underlying data is the same, the number of peaks/compounds produced by the SPs after alignment differ as indicated by Table 5. To ensure only significant substances are considered in further chemometrical analysis, normal distribution of the data was ensured as illustrated in Figure S13 in the supporting information.

Table 5 list the skewness of the residual distribution prior and post the standardization and transformation procedures as investigative and decision criterion for data transformation processes. Thereby, the raw distributions deviate to various degrees from the normal distribution (e.g. the data obtained from SP C is more normally distributed than the data obtained by SP F) and would therefore demand for different data transformations. However, to keep the treatment the same for each SP the obtained peak lists were all logarithmic transformed resulting in a good compromise solution (-1< skewness < 1).

	A	**C**	**D**	**E**	**F**	**H**
post Alignment	613	1318	133	5327	286	1100
pre skewness	1.2	0.56	1.02	1.03	2.16	2.08
post-skewness	0.07	0.38	-0.01	-0.01	-0.63	-0.6
post-ANOVA	420	264	55	606	137	221

Table 5: Number of peaks/compounds after the alignment and after the ANOVA actions. Pre- and post-skewness show the effect of the standardization and normalization procedure.

Applying ANOVA (with $\alpha = 0.05$) allows to reduce the aligned compound lists to substances responsible for the differences within the individual chocolate types. This represents a drastic data reduction (Table 5, post-ANOVA column). In Figure 7 the PCAs for theses pre filtered compound sets are displayed. Based on the hypothesis, similar flavoured chocolates will cluster closely, the software packages A, C, D and H provide comparable results with close clusters according to chocolate type and/or filling.

Regarding the explained variance, each analysis performs similarly, ranging between 53 and 72%. Thereby, the density of the clusters varies within the individual SPs. However, the general tendencies shown by each analysis are similar, clear separation of chocolates with orange and mint flavour from the bulk based on principal component 1 (PC1). According to the ingredients list the mint flavoured chocolate contains 80% cocoa whilst the rest of the chocolates range between 40 and 60%. The orange chocolate covering the lower end of this range. This suggests that the principle component 2 (PC2) separates the chocolates according to their cocoa content. Thereby, the strongly correlated features (grey) are the same compounds within each analysis when they are detected.

Even though approaches for data alignment varied, the visualization of the results via PCA revealed consistent trends among all SPs. Although the underlying data is the same, the different alignment procedures yielded different amounts of features as input for the PCAs, resulting in differences in cluster density or cluster separation. Based on such pre-filter and visualization techniques, conclusions for marker substances are typically drawn. Thereby, the utilized statistical technique poses great influence on such selection processes. Therefore, it is important that the compounds responsible for the formation of the cluster are still present after alignment.

Considering the degree of overlap between the aligned lists would allow the investigation of this measure. However, the dataset is not set-up to investigate such a measure in detail and might be extended in the near feature (e.g. artificial created matrix on-purpose differences). Regardless of the utilized approach for alignment, the aligned features should be consistent especially considering the further use of statistical techniques such as e.g. marker identification.

3.6 “In-built” statistics

Figure S14 (supporting information) illustrates the result of “in-built” of the SP allowing for PCA. Comparing these plots to the above reported results reveal similar trends. However, the customizability of these plots varies within each SPs and sometimes necessitates the export of the data to third-party programs. For example, the SP D allows no customization resulting in PCA charts that are hard to read due to small font and marker sizes or colours. For comparison purposes, the Eigenvalues were extracted and plotted externally in Matlab R2018a.

Other features such as pre-filter of the matched peak lists, using e.g. Fisher Ratios, t-test, Vulcano charts etc. have not been exploited since each SP follows different calculation guidelines rendering these features incomparable. It also became evident that none of the investigated SPs allow for normal distribution investigations, which is a basic assumption for most of the chemometric analyses.

SPs that allow for in-built statistics and post-data treatment are of particular appeal especially when considering GC×GC for “out-of-academia” applications. Although the general trends appear similar to the manually obtained results, the limited customizability and the lack of transparence of the applied techniques still require the use of third-party software for adequate display or more detailed and accurate analysis. The individual SPs differ widely in the extent of their additional in-build statistics and thus allowing only a partial evaluation of the complete chemometric capabilities.

4. CONCLUSION

A fixed data analysis framework and benchmark data enables the objective comparison of different data processing options. As demonstrated above, differences and tendencies within individual software packages can be identified and addressed. Thereby, it is of utmost importance that the underlying data is the same and of high quality. Additionally, the authors acknowledge that, despite guidance from program authors regarding program parameters for optimal performance were sought and given, optimal performance for each SP might not have been achieved. However, within this study, it became evident, that different SPs show differences in $^{1}t_r$, $^{2}t_r$ and RI results, indicating differences in the peak placement or reconstruction algorithms. This fact should not be overlooked, particularly if these parameters are used for identification purposes. The presented dataset might help developers to overcome these differences, ensuring inter-laboratory and inter-experimental comparability.

Generally, the detector response represents the parameter that is used for further chemical analysis such as quantitative, qualitative or even chemometric analyses. Therefore, the response is required to be stable and comparable. As demonstrated, the SPs influence the quality of the mass spectra by alteration of the mass spectra via deconvolution or background subtracting method. These methods often differ from manufacturer to manufacturer and are often proprietary. To ensure a certain degree of harmonization and standardization, a globally recognized benchmark dataset is needed.

With the increase in GC×GC's popularity, the demand for sophisticated post-data treatment increased as well. Currently such analyses are time consuming and need to be performed by trained experts. Therefore, "all-in-one" solutions are highly desirable, which demands that the

494 user can be aware of the underlying algorithms that are used especially when in-built statistic
495 tools are included. However, only few of the available SPs offer this capability. Considering
496 routine or designed applications with several replicates, alignment procedures are essential.
497 However, as demonstrated by this dataset the different strategies lead to different alignment
498 results and, although general tendencies appear to be consistent, different number of
499 compounds are aligned and used as input for post-analysis. It is evident that for the core strengths
500 of GC×GC, such as identification of marker substances, the outcome of the alignment procedures
501 needs to be coherent independently of the applied algorithms.

502 Certainly, the presented data analysis framework and data sets do not show the complexity and
503 scope reached by other GC × GC studies, but allow for performance assessment of different
504 processing tools as well as a didactic entrance in evaluation of multidimensional data. Moreover,
505 not each and every special case in terms of processing is represented by the dataset, on the
506 contrary the dataset needs to be evolved, adapted and made available for future elevation in the
507 field of GC×GC data analysis.

508 More information of this data can be found in the supporting information and on the homepage
509 of the obiachem group (https://www.obiachem.uliege.be/cms/c_5882500/en/data-
510 visualization-projects). In conclusion, the presented data framework and benchmark data
511 represent a valuable opportunity to test, harmonize and improve existing and future features in
512 GC×GC SPs.

513 **5. ACKNOWLEDGEMENT**

514 Authors are grateful towards JEOL Benelux for providing the mass spectrometer employed in this
515 study.

6. FUNDING

This research was supported by the Fonds de la Recherche Scientifique - FNRS and the Fonds Wetenschappelijk Onderzoek - Vlaanderen (FWO) under EOS Project n°30897864

6. REFERNCES

[1] J.C. Giddings, Sample dimensionality: A predictor of order-disorder in component peak distribution in multidimensional separation, J. Chromatogr. A. 703 (1995) 3–15. https://doi.org/10.1016/0021-9673(95)00249-M.

[2] J. Jáčová, A. Gardlo, J.M.D. Dimandja, T. Adam, D. Friedecký, Impact of sample dimensionality on orthogonality metrics in comprehensive two-dimensional separations, Anal. Chim. Acta. 1064 (2019) 138–149. https://doi.org/10.1016/j.aca.2019.03.018.

[3] B.A. Weggler, B. Gruber, J.-F. Focant, Comprehensive Two-dimensional gas-chromatography to study the human exposome: Current trends and perspectives, Curr. Opin. Environ. Sci. Heal. (2020). https://doi.org/10.1016/j.coesh.2020.02.011.

[4] B. Gruber, B.A. Weggler, R. Jaramillo, K.A. Murrell, P.K. Piotrowski, F.L. Dorman, Comprehensive two-dimensional gas chromatography in forensic science: A critical review of recent trends, TrAC Trends Anal. Chem. 105 (2018) 292–301. https://doi.org/10.1016/J.TRAC.2018.05.017.

[5] M.E. Machado, Comprehensive two-dimensional gas chromatography for the analysis of nitrogen-containing compounds in fossil fuels: A review, Talanta. 198 (2019) 263–276. https://doi.org/10.1016/j.talanta.2019.02.031.

[6] S.E. Prebihalo, K.L. Berrier, C.E. Freye, H.D. Bahaghighat, N.R. Moore, D.K. Pinkerton, R.E. Synovec, Multidimensional Gas Chromatography: Advances in Instrumentation, Chemometrics, and Applications, Anal. Chem. (2017) acs.analchem.7b04226. https://doi.org/10.1021/acs.analchem.7b04226.

[7] B.J. Pollo, G.L. Alexandrino, F. Augusto, L.W. Hantao, The impact of comprehensive two-dimensional gas chromatography on oil & gas analysis: Recent advances and applications in petroleum industry, TrAC - Trends Anal. Chem. 105 (2018) 202–217. https://doi.org/10.1016/j.trac.2018.05.007.

[8] A.M. Muscalu, T. Górecki, Comprehensive two-dimensional gas chromatography in environmental analysis, TrAC - Trends Anal. Chem. 106 (2018) 225–245. https://doi.org/10.1016/j.trac.2018.07.001.

[9] M.S.S. Amaral, Y. Nolvachai, P.J. Marriott, Comprehensive Two-Dimensional Gas Chromatography Advances in Technology and Applications: Biennial Update, Anal. Chem. (2020). https://doi.org/10.1021/acs.analchem.9b05412.

[10] P.H. Stefanuto, K.A. Perrault, S. Stadler, R. Pesesse, H.N. Leblanc, S.L. Forbes, J.F. Focant, GC × GC-TOFMS and supervised multivariate approaches to study human cadaveric decomposition olfactive signatures, Anal. Bioanal. Chem. 407 (2015) 4767–4778. https://doi.org/10.1007/s00216-015-8683-5.

[11] A. Giri, M. Coutriade, A. Racaud, K. Okuda, J. Dane, R.B. Cody, J.F. Focant, Molecular Characterization of Volatiles and Petrochemical Base Oils by Photo-Ionization GC×GC-TOF-MS, Anal. Chem. 89 (2017) 5395–5403.

https://doi.org/10.1021/acs.analchem.7b00124.

[12] S.E. Reichenbach, M. Ni, V. Kottapalli, A. Visvanathan, Information technologies for comprehensive two-dimensional gas chromatography, Chemom. Intell. Lab. Syst. 71 (2004) 107–120. https://doi.org/10.1016/j.chemolab.2003.12.009.

[13] A.A.S. Sampat, M. Lopatka, G. Vivó-Truyols, P.J. Schoenmakers, A.C. van Asten, Towards chemical profiling of ignitable liquids with comprehensive two-dimensional gas chromatography: Exploring forensic application to neat white spirits, Forensic Sci. Int. 267 (2016) 183–195. https://doi.org/10.1016/j.forsciint.2016.08.006.

[14] E.A. Higgins Keppler, C.L. Jenkins, T.J. Davis, H.D. Bean, Advances in the application of comprehensive two-dimensional gas chromatography in metabolomics, TrAC - Trends Anal. Chem. 109 (2018) 275–286. https://doi.org/10.1016/j.trac.2018.10.015.

[15] A. Sampat, M. Lopatka, M. Sjerps, G. Vivo-Truyols, P. Schoenmakers, A. van Asten, Forensic potential of comprehensive two-dimensional gas chromatography, TrAC - Trends Anal. Chem. 80 (2016) 345–363. https://doi.org/10.1016/j.trac.2015.10.011.

[16] C. Cordero, J. Kiefl, S.E. Reichenbach, C. Bicchi, Characterization of odorant patterns by comprehensive two-dimensional gas chromatography: A challenge in omic studies, TrAC - Trends Anal. Chem. 113 (2019) 364–378. https://doi.org/10.1016/j.trac.2018.06.005.

[17] S.E. Reichenbach, X. Tian, Q. Tao, E.B. Ledford, Z. Wu, O. Fiehn, Informatics for cross-sample analysis with comprehensive two-dimensional gas chromatography and high-resolution mass spectrometry (GCxGC-HRMS), Talanta. 83 (2011) 1279–1288. https://doi.org/10.1016/j.talanta.2010.09.057.

578 [18] K.M. Pierce, B. Kehimkar, L.C. Marney, J.C. Hoggard, R.E. Synovec, Review of chemometric
579 analysis techniques for comprehensive two dimensional separations data, J. Chromatogr.
580 A. 1255 (2012) 3–11. https://doi.org/10.1016/j.chroma.2012.05.050.

581 [19] S.E. Reichenbach, X. Tian, C. Cordero, Q. Tao, Features for non-targeted cross-sample
582 analysis with comprehensive two-dimensional chromatography, J. Chromatogr. A. 1226
583 (2012) 140–148. https://doi.org/10.1016/j.chroma.2011.07.046.

584 [20] P.E. Sudol, D. V. Gough, S.E. Prebihalo, R.E. Synovec, Impact of data bin size on the
585 classification of diesel fuels using comprehensive two-dimensional gas chromatography
586 with principal component analysis, Talanta. 206 (2020) 120239.
587 https://doi.org/10.1016/j.talanta.2019.120239.

588 [21] K.M. Pierce, L.F. Wood, B.W. Wright, R.E. Synovec, A comprehensive two-dimensional
589 retention time alignment algorithm to enhance chemometric analysis of comprehensive
590 two-dimensional separation data, Anal. Chem. 77 (2005) 7735–7743.
591 https://doi.org/10.1021/ac0511142.

592 [22] Y. Koh, K.K. Pasikanti, C.W. Yap, E.C.Y. Chan, Comparative evaluation of software for
593 retention time alignment of gas chromatography/time-of-flight mass spectrometry-based
594 metabonomic data, J. Chromatogr. A. 1217 (2010) 8308–8316.
595 https://doi.org/10.1016/J.CHROMA.2010.10.101.

596 [23] W. Niu, E. Knight, Q. Xia, B.D. McGarvey, Comparative evaluation of eight software
597 programs for alignment of gas chromatography–mass spectrometry chromatograms in
598 metabolomics experiments, J. Chromatogr. A. 1374 (2014) 199–206.

https://doi.org/10.1016/J.CHROMA.2014.11.005.

[24] B.A. Weggler, L.M. Dubois, N. Gawlitta, T. Gröger, J. Moncur, L. Mondello, S.E. Reichenbach, P.Q. Tranchida, Z. Zhao, R. Zimmermann, M. Zoccali, J.-F. Focant, Standard Dataset 1: Calibration Curve, Fragrance and Allergenes, (2020). https://doi.org/https://doi.org/10.7910/DVN/KA5BTU.

[25] B.A. Weggler, L.M. Dubois, N. Gawlitta, T. Gröger, J. Moncur, L. Mondello, S.E. Reichenbach, P.Q. Tranchida, Z. Zhao, R. Zimmermann, M. Zoccali, J.-F. Focant, Dataset 2: Chocolate Data, (2020). https://doi.org/https://doi.org/10.7910/DVN/AKT6BH.

[26] ASTM, Standard Specification for Analytical Data Interchange Protocol for Chromatographic, 98 (2004) 1–8.

[27] C. Müller, F. Vetter, E. Richter, F. Bracher, Determination of caffeine, myosmine, and nicotine in chocolate by headspace solid-phase microextraction coupled with gas chromatography-tandem mass spectrometry, J. Food Sci. 79 (2014). https://doi.org/10.1111/1750-3841.12339.

[28] G.C.P. Schaafsma, M. Vihinen, Representativeness of variation benchmark datasets, BMC Bioinformatics. 19 (2018) 461. https://doi.org/10.1186/s12859-018-2478-6.

[29] M.R. Aniba, O. Poch, J.D. Thompson, Issues in bioinformatics benchmarking: the case study of multiple sequence alignment, Nucleic Acids Res. 38 (2010). https://doi.org/10.1093/nar/gkq625.

[30] M. Jiang, Facile approach for calculation of second dimensional retention indices in comprehensive two dimensional gas chromatography with single injection, Anal. Chem.

(2019). https://doi.org/10.1021/acs.analchem.8b05717.

[31] C. Veenaas, P. Haglund, A retention index system for comprehensive two-dimensional gas chromatography using polyethylene glycols, J. Chromatogr. A. (2018). https://doi.org/10.1016/j.chroma.2017.08.062.

[32] D.M. Mazur, I.G. Zenkevich, V.B. Artaev, O. V. Polyakova, A.T. Lebedev, Regression algorithm for calculating second-dimension retention indices in comprehensive two-dimensional gas chromatography, J. Chromatogr. A. (2018). https://doi.org/10.1016/j.chroma.2018.07.038.

CAPTIONS

Figure 1: *GC×GC Chromatogram for A) the standard Dataset (1ppb) and B) a representative chocolate measurement (mint and lime flavor). For the standard dataset more than the 36 standard compounds and the added alkane standard can be observed as result of manufacturer sided impurities and injection artefacts.*

Figure 2: *Conceptual and general data analysis framework used in this study separated whether the necessary steps and actions were performed within the used GC×GC software packages or in external software. In the bottom row the extracted chemical information is listed that was used to derive measures for comparative purposes.*

Figure 3: *A) distribution of* $^{1}t_r$ *for d-limonene, linalool, lilial and benzylcinnamate (Substances # 1, 5, 24, 36). these substances represent the spectrum of standard compounds and encompass the entire 2D separation space. B) Software dependent ranking of the* $^{1}t_r$ *values for each compound (highest to lowest value). Thereby, 8 represents the highest value, 1 the lowest. Different colours represent individual software packages. For example, for compound 36 software C yields the highest and D the lowest value. C) Worst-case scenario for changing the software. The mean value, represented by the dashed line, thereby resembles the applied modulation time.*

Figure 4: *A) distribution of* $^{2}t_r$ *for d-limonene, linalool, lilial and benzylcinnamate (Substances #1, 5, 24 and 36). B) Software dependent ranking of the* $^{2}t_r$ *values for each compound (highest to lowest value). Thereby, 8 represents the highest value, 1 the lowest. Different colours represent individual software packages. C) Worst-case scenario for changing the software. Mean value represented by the dashed line. Gap in-between as result from compounds on the verge of*

wrapping. Several SPs report these compounds as unwrapped whilst the other report them as wrapped.

***Figure 5**: A) distribution of RI for d-limonene, linalool, lilial and benzylcinnamate. B) Software dependent ranking of the RI values for each compound (highest to lowest value). Thereby, 6 represents the highest value, 1 the lowest. Different colours represent individual software packages. C) Worst-case scenario for changing the software. Mean value represented by the dashed line. The gap in the line represents the second isomer of lyral for which the majority of the software packages could not determine a plausible retention index.*

***Figure 6:** A) Comparison of the m/z area 160-165 in the spectrum of safrole. Grey shaded spectrum (middle) is extracted from the raw data. B) Formation of m/z 131 and 135 fragments of safrole due to CH_3O and C_2H_3 separation. C) Comparison of the m/z area 130-135. Differences between software A and D visible.*

***Figure 7:** PCAs after ANOVA with $\alpha = 0.05$ for the SPs allowing for overall alignment (software A, C, D, E, F, H). Grey Stars represent the matched features used as input for the different chocolate types.*