Journal of Chromatography A 1699 (2023) 464010

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Augmented visualization by computer vision and chromatographic fingerprinting on comprehensive two-dimensional gas chromatographic patterns: Unraveling diagnostic signatures in food volatilome

Andrea Caratti[a,1], Simone Squara[a,1], Carlo Bicchi[a], Qingping Tao[b], Daniel Geschwender[b], Stephen E. Reichenbach[b,c], Francesco Ferrero[d], Giorgio Borreani[d], Chiara Cordero[a,*]

[a] Dipartimento di Scienza e Tecnologia del Farmaco, Università di Torino, Via Pietro Giuria 9, Turin I-10125, Italy
[b] GC Image LLC, Lincoln, NE, USA
[c] Computer Science and Engineering Department, University of Nebraska – Lincoln, Lincoln, NE, USA
[d] Department of Agricultural, Forestry and Food Sciences, Università di Torino, Grugliasco TO, Italy

ARTICLE INFO

Article history:
Received 13 March 2023
Revised 18 April 2023
Accepted 20 April 2023
Available online 23 April 2023

Keywords:
Comprehensive two-dimensional gas chromatography
Augmented visualization
Ut fingerprinting
Butter volatilome analysis
Chromatograms registration and alignment

ABSTRACT

Computer Vision is an approach of Artificial Intelligence (AI) that conceptually enables "*computers and systems to derive useful information from digital images*" giving access to higher-level information and "*take actions or make recommendations based on that information*". Comprehensive two-dimensional chromatography gives access to highly detailed, accurate, yet unstructured information on the sample's chemical composition, and makes it possible to exploit the AI concepts at the data processing level (*e.g.*, by Computer Vision) to rationalize raw data explorations. The goal is the understanding of the biological phenomena interrelated to a specific/diagnostic chemical signature. This study introduces a novel workflow for Computer Vision based on pattern recognition algorithms (*i.e.*, combined untargeted and targeted UT fingerprinting) which includes the generation of composite Class Images for representative samples' classes, their effective re-alignment and registration against a comprehensive feature template followed by Augmented Visualization by comparative visual analysis. As an illustrative application, a sample set originated from a Research Project on artisanal butter (from raw sweet cream to ripened butter) is explored, capturing the evolution of volatile components along the production chain and the impact of different microbial cultures on the finished product volatilome. The workflow has significant advantages compared to the classical one-step pairwise comparison process given the ability to realign and pairwise compare both targeted and untargeted chromatographic features belonging to Class Images resembling chemical patterns from many different samples with intrinsic biological variability.

1. Introduction

The process of chromatographic fingerprinting on patterns obtained by comprehensive two-dimensional chromatography (C2DC) aims at detecting, re-aligning, and comparing features extracted from 2D peaks across a series of analyzed samples [1]. The goal is to have access to higher-level information related to unique and/or distinctive composition ant its relation to the phenomenon under study.

The representation of C2D chromatograms in digital images is obtained by *rasterization*, i.e., ordering detector data values from single modulation periods (or cycles) as a column of pixels (picture elements). Note that, when mass spectrometry (MS) with suitable analyzers is used, each pixel corresponds to a spectrum, thus information on feature identity is left available for further investigation. Rasterized pixel columns are then sequenced along the X-axis (from left to right) according to the first dimension (^{1}D) retention, while the second dimension (^{2}D) data are presented in a right-handed Cartesian coordinate system, and the ^{2}D retention time is plotted on the Y-axis (from bottom-to-top) [1]. Chromatographic

* Corresponding author.
E-mail address: chiara.cordero@unito.it (C. Cordero).
[1] Andrea Caratti and Simone Squara, listed in alphabetical order equally contributed to this work.

https://doi.org/10.1016/j.chroma.2023.464010

images can be processed in various ways, adopting different feature types and related tools [1,2].

The peak feature approaches, the most used tools implemented in all commercial data processing software, collect and sum data with associated metadata for each analyte peak. When multiple chromatograms are examined, peak features are then matched across multiple samples enabling detailed profiling of their components. If they correspond to targeted analytes (known constituents), their peak is explicitly matched across chromatograms by their target name. If they are not *a priori* targeted (*i.e.*, untargeted analytes), the cross-comparative analysis is more challenging and unidentified peaks across many chromatograms must be matched by the software in a process usually referred to as peak matching or peak tracking [3]. An almost comprehensive review of the peak features approaches for C2DC has been recently proposed by Stilo et al. [1].

To compensate for peak matching inconsistencies occurring for large data sets, region and peak region feature concepts have been introduced. Within region feature approaches, the tile-based Fisher ratio analysis [4–6] enables effective cross-comparative analysis by adopting rectangular tiles to generate features over the chromatographic space and inspecting tile's detector response (including single *m/z* channels) for meaningful variations across samples (e.g., Fisher discriminant analysis). By peak region features, tiles are generated around 2D peaks footprint with some advantages when chromatographic misalignment occurs. Further insights on this feature type are provided below.

Independent of the feature type adopted, the process that realigns and compares 2D peak patterns across many chromatograms is referred to as chromatographic fingerprinting [1]. However, if it can provide visual evidence of compositional differences between samples (*e.g.*, by comparative visualization [7]) it can also be considered as a Computer Vision (CV) approach. CV is defined as "*a field of artificial intelligence (AI) that enables computers and systems to derive meaningful information from digital images… and take actions or make recommendations based on that information. If AI enables computers to think, computer vision enables them to see, observe and understand*" [8].

Comparative visualization, a type of CV, was one of the earliest approaches applied to reveal compositional differences between pairs of 2D chromatograms. However, data point features which comprehensively cover all detected/untargeted analytes with the highest precision on 2D fingerprint, result in greater computational complexity since analytes are represented by multiple features (i.e., multiple spectra per analyte), and/or might be imprecise due to retention-times inconsistencies which confound feature matching across chromatograms.

In this study, a novel workflow for effective comparative visualization of multiple chromatograms is proposed. In particular, CV is performed by combined untargeted and targeted (UT) fingerprinting [9,10] with a fully automated procedure while Augmented Visualization (AV) [11] is made possible by 2D chromatograms alignment on reliable peaks and peak regions, information (*i.e.*, feature metadata) which is available at each stage.

As a challenging test bench, butter volatilome evolution along the production chain is studied. The complexity of the analyzed mixture relates to the large dynamic range of concentrations for characteristic chemical classes (*e.g.*, short-chain fatty acids, alcohols, aldehydes, esters and lactones, etc.) associated to a limited volatility range. A CV approach with molecular resolution is of great interest for food technology, it prompts information on compositional differences induced by a specific process, or technological conditions, supports decision-making strategies, and facilitates process optimization and product *identitation* [12].

2. Materials and methods

2.1. Reference standards and solvents

Pure standards of *n*-alkanes (from *n*-C7 to *n*-C30) used for Linear Retention Indices (I^T) calibration, of α/β-thujone and 2-methyloctynoate used as Internal Standards (ISTDs), and solvents (cyclohexane and dibutyl phthalate – 99% of purity) were from Merck (Milan, Italy).

For key-aroma compounds and potent odorants investigated in the study, identity confirmation was by authentic standards available from Merck (Milan, Italy).

2.2. Reference solutions and calibration mixtures

Standard stock solutions of reference analytes were prepared at a concentration of 10 g/L in cyclohexane and stored at −18 °C for one week.

The Reference Solution for analytes identity confirmation was prepared by mixing suitable amounts of standard stock solutions to reach the concentration of 0.100 g/L using cyclohexane as solvent.

The *n*-alkanes solution for I^T calibration was prepared by mixing suitable amounts of standard stock solutions of pure standards to reach the concentration of 0.050 g/L using cyclohexane as solvent.

ISTDs working solution for standard-in fiber pre-loading [13] was prepared at 0.100 g/L in dibutyl phthalate and stored at −18 °C in sealed vials. ISTDs were used for validation purposes (method precision and repeatability) and to normalize the analytes' absolute responses (*i.e.*,% normalized response).

2.3. Cream and butter samples

Sweet cream from raw cow milk and butter samples were provided by Dr. Fabio Bruno and were part of an experimental production of artisanal butter by Beppino Occelli Agrinatura S.r.l (Farigliano, Cuneo, Italy). Samples consisted of raw sweet cream from cow milk from two different farms located in Piedmont, Italy (Farm #1 and Farm #2) which were then mixed in a ratio 1:1 before pasteurization (95 °C/10 s) and inoculum with selected cultures (Standard - STD and TEST) for butter production. After 18 h from the inoculum, samples were cooled to allow butterfat crystallization, which facilitates the churning and improves butter texture (process step generally referred to as *aging*). Once ready, samples were transferred to a continuous butter-maker via a plate heat exchanger. The butter obtained, after draining and washing, was then ripened until 40 days.

Quality control samples (QCs) were obtained from a commercial butter which was gently melted and homogenized before storage at −80 °C until analyzed. The melting process, expected to have an impact on the volatile fraction, was performed to obtain a truly homogeneous QC for analytical checks. The butter matrix is known to have a certain inhomogeneity due to the nonuniform bacterial growth within the crystallized fat.

Details on the sample set under study are provided in Table 1 together with the acronyms adopted in the text; a schematic diagram of available samples together with biological and technical replicates is shown in Supplementary Figure 1.

2.4. Headspace solid-phase microextraction: devices and conditions

Volatiles from cream and butter samples were extracted by HS-SPME with a divinylbenzene/carboxen/polydimethyl siloxane (DVB/CAR/PDMS) fiber (d_f 50/30 µm; 2 cm length) from Merck Lifesciences - Supelco (Bellefonte, PA, USA) chosen on the basis of

Table 1
List of samples together with processing details, acronyms adopted in the text, and corresponding class image generated for the computer vision (CV).

Production stage	Samples details	Analyzed samples		Class Images
		Farm #1	Farm #2	
Sweet cream - raw		*SCR #1 (n=6)*	*SCR #2 (n=6)*	*SCR*
		Mixed 1:1 ratio		
Pasteurized cream	Standard pasteurization 95 °C/15 s Inoculum "standard" for cream A or "test" for cream B – time 0	*PC STD -T0 (n=6)*	*PC TEST -T0 (n=6)*	*PCT0*
Inoculated cream	Inoculum "standard" for cream A or "test" for cream B - aging for 18 h	*PC STD –T18 (n=6)*	*PC TEST –T18 (n=6)*	*PCT18*
Butter	Butter at time 0	*B STD-T0*	*B TEST-T0*	*BT0*
	Butter after 8 days of ripening	*B STD-T8*	*B TEST-T8*	*BT8*
	Butter after 20 days of ripening	*B STD-T20*	*B TEST-T20*	*BT20*
	Butter after 40 days of ripening	*B STD-T40*	*B TEST-T40*	*BT40*

previous studies [14]. The SPME fiber was conditioned before use as recommended by the manufacturer.

The ISTDs for response normalization and quality control were preloaded onto the SPME device [15] by sampling 5.0 µL of α/β-thujone and methyl 2-octynoate ISTDs solution (0.100 g/L) placed in a 20 mL headspace vial. ISTDs pre-loading was performed by exposing the SPME device in the HS at a temperature of 50 °C for 5 min.

Sampling was carried out on 0.100 ± 0.005 g of sample, precisely weighed, in 20 mL headspace vials, kept at 50 °C for 50 min under constant agitation. The very low amount of sample was chosen to match the HS linearity conditions for most of the characteristic analytes of the dairy fat volatilome. After extraction, the SPME device was automatically transferred to the split/splitless injection port of the GC × GC system, kept at 250 °C, and submitted to thermal desorption was for 5 min. Samples were analyzed in three replicates randomly distributed over a two-week time frame.

2.5. GC × GC-TOFMS: instrument set-up and conditions

Comprehensive two-dimensional GC analyses were carried out with an Agilent 7890B GC chromatograph (Agilent Technologies, Wilmington, DE, USA) coupled with a Markes BenchTOF Select™ mass spectrometer featuring Tandem Ionization™ (Markes International, Llantrisant, UK). The GC transfer line was set at 270 °C. The TOFMS was tuned for single ionization at 70 eV and the scan range was set between 35−350 *m/z* with a spectrum acquisition frequency of 100 Hz. The thermal modulator was a loop-type, two-stage KT 2004 (Zoex Corporation, Houston, TX) cooled with liquid nitrogen and controlled by Optimode, v2.0 (SRA Instruments, Cernusco sul Naviglio, Milan, Italy). The modulation period (P_M) was set at 2.5 s, while the hot-jet pulse duration was set at 250 ms. The cold-jet stream at the mass flow controller (MFC) was programmed to linearly reduce the total flow (*i.e.*, 20 L/min) from 40% to 8% along the analytical run.

The column set consisted of a ^{1}D HeavyWax™ column (100% polyethylene glycol – PEG; 30 m × 0.25 mm d_c × 0.25 µm d_f) coupled with a ^{2}D DB17 column (50% phenyl-methylpolysiloxane; 1.0 m × 0.10 mm d_c × 0.10 µm d_f), both supplied by Agilent Technologies (Wilmington, DE, USA). A fused silica capillary loop (1.0 m × 0.1 mm d_c) was used in the modulator slit. SilTite™ µ-unions (Trajan Scientific and Medical, Melbourne, Australia) were used to connect the columns with the capillaries.

The GC split/splitless injector port was set at 250 °C and operated in pulsed-split mode (250 kPa overpressure applied to the injection port until 2 min) with a 1:20 split ratio. A special design liner for SPME thermal desorption (Merck Lifesciences) was used to improve the transfer of the analytes to the ^{1}D column and to limit band broadening in space. Helium was used as the carrier gas at a nominal flow of 1.3 mL/min. The oven temperature program was set as follows: from 50 °C (1 min) to 260 °C (5 min) at 4.0 °C min^{-1}, to 270 °C (10 min) at 10 °C min^{-1}.

The *n*-alkanes solution for I^Ts determination was analyzed under the following conditions: split/splitless injector in split mode, split ratio: 1:50, injector temperature: 250 °C, and injection volume: 1 µL.

2.6. Combined untargeted and targeted (UT) fingerprinting by smart templates

A template is a pattern of 2D peaks and/or graphic objects (*i.e.*, features) generated over a reference chromatogram or image (single or cumulative image), which is used to identify similar patterns of 2D peaks in a set of analyzed chromatograms or images [10,16]. This procedure is guided by specific matching functions that establish correspondences between features across many chromatograms; specificity is achieved by defining confidence thresholds on retention times and MS spectral similarity to compensate for variable peak detection, and through suitable transform functions that successfully compensate for retention time inconsistencies between runs [17,18]. Once the correspondence between features is established, the template object metadata (*i.e.*, chemical name, retention times, mass spectra, informative ions and their relative ratios) are transferred to candidate peaks and/or graphic objects (*i.e.*, peak regions) in the analyzed chromatogram.

The feature template including untargeted reliable peaks and peak regions is obtained with a fully automated workflow in GC Image Investigator™ (GC Image™, GC Image LLC).

The workflow includes the following steps:

a. Peak patterns are matched between all chromatogram pairs from a set of representative samples. Matching constraints were set for this study based on optimized parameters validated by Squara *et al.* [18], *i.e.*, template peaks inclusion criterion S/N threshold of 50 or higher; peak reference spectra from *peak spectrum*; MS constraint set at 700 for the direct match factor (DMF) and 700 for the reverse match factor (RMF) based on the NIST similarity algorithm [19].
b. Reliable peaks across the set of analyzed chromatograms are then selected. The template of reliable peaks is used at the successive step for chromatogram re-alignment in the temporal domain. The criterion for inclusion was *relaxed* reliability [20] which selects 2D peaks that match at least 50% +1 of the chromatograms.
c. Chromatograms of representative samples are aligned consistently with the average retention times of the reliable peaks and then summed/fused into a composite chromatogram.
d. From the composite chromatogram, a comprehensive untargeted feature template is generated [20] including reliable

peaks and peak regions. These last are defined over the footprint of all detected peaks in the composite chromatogram.

This process was applied to generate composite class images from samples with common characteristics, e.g. same processing step. In particular, class images were built for raw sweet cream (*SCR*); pasteurized sweet cream inoculated with selected cultures and analyzed at time 0 and after 18 h (*PCT0* and *PCT18*); butter samples analyzed along ripening at 0–8–20–40 days (*BT0, BT8, BT20, BT40*). Acronyms and samples' characteristics are reported in Table 1 while a schematic diagram of butter making process is illustrated in Supplementary Figure 1. The data processing workflow is illustrated in Fig. 1 and further details are reported in Supplementary Figure 2. In the *Results and Discussion* section insights on the novel workflow are provided.

Identification of targeted features was by matching EI-MS fragmentation patterns collected from the peak apex (NIST MS Search algorithm, version 2.0 [19]) with those of commercial and in-house databases. Identification thresholds were 900 for DMF and 950 for RMF. A further constraint was applied for reliable identification with ±20 units of tolerance for experimental vs. tabulated linear retention indices (I^T) along the ^{1}D.

The list of targeted features is reported in Table 2 together with chemical names, ^{1}D and ^{2}D retention times (^1t_R; ^2t_R), experimental I^T, tabulated I^T, and odor quality as from reference literature.

2.7. Comparative visualization by visual features

Visual features fingerprinting, with pair-wise image comparison, is a comparative visualization tool that computes the difference at each data point (*i.e.*, the output of the detector at a point in time) between pairs of 2D chromatograms. Absolute or relative response differences are mapped into hue-intensity-saturation (HIS) color space to generate an image for visualizing the compositional differences between image pairs in the retention times plane. The approach enables comparative visualization with compositional differences rendered with an array of colorization modes.

In practice, the algorithm represents GC × GC raw data as an *a [m, n]* matrix, where *a* is the *analyzed* chromatogram with indexed pixels by ^{1}D retention time, *m*, and ^{2}D retention time, *n*. Each pixel corresponds to a detector data package and its colorization reveals differential detector response between image pairs (*analyzed* vs. *reference*) [7,16]. In this study, comparative visualization is by the grayscale *fuzzy difference* rendering. It subtracts pixel-by-pixel the *analyzed* image minus the *reference* image. The range of pixel values is mapped linearly to grayscale from black to white (0 to 255), with dark values (0–127) for negative differences, gray (value 128) for zero difference, and bright values (128–255) for positive differences [21]. This difference is non-zero only if the pixel in one image is outside the range of pixel values in a small neighbourhood in the other image.

2.8. Data acquisition and 2D data processing

Data were acquired by TOF-DS™ (Markes International, Llantrisant, UK) and processed by GC Image® suite, Release 2021r1 (GC Image, LLC, Lincoln NE, USA).

Data Processing was by a laptop with a standard processor Intel(R) Core™ i7–7500CPU@2.70 GHz, 2904 Mha, 2 core, 4 logic processors; RAM 16Gb; NVIDIA GeForce 940MX.

Statistical analysis and chemometrics were performed using Matlab R2021a (The MathWorks, Inc., Natick, Massachusetts, United States) with the following packages: PCA toolbox (v1.5) [22] and Classification toolbox (v6.0) [23], and XLSTAT statistical and data analysis solution software (Addinsoft 2020, New York, USA).

3. Results and discussion

The intriguing concept of Computer Vision to highlight compositional differences between sample pairs is not new in the scenario of data processing tools for C2DC [1,24]. It can be achieved by combining comparative visualization [7] with a comprehensive mapping of features, *i.e.*, *UT* fingerprinting [9]. However, even for uniform batch analyses, such as those acquired in a limited time frame without any change in the system hardware (*e.g.*, column set, modulation loop alignment, MS calibration and tuning), retention times misalignment might occur preventing consistent comparisons. An example of comparative visualization by colorized fuzzy difference of two slightly misaligned images is represented in Fig. 2A–C. In Fig. 2A the misalignment is along the ^{1}D in an analysis with a retention offset of 4.05% in the γ-terpinene region, Fig. 2B shows a misalignment in the ^{2}D with an offset of 15.71% in the 3-methyl butanoic acid elution region, and Fig. 2C shows a misalignment in both dimensions. Comparative visualization, if supervised by an analyst could lead to coherent conclusions due to experience and knowledge. If Computer Vision is systematized in a decision-making process, misinterpretation could occur deriving wrong conclusions about the actual compositional differences between samples pair.

In this scenario, approaches that examine the response recorded for peak and/or peak region features are preferable. By applying retention times tolerance windows for features alignment and matching, correct correspondences can be established. Additional specificity is also achieved by including MS similarity constraints above a pre-defined threshold [25]. However, while peak and peak region patterns can be re-aligned and rationalized in peak tables for data mining/data analysis, chromatogram images cannot.

To solve this issue and enable an effective and reliable CV, chromatogram alignment and registration in the temporal domain are mandatory. Reichenbach and co-workers [10,26] and Schmarr et al. [27] described similar approaches for the geometrical alignment of features across many chromatograms. Both methodologies adopt peak region features obtained by the alignment and combination of source chromatograms into a single composite chromatogram (*e.g.*, simply by addition or other fusion operations [27]). The feature template, generated by outlining footprint regions around each detected peak in the composite chromatogram, is then used to geometrically remap each source chromatogram. To note, the transform operation to align a chromatogram for compositing is bidirectional and can be used to map the feature template back to the source chromatogram. An example of the transform operation from the template to the chromatogram image is illustrated in Fig. 2D–F where for the different misalignments described above, the template is matched to the correct peak feature by its transform.

The alignment consistency is improved if the feature template matching is guided by appropriate transform models and additional constraints related to spectral similarity when the MS channel is available. Available transformation models include the global affine transform [28], capable of compensating for linear shifts in both chromatographic dimensions, and the global, low-degree polynomial alignment transformations which have better performances with non-linear and/or severe misalignments [29]. Both these algorithms are available in commercial software (GC Image™, GC Image LLC) as those proposed by Pierce et al. [30] and Zhang et al. [31]. Other approaches, not yet available in commercial data processing platforms for C2DC, include the piecewise linear interpolation/extrapolation for ^{1}D shifts [32]; natural-neighbor interpolation for the ^{2}D [33]; semi-parametric warping [34]; and BARCHAN that uses a non-rigid transformation in the first dimension and a rigid transformation in the second dimension [35].

Table 2
List of targeted features together with ^{1}D and ^{2}D retention times ($^{1}t_R$; $^{2}t_R$), experimental I^T and tabulated I^T, odor quality as from reference literature. Identification criteria: (a) by authentic reference compounds or (b) tentative identification combining $I^T \pm 10$ and spectral similarity direct match factor ≥900.

Compound Name	Identification	$^{1}t_R$ (min)	$^{2}t_R$ (sec)	Experim. I^T	Tabulated I^T	Odor quality
Hexane	a	4.38	0.54	600	/	/
Heptane	a	4.78	0.7	700	/	/
Dimethyl sulfide	b	4.81	0.5	729	754	
Cyclohexane	a	4.9	0.72	811	766	/
Triethylamine	a	5.13	0.76	821	780	Fishy
Octane	a	5.48	1	835	/	/
Acetone	a	5.72	0.54	845	826	Solvent
1-Chlorobutane	b	6.07	0.72	860	842	/
4-Methyloctane	b	6.24	1.28	868	850	/
Butanal	a	6.59	0.64	883	867	Cocoa
Acetic acid ethenyl ester	b	6.72	0.65	887	890	
Ethyl acetate	a	6.83	0.66	893	884	Ethereal
Nonane	a	6.88	1.46	895	/	/
2-Butanone	b	7.05	0.56	907	907	
3-Methylbutanal	a	7.35	0.76	915	917	Aldehydic, fruity
Ethanol	a	7.58	0.5	925	926	Alcoholic
1-Nonene	b	7.82	1.42	935	931	/
2,2,4,6,6-Pentamethylheptane	b	8.17	1.9	950	954	/
Butyl ether	a	8.34	1.46	957	963	Ethereal
3-Methylnonane	b	8.34	1.82	957	966	/
Diacetyl	a	8.87	0.6	979	980	Buttery
Decane	a	9.16	2.02	992	/	/
Isobutyl acetate	a	9.74	0.92	1012	1013	Fruity
α-Pinene	a	9.92	1.66	1018	1023	Herbal, woody
Tetrachloroethylene	b	9.98	1.12	1020	1024	/
Chloroform	a	10.03	0.6	1021	1024	/
1-Propanol	a	10.33	0.52	1030	1031	Alcoholic
Ethyl butanoate	a	10.44	1.08	1034	1042	Fruity
1-Decene	b	10.44	1.88	1034	1044	/
Camphene	a	11.26	1.82	1060	1060	Woody, camphor
Isopentyl acetate	b	11.67	1.12	1080	1080	Fruity
Hexanal	a	11.9	0.98	1083	1083	Grassy, apple
Isobutyl alcohol	a	12.02	0.54	1091	1085	Ethereal
Undecane	a	12.25	2.44	1091	/	/
β-Pinene	a	12.66	1.9	1103	1109	Herbal
2-Pentanol	a	13.01	0.62	1112	1113	Mild, musty
Sabinene	a	13.18	1.68	1117	1121	Woody
1-Methoxy-2-propanol	b	13.36	0.62	1122	1126	/
(E)−2-Pentenal	a	13.53	0.84	1126	1127	Green, apple
4-Methyl-3-penten-2-one	b	13.65	0.92	1129	1131	Pungent, potato
p-Xylene	a	13.71	1.2	1131	1136	/
1-Butanol	a	13.88	0.58	1136	1138	Fermented, fruity
m-Xylene	a	13.94	1.14	1137	1141	/
δ−3-Carene	a	14.18	1.86	1143	1147	Citrus
3-Heptanone	a	14.41	1.18	1150	1148	Green, ketonic
3-Methylundecane	b	14.7	2.62	1157	/	/
2-Methylpropyl butanoate	b	14.7	1.4	1157	1158	Fruity
α-Phellandrene	a	14.76	1.74	1159	1160	Terpenic
β-Myrcene	a	14.76	1.52	1159	1162	Peppery, woody
2-Methoxy-1-propanol	b	14.76	0.68	1159	/	/
Pyridine	a	15.17	0.8	1170	1175	Fishy
α-Terpinene	a	15.34	1.76	1174	1177	/
2-Heptanone	a	15.46	1.1	1178	1178	Fruity, sweet
Heptanal	a	15.63	1.18	1182	1181	Oily, fatty
2-Ethyl hexanal	a	15.69	1.34	1184	1188	/
Methyl hexanoate	a	15.75	1.24	1185	1188	Fruity
Dodecane	a	15.93	2.74	1190	1200	/
Limonene	a	16.04	1.74	1193	1202	Terpenic, herbal
3-Methyl-1-butanol	a	16.22	0.6	1198	1207	Fermented, fusel
1,8-Cineole	a	16.39	1.84	1202	1211	Herbal, minty
2-Hexanol	a	16.74	0.68	1211	1220	Chemical, winey
(E)−2-Hexenal	a	16.86	0.94	1215	1223	Bitter almond, green
2-Pentylfuran	a	17.38	1.3	1228	1235	Fruity, green
β-Ocimene	a	17.44	1.58	1230	1238	Floral, green
Ethyl hexanoate	a	17.5	1.44	1231	1240	Sweet, fruity
1-Dodecene	b	17.62	2.28	1234	1242	/
γ-Terpinene	a	17.85	1.78	1240	1249	/
1-Butoxy-2-ethylhexane	b	18.14	2.62	1248	/	/
3-Octanone	a	18.2	1.3	1249	1250	Fresh, mushroom
Styrene	a	18.32	1.04	1252	1257	Balsamic, gasoline
Isoamyl butyrate	b	18.73	1.58	1263	1267	Fruity, waxy
Hexyl acetate	a	19.02	1.3	1271	1274	Fruity
4-Heptanol	b	19.13	0.78	1274	1281	Alcoholic
α-Terpinolene	a	19.25	1.82	1277	1283	Fresh, woody
1,3,5-Trimethylbenzene	b	19.25	1.34	1277	1285	/
3-Hydroxy-2-butanone (acetoin)	a	19.54	0.58	1284	1287	Buttery, creamy
Octanal	a	19.6	1.24	1286	1291	Fatty, sharp
Tridecane	a	19.72	2.86	1289	1300	/

(*continued on next page*)

Table 2 (*continued*)

Compound Name	Identification	^1t_R (min)	^2t_R (sec)	Experim. I^T	Tabulated I^T	Odor quality
2-Ethyl-2-hexenal	a	19.95	1.24	1295	1314	/
Hexanenitrile	b	20.01	0.9	1296	1305	/
2-Heptanol	a	20.48	0.74	1309	1318	Fresh, fruity
3-Methyl-2-buten-1-ol	b	20.59	0.6	1312	1321	Fruity, sweet
N,N-Dimethyl formamide	b	20.65	0.78	1313	1326	/
(Z)−2-Hexenal	a	20.88	1.02	1319	1326	Leaves, cut grass
α-Methyl styrene	b	21.12	1.18	1325	1327	/
6-Methyl-5-hepten-2-one	a	21.41	1.12	1333	1339	Fatty, green
1-Hexanol	a	21.76	0.68	1342	1340	Fruity, banana
4-Hydroxy-4-methyl-2-pentanone	b	22.34	0.68	1358	1366	/
Dipropyl disulfide	b	22.87	1.64	1371	1372	Sulfury. earthy
2-Nonanone	a	23.33	1.32	1384	1390	Fruity, cheesy
Nonanal	a	23.51	1.3	1388	1397	Fatty, waxy
Tetradecane	a	23.54	2.86	1388	1400	Mild, waxy
2-Butoxyethanol	a	23.74	0.72	1395	1403	/
α-Thujone	a	24.68	1.4	1420	1421	Thujonic
(E)−2-Octenal	a	24.79	1.08	1423	1428	Green, nut
Benzene, 1-ethenyl-3-ethyl-	b	25.01	1.01	1427	1424	
p-Cymenene	b	25.08	1.28	1431	1430	Phenolic, spicy
Ethyl octanoate	a	25.08	1.58	1431	1431	Fruity, waxy
β-Thujone	a	25.38	1.3	1439	1440	Thujonic
1-Heptanol	a	25.55	0.72	1443	1444	Herb
1-Octen-3-ol	a	25.78	0.82	1449	1450	Mushroom, herbal
(E,E)−2,4-Heptadienal	a	26.08	0.9	1457	1465	Fatty, rancid
Furfural	b	26.13	0.66	1459	1466	Bready, brown
Butyl-2-ethylhexanoate	b	26.25	2	1462	1471	/
Acetic acid	a	26.72	0.4	1475	1479	Sour, vinegary
Decanal	a	27.36	1.42	1492	1490	Sweet, waxy
2-Nonanol	a	27.88	0.86	1507	1508	Waxy
(E,E)−3,5-Octadien-2-one	b	28	1	1510	1518	Geranium-like
Camphor	a	28.06	1.44	1512	1512	Camphoreus
Benzaldehyde	a	28.23	0.86	1517	1521	Fruity, sweet
1,3-Butanediol	a	28.53	0.5	1525	1540	/
(E)−2-Nonenal	a	28.64	1.14	1528	1532	Paper-like, fatty
Linalool	a	28.93	0.86	1537	1537	Floral, citrus
1-Octanol	a	29.23	0.78	1545	1547	Nut, mushroom
5-Methyl-2-furancarboxaldehyde	a	29.51	1.05	1548	1570	Caramel-like
2,3-Butanediol	b	29.75	0.5	1560	1569	Creamy
2-Undecanone	a	30.8	1.42	1590	1592	Fruity, waxy
Undecanal	a	31.03	1.48	1597	1606	Waxy, aldehydic
Benzonitrile	b	31.15	0.8	1600	1597	/
2-(2-ethoxyethoxy)-ethanol	b	31.38	0.7	1607	1599	/
Butanoic acid, 4–chloro-	b	31.85	0.74	1621	/	/
Menthol	a	32.08	0.96	1628	1632	Mentholic, cooling
Butyrolactone	b	32.15	1.05	1629	1632	
(E)−2-Decenal	a	32.32	1.22	1635	1642	Fishy, fatty
Acetophenone	a	32.55	0.92	1642	1645	Floral, powdery
Methyl-2-octynoate	a	32.61	1.16	1644	/	/
Butanoic acid	a	32.84	0.44	1650	1647	/
3-Methylbutanoic acid	a	33.75	0.58	1668	1666	Cheese, sweat
α-Terpineol	a	34.01	0.98	1685	1692	Terpenic, citrus
γ-Hexalactone	a	34.3	0.88	1694	1703	Herbal, coconut
Dodecanal	a	34.53	1.52	1701	1709	Soapy, waxy, citrus
Naphthalene	a	35.47	1.14	1730	1740	Pungent
(E)−2-Undecenal	a	35.88	1.26	1743	1750	Fruity, waxy
α,α-Dimethylbenzenemethanol	b	35.99	0.76	1747	1766	/
Methoxy phenyl oxime	b	35.99	0.44	1747	/	/
δ-Hexalactone	a	37.1	0.9	1781	1893	Creamy, chocolate
1,3-Propanediol	b	38.21	0.6	1792	1789	
Butyl benzoate	a	39.43	1.28	1857	1870	Balsamic, fruity
Hexanoic acid	a	40.02	0.48	1876	1880	Cheesy, fatty
Dimethyl sulfone	b	40.37	0.56	1887	1895	/
Benzene ethanol	b	40.66	0.7	1897	1906	Floral
γ-Octalactone	a	40.95	0.96	1907	1925	Coconut, lactonic
Diethylene glycol	a	42.29	0.52	1952	1968	/
δ-Octalactone	a	42.35	1.04	1954	1946	Coconut, creamy
1-Dodecanol	b	42.58	1.28	1960	1966	
Biphenyl	b	43.05	1.22	1978	1986	Green, floral
Heptanoic acid	a	43.52	0.48	1994	1997	Rancid, cheesy
γ-Nonalactone	a	44.22	1.02	2019	2038	Fatty, coconut
Octanoic acid	a	46.55	0.5	2105	2088	Rancid, soapy
2-Phenoxyethanol	b	47.25	0.68	2130	2110	Floral
δ-Decalactone	a	48.71	1.12	2181	2160	Coconut, peach
Nonanoic acid	a	49.58	0.52	2213	2197	Fatty, waxy, cheesy
Farnesol	a	51.16	1.66	2276	2293	Floral, fresh
Decanoic acid	a	52.44	0.54	2328	2309	Soapy, waxy
δ-Undecalactone	b	52.25	1.58	2340	2341	Sweet, coconut
1(3H)-Isobenzofuranone	b	52.85	0.84	2344	2356	/
γ-Dodecalactone	b	53.06	1.2	2364	2365	Creamy, sweet
Diethyl Phthalate	a	53.32	1.12	2363	2378	/
δ-Dodecalactone	a	54.6	1.2	2415	2420	Creamy, sweet
Dodecanoic acid	a	57.87	0.58	2554	2537	Fatty, soapy, coconut

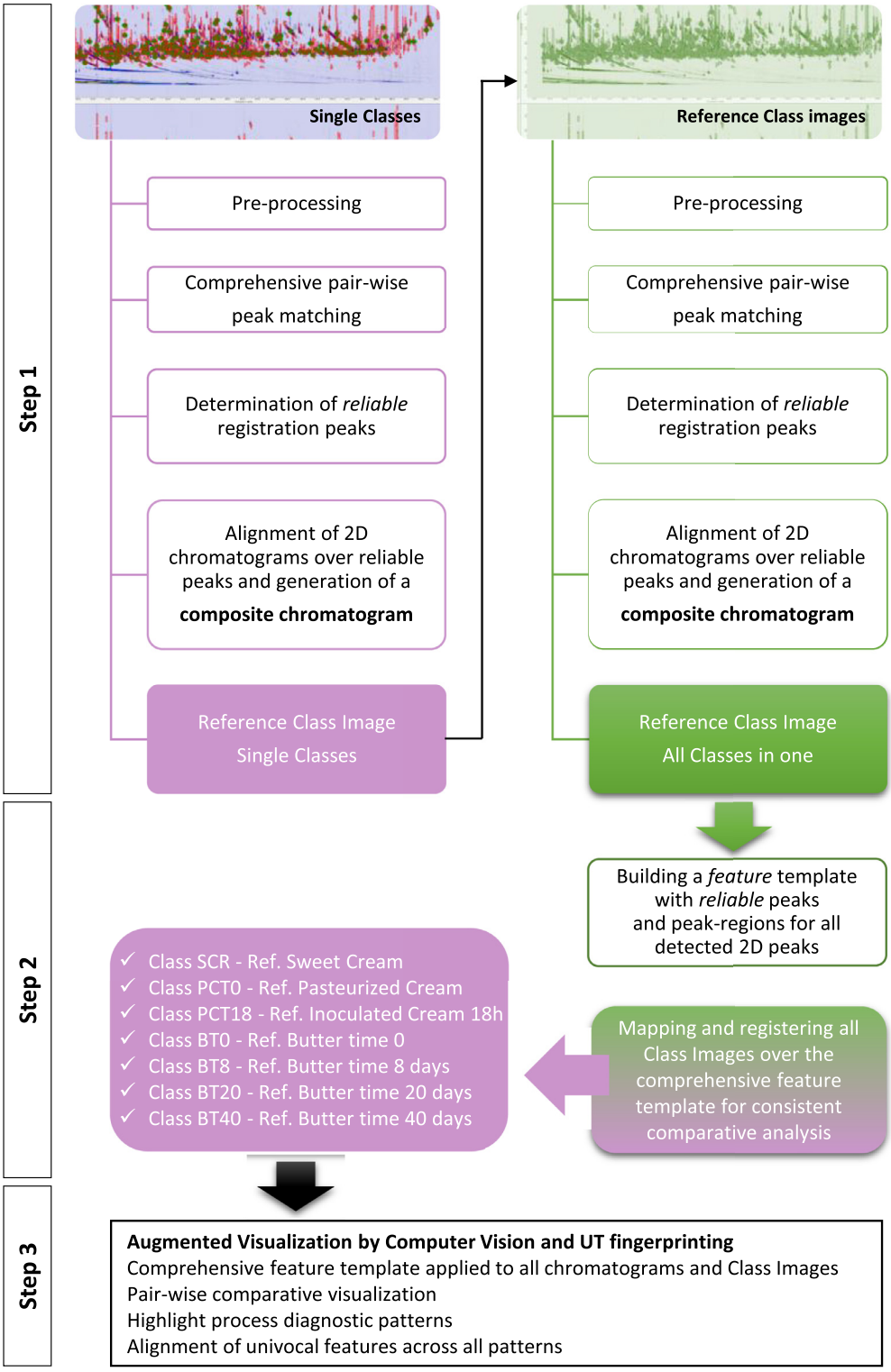

Fig. 1. Multistep CV data processing workflow: step 1 includes all the data pre-processing operations needed to create composite class images for representative samples' classes. Subsequently, the class images are effectively re-alignment and registered against a comprehensive feature template. Subsequently, by comparative visual analysis, samples are pairwise compared to highlight process diagnostic patterns after the alignment of univocal features across all patterns.

The next sections include an insight into the CV challenges faced approaching a large dataset of chromatograms as those collected along the butter-making process, followed by the description of the novel workflow for Augmented Visualization with molecular resolution. Results are presented and validated against a conventional peak feature approach.

3.1. Computer vision by datapoint features: chromatograms transform and registration

The dataset adopted to develop and validate the CV workflow included 84 chromatograms (7 butter processing steps × 2 inocula × 2 biological replicates × 3 analytical replicates) plus ad-

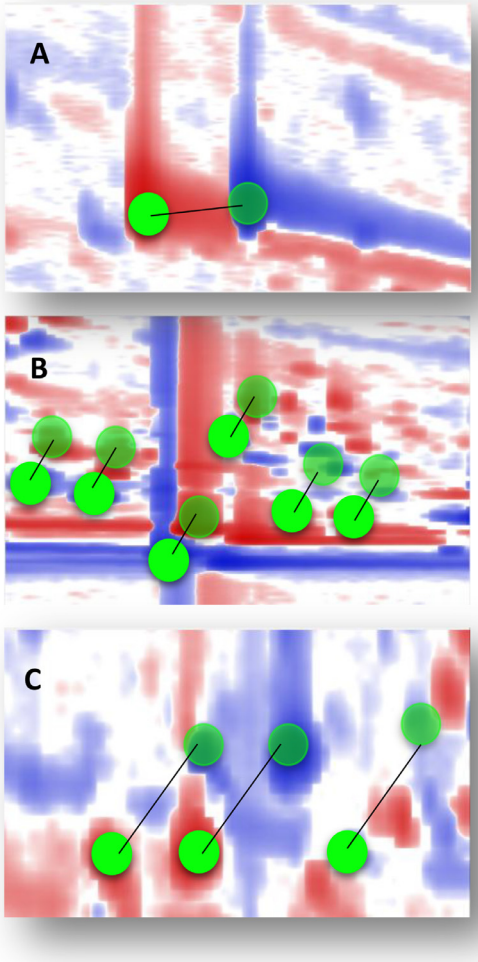

Fig. 2. (A–C) Comparative visualization as colorized fuzzy ratio of three regions of misaligned chromatograms with ^{1}D retention shift (A), ^{2}D retention shifts (B) and both (C). Colored circles indicate the 2D peaks centroids and their relative position in the two chromatograms. Black lines connect peaks of the same feature in the misaligned images and ideally illustrate how template transformation operates to re-align patterns.

ditional quality control samples (QCs) from a commercial butter sample analyzed to check for analytical system consistency over time (n=12). Once the alignment of QCs in terms of peak features response was examined, they were excluded by the CV workflow.

Peaks and peak regions in the comprehensive feature template (see Section 3.2 for template construction) were 346, of them 160 were putatively targeted according to criteria described in Section 2.6 and 186 were left untargeted by assigning them a unique identifier. The feature template was matched, transformed, and aligned over the 84 source chromatograms. Features metadata were collected in a spreadsheet for data evaluation and machine learning (Section 3.3).

Retention time fluctuations along the two chromatographic dimensions were at first examined since they would prevent a direct image comparison (cf. Fig. 2A–C). Along the ^{1}D, retention times accounted for a% relative standard deviation (%RSD) computed on their absolute values of 0.43 mean/0.25 median with a minimum value of 0.07 for γ-dodecalactone and a maximum of 5.82 for γ-terpinene. The non-normal distribution of $^{1}t_R$ values was connected to the retention variability/inconsistency of higher volatiles likely impacted by extra-chromatographic phenomena related to SPME fiber desorption kinetic into the GC injector. On the other hand, ^{2}D misalignment showed%RSD values of 8.77 mean/8.27 median with a minimum value of 0.1 for n-C14 (tetradecane) and a maximum of 16.7 for 3-methyl butanoic acid. This variability, as expected, was higher because of the simplified system configuration adopted that did not include a secondary oven isolated by the loop-type thermal modulator operating with liquid nitrogen as a coolant.

Fig. 3A shows an enlarged area of a sample's chromatogram (*B STD-T20* R2_rep #3) with the superimposed comprehensive feature template, containing peak features objects (red areas). By matching and transforming with polynomial second-order function (Fig. 3B) the template compensates for retention time shifts. To note, the enlarged area corresponds to the elution region of butanoic acid (larger peak) and 3-methylbutanoic acid (black arrow) that was affected by a severe misalignment across the chromatograms set.

Comparative visualization is highly dependent on chromatogram registration by feature template matching, otherwise it would fail in precisely informing about compositional differences. As shown in Fig. 3C without registration, several peak regions would be misaligned (yellow arrows); after registering, they were effectively aligned as shown in Fig. 3D.

After the temporal misalignment issue was solved, the possibility of achieving an Augmented Visualization with access to chromatogram metadata, *i.e.*, information about features' identity, was considered. The concept of AV refers to "*computational techniques for visualizing what cannot be seen with raw image input*" [11]. With the software platform here adopted, comparative visualization is linked to feature (data points and peaks) metadata. Therefore, once the information contained in the feature template is linked to the chromatographic image, comparative visualization is *augmented* by molecular information giving access to higher-level information about the phenomenon under study.

3.2. Augmented visualization by ut fingerprinting on composite class images: principle and workflow

The novel workflow was based on the UT fingerprinting automated process, a routine available in GC Image Investigator™ (GC Image™, GC Image LLC). The procedure is described in detail in the experimental section. In practice, samples were grouped in *classes* as a function of the processing stage: raw sweet cream (*SCR*); pasteurized sweet cream inoculated with selected cultures and analyzed at time 0 and after 18 h (*PCT0* and *PCT18*); butter samples analyzed along ripening at 0–8–20–40 days (*BT0, BT8, BT20, BT40*).

The process was carried out with the primary aim of samples' chromatogram re-alignment and combination into a composite class image representing all detected features in that class. The number of source chromatograms for each class was homogeneous, 6 chromatograms (2 biological replicates × 3 analytical replicates) were automatically pre-processed (file import, rasterization, colorization, baseline correction, 2D peaks detection and integration) with method parameters optimized by the analyst according to the sample compositional complexity, chromatographic resolution and absolute sensitivity. S/N threshold was set at 50 and baseline corrected according to Reichenbach *et al.* using the *default* algorithm [36]. Then, a comprehensive pair-wise peak matching was run for a total of 30 cross-matches [*i.e.*, (6 × 6)–6] and a computational time of about 30 min. The determination of reliable registration peaks was by most relaxed constraint and included 2D peaks that matched in at least 50% +1 of the chromatograms. This last operation took less than one minute and selected on average 1000 peaks per class. Note that the higher the number of anchor points for template matching, the better the image registration. The 2D chromatograms were then aligned according to the set of reliable

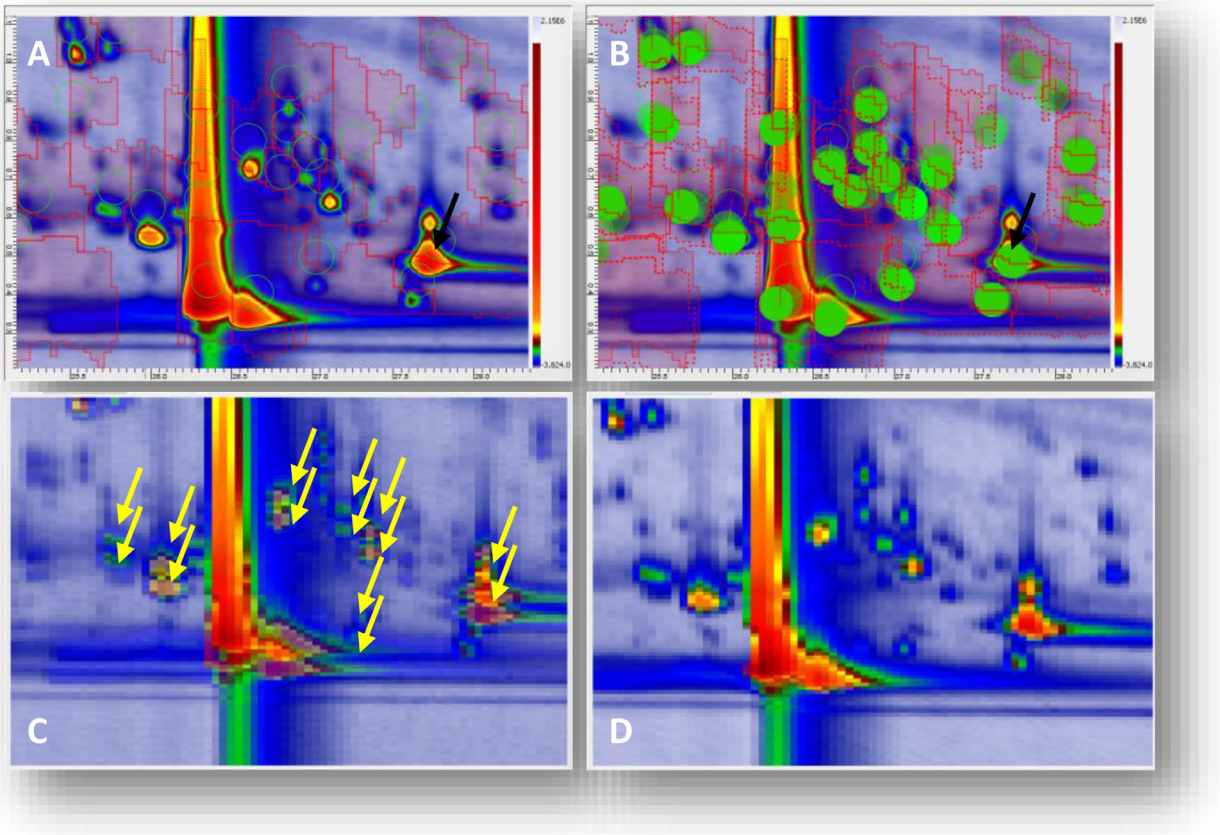

Fig. 3. Enlarged section of the B STD-T20 R2_rep n°3 2D chromatogram, between 25 and 28 min (^{1}D), 0.2 and 1.1 s (^{2}D). In Fig. 3A the feature template is over imposed and (3B) the retention time shifts corrected operated by the template matching algorithm. Comparative analysis between two misaligned chromatograms visualized in transparency before (3C) and after (3D) image registration.

registration peaks, transformed and registered before saving data, and making them available for combination into a composite image. For each chromatogram, this process took about 2 min. As the last step, the composite chromatogram (Class Image) was generated by the combination of the registered images in a single one, which was then processed for baseline correction and 2D peaks detection. The reliable template was matched and graphic objects (*i.e.*, areas corresponding to peak region features, shown with the red contour in Fig. 3A, B) were generated around all detected 2D peaks before their inclusion in the final feature template. The total computational time from raw data pre-processing until the final step with feature template composition was about 50 min per Class.

A schematic diagram of the workflow is illustrated in Fig. 1 (left side) while an in-depth scheme is provided in Supplementary Fig. 2.

The seven Class Images were then used to feed a new UT fingerprinting workflow (Fig. 1, right side) at this stage to re-align them against a comprehensive feature and vice-versa, in a way to obtain transformed and realigned Class Images for effective pair-wise comparisons. Moreover, the comprehensive feature template obtained at this stage was targeted by adding putative identifications for known components (see Section 2.6 for identification criteria) and matched back to all single chromatograms (i.e., single samples n=84 and Class Images n=7). This global re-alignment enabled AV with access to the higher-level information and traditional data mining by machine learning on the complete data set (i.e., 84 chromatograms × 346 features).

The next section illustrates as proof of concept some investigation possibilities offered by AV with registered Class Images. The flexibility of the workflow is so high that the analyst can answer many different questions just by properly designing the Class Images composition and consequent pair-wise matches.

3.3. Molecular patterns in butter processing

The volatile fraction of butter and butter oil has been the object of several studies since the mid-1950s; one of the most complete lists of known components (*i.e.*, targeted volatiles) accounts for 287 volatiles [37]. Within them, the sub-group of potent odorants and key-aroma compounds play a crucial role in consumers' acceptability and preferences [38]. However, many other volatiles contribute to the definition of butter quality. For example, terpenes and terpenoids, transferred from cow feeding to milk and sweet cream, inform about feeding systems (e.g., pasture or indoor feeding [39]), esters naturally formed during fermentation and cream acidification, are characteristic of the inocula, since precursors (*i.e.*, primary and secondary alcohols and acids) are formed by specific metabolic pathways [40]. Lactones, formed by the intra-esterification of hydroxyl acids [41], leave a diagnostic signature on many dairy products.

The evolution of the detectable volatilome along the butter production chain was examined by sequential pair-wise comparisons of Class Images. AV captured the compositional differences of products against the primary material, *i.e.*, the raw sweet cream. By these approaches, volatiles with increasing or decreasing trends are

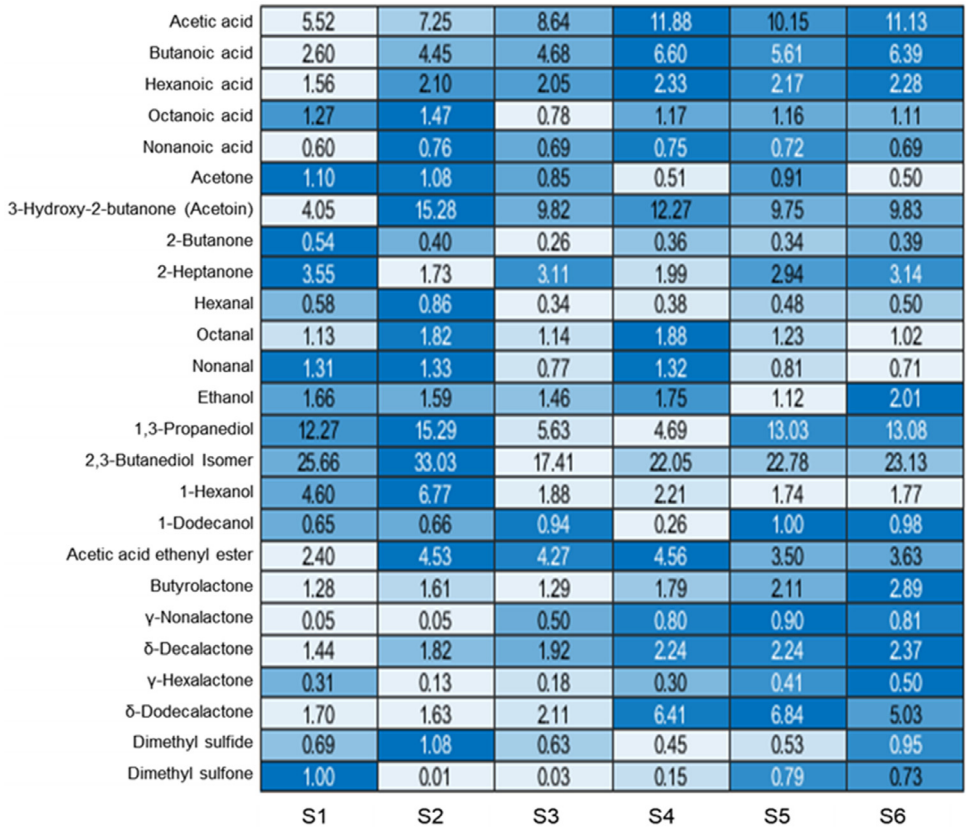

Fig. 4. Heatmap of target analytes showing the largest variations (indicated as fold change) relative to raw sweet cream. The heatmap colorization (light blue to dark blue) is scaled on the row values. Comparisons are between PCT0 vs. SCR – S1, PCT18 vs. SCR – S2, BT0 vs. SCR – S3, BT8 vs. SCR – S4, BT20 vs. – S5, BT40 vs. SCR – S6.

easily mapped over the chromatographic space thanks to specific colorizations, while metadata inspection enables access to higher-level information. Fig. 4 shows as a heat map the fold change for a selection of features that showed the largest variations within the processing steps. Fold change is calculated through the ratio of the target analytes' absolute responses from Class Images and taking the raw sweet cream Class Image as a reference. Comparisons are between *PCT0* vs. *SCR* (pasteurized cream at T0 vs. raw sweet cream – S1), *PCT18* vs. *SCR* (pasteurized cream at T18 vs. raw sweet cream – S2), *BT0* vs. *SCR* (butter at T0 vs. raw sweet cream – S3), *BT8* vs. *SCR* (butter at T8 vs. raw sweet cream – S4), *BT20* vs. *SCR* (butter at T20 vs. raw sweet cream – S5), *BT40* vs. *SCR* (butter at T40 vs. raw sweet cream – S6).

Analytes that showed the largest variations, with a progressive increment during processing are volatile acids (acetic acid, butanoic acid, and hexanoic acid) reaching a relative abundance in the ripened butter after 40 days, which corresponds to one order of magnitude higher than in the raw sweet cream. They are formed by both fermentation of milk primary metabolites and by enzymatic hydrolysis of triglycerides. Another analyte of interest for butter aroma identity is acetoin (3–hydroxy-2-butanone) [42,43]; it is formed from sugars through the pyruvate pathway by the lactic acid bacteria metabolism [44]. It reaches a maximum after inoculation of ferments in pasteurized raw cream (S2 comparison – 15.3 fold-change) and then stays stable along butter ripening. Interestingly, its reduction product, 2,3-butanediol, has a similar trend. Another group of analytes showing meaningful variations are lactones, they contribute to the *creamy* and *sweet* aroma of milk and dairy products [45]; γ-nonalactone, δ-decalactone and δ-dodecalactone have an incremental trend along processing.

AV facilitates the prompt capture of these information, Fig. 5A shows the comparative visualization rendered as *grayscale fuzzy ratio* between *PCT0* vs. *SCR* (pasteurized cream at T0 vs. raw sweet cream – S1). The yellow boxes over imposed on the chromatographic plane (zoomed area with green lines) highlight the peak region of acetoin (3–hydroxy-2-butanone). The access to relative and absolute responses, bottom tabular area, is facilitated and eventually guided by analysts with a mouse click on the chromatogram image. In Fig. 5B, another possibility is a comparative side-by-side visualization. The red box indicates the butanoic acid peak regions in the comparison between Class Images *BT40* vs. *SCR* (butter at T40 vs. raw sweet cream – S6).

Validation of the CV results was by conventional UT fingerprinting on the complete data matrix including 84 chromatograms $\times$ 28 features (having Fisher ratio values $>$ 8). Peak region features absolute responses for the three randomized replicated analyses were examined by Partial Least Square Discriminant Analysis (PLS-DA) after range scaling to highlight variables capable of classes discrimination (*i.e.*, *SCR, PCT0, PCT18, BT0, BT8, BT20*, and *BT40*). All PLS-DA models were cross-validated via Montecarlo validation excluding 20% of the dataset out of the model construction and using it as a validation set; this validation process was reiterated 1000 times to satisfy robustness requirements. PLS-DA results are visualized in Fig. 6A; from left to right processing steps are distributed over the Cartesian space delineated by the first two components. Sweet raw cream and pasteurized cream at t0 form clear clusters, due to an almost similar composition of their volatile fraction; after 18 h from the inoculum (*PCT18* – yellow indicators) the volatilome is evolving/changing and samples dispersion increases coherently with the logarithmic development of lactic bacteria. Then, after fat crystallization and butter ripening (*BT0* -> *BT40*), sample clusters show some overlapping. Of the most informative variables, according to Variable Importance for the Projections (VIPs), capable of effectively discriminating all classes (i.e.,

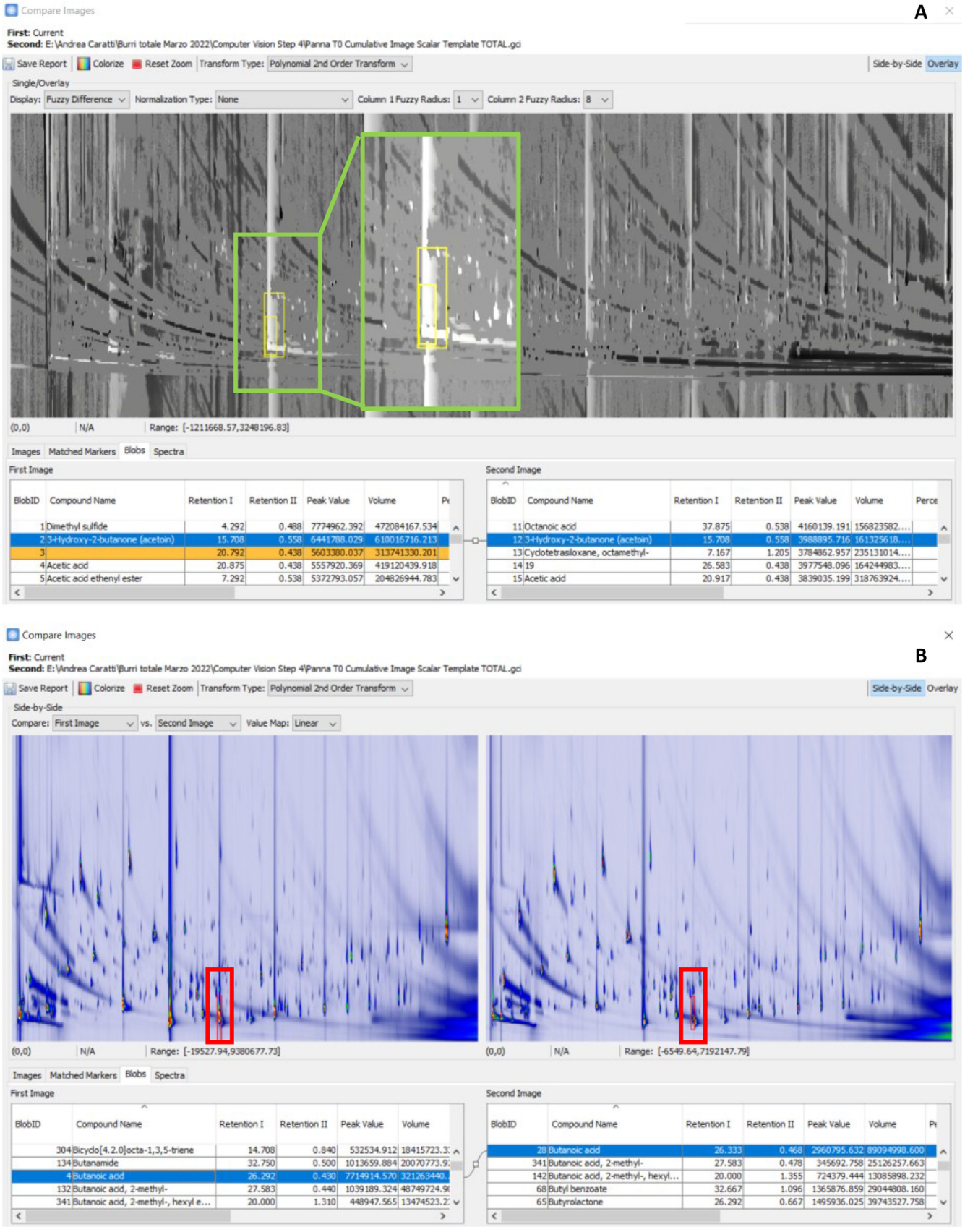

Fig. 5. (5A) Comparative overlayed visualization rendered as grayscale fuzzy ratio between PCT0 vs. SCR – S1. The peak region of 3–hydroxy-2-butanone (Acetoin – yellow boxes) is highlighted and zoomed within green areas, and its metadata in the two chromatograms are reported in the tabular area. (5B) Comparative side-by-side visualization between BT40 vs. SCR – S6. Butanoic acid peak regions with its metadata in both chromatograms is highlighted by red boxes.

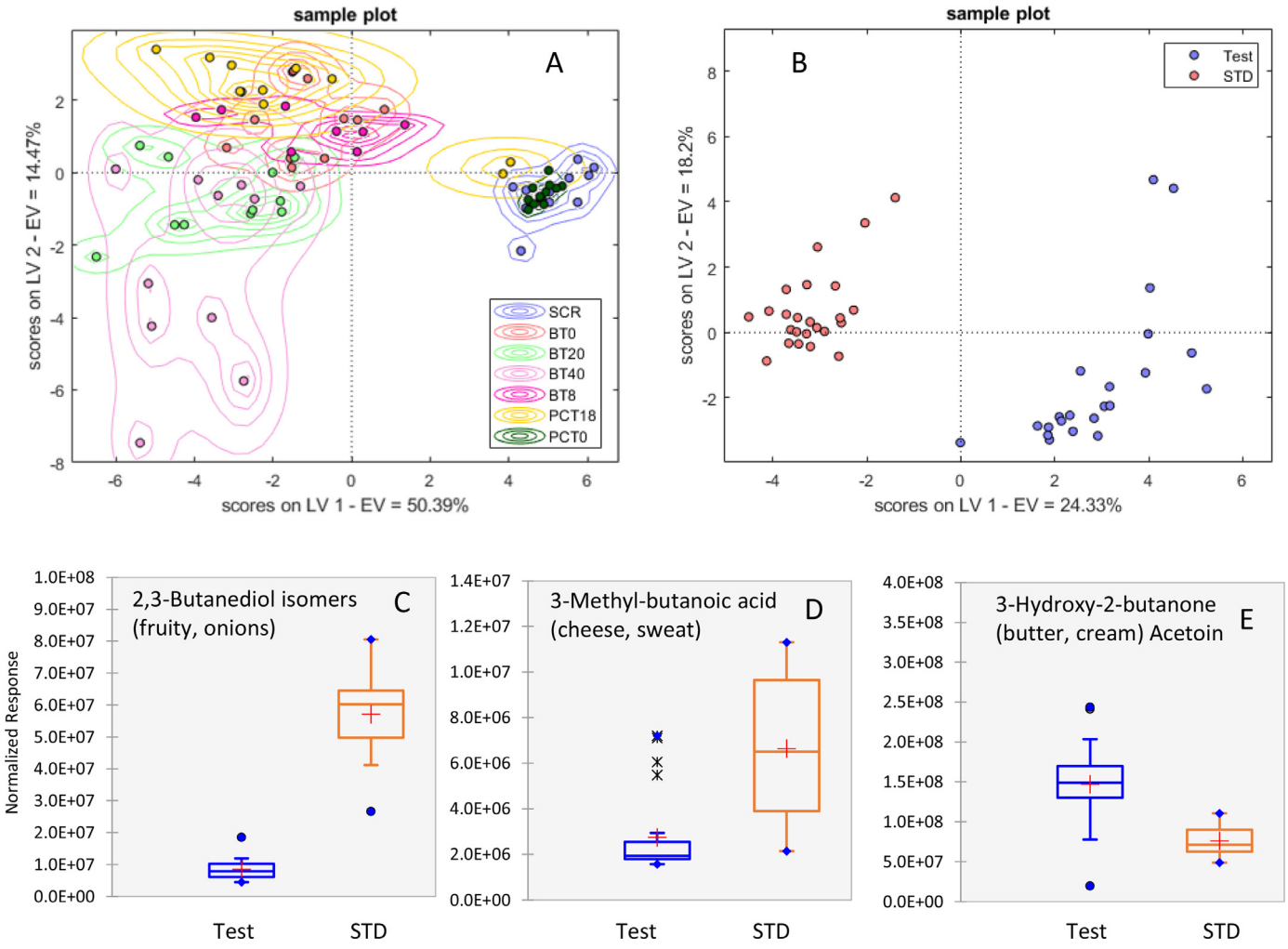

Fig. 6. (6A) PLS-DA (82 samples x 28 variables with Fisher ratio > 8) score plot on the first two latent variables of the various steps of artisanal butter production and ripening from raw sweet cream after range scaling as pre-processing step. Montecarlo validation 20% out was conducted for model validation and performance evaluation. (6B) PLS-DA (48 samples x 49 variables with Fisher ratio > 4) score plot on the first two latent variables of the two butter samples deriving from different microbial inocula (Test vs STD) after range scaling as pre-processing step. Montecarlo validation 20% out was conducted for model validation and performance evaluation. (6C, 6D, 6E) boxplots showing the absolute response distribution of the most impacting key odorants in Test vs STD butter samples.

VIPs ± SD >1), acetic acid ranked second after an untargeted feature (#115 not identified). They are then followed with decreasing VIPs values by: 1-heptanol; butanoic acid; propanoic acid; 1-octen-3-ol; feature #48; acetic acid ethenyl ester; feature #285; 3–hydroxy-2-butanone (acetoin); δ-decalactone; 3-methylbutanoic acid; hexanoic acid; γ-nonalactone; and 2,3-butanediol isomers. Most of them were promptly revealed by CV by Class Images comparison. PLS-DA in Fig. 6A was built with 7 latent variables (LVs), and achieved good sensitivity, specificity, and precision results; in particular, the best classification performances are achieved for the raw and pasteurized sweet cream, while worse discriminating capabilities are achieved when trying to discriminate butter samples according to shelf life. As reported in the confusion matrix and in the ROC curves (Supplementary Figure 3), the greatest mistakes happen within successive time points, this may be recollected to the fact that different samples in different physical areas of the butter are at different ripening stages due to inhomogeneous distribution of microbial in the matrix.

By further investigating the differential impact of the two inocula (*i.e.*, Test vs. STD), a PLS-DA was conducted on UT peak region features normalized responses from butter samples (from t0 to t40 days of ripening). Results are visualized in Fig. 6B where the sample classes are represented by pink and purple indicators. The model is guided by 28 UT features with a VIPs ± SD >1, of them, those with the highest odor impact are 2-heptanol; 2,3-butanediol isomers; 1,3-propanediol; octanal; 3-methylbutanoic acid; acetic acid; 3–hydroxy-2-butanone (acetoin); 1-octen-3-ol; 2-heptanone; 2-butanone; 5-methyl-2-furancarboxaldehyde; δ-undecalactone; and γ-butyrolactone. Key-odorants differential distribution is rendered in box-plots of Fig. 6C–E [2,3-butanediol isomers;3-methyl butanoic acid; and 3–hydroxy-2-butanone (acetoin)]. The second model created on two butter samples deriving by different microbial inocula (STD vs Test) shows optimal classification parameters as reported in Supplementary Figure 3. This confirms how the volatile fraction is heavily influenced by the microbial culture.

Augmented visualization results on Class Images were validated by conventional data mining (PLS-DA) on UT fingerprinting on single chromatogram Images. AV by combining data from different samples minimizes the effect of confounding/secondary variables (*e.g.*, sweet cream suppliers, inoculum bacteria, and timing) while facilitating the prompt capture of common features. Analytes with a clear increment along processing are also those with a classification role in the class model; they are acetic acid; butanoic acid; hexanoic acid; 3–hydroxy-2-butanone (acetoin); 2,3-butanediol isomers; acetic acid ethenyl ester; γ-nonalactone; and

δ-decalactone. Those which were not cross-validated have a secondary information role most related to the secondary variables' impact.

4. Conclusions

This study demonstrates how CV combined with chromatographic fingerprinting can effectively enable Augmented Visualization facilitating access to samples' compositional data and interpretation of the phenomena related to it. The capture of diagnostic patterns within the complex fraction of volatiles from food (food volatilome) is complicated by the concurrent effect of many biological phenomena which could hide informative patterns.

The multi-step workflow proposed is successful in realigning and combining chromatograms from samples belonging to specific classes, pre-defined based on the investigation questions. Once generated, Class Images can be pairwise compared with easy access to peak feature metadata. As a bench test, the butter volatilome was investigated for its evolution along the production chain. Besides the presence of confounding variables, stepwise comparisons promptly revealed analytes and analyte patterns with increasing trends from primary material (i.e. sweet cream) to the final product (i.e., ripened butter at 40 days).

The workflow is highly flexible, reliable, and robust by compensating for temporal misalignments occurring in large time-frame studies. Validation against a conventional approach based on UT fingerprinting on peak and peak region features confirms its consistency, supports its adoption as decision-making tool in food quality applications, and suggests its application in many other fields.

In a scenario where AI algorithms can support decisions in many fields, the high-resolution separation and multi-level information provided by C2DC are ideally suited to the Augmented Visualization processes.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests.

CRediT authorship contribution statement

Andrea Caratti: Formal analysis, Data curation, Validation, Writing – original draft, Writing – review & editing. **Simone Squara:** Formal analysis, Data curation, Validation, Writing – original draft, Writing – review & editing. **Carlo Bicchi:** Conceptualization, Supervision, Writing – review & editing. **Qingping Tao:** Conceptualization, Software, Writing – review & editing. **Daniel Geschwender:** Software, Writing – review & editing. **Stephen E. Reichenbach:** Conceptualization, Software, Writing – review & editing. **Francesco Ferrero:** Writing – review & editing. **Giorgio Borreani:** Funding acquisition, Project administration, Writing – review & editing. **Chiara Cordero:** Funding acquisition, Conceptualization, Supervision, Writing – original draft, Writing – review & editing.

Data availability

Data will be made available on request.

Acknowledgments

This work was carried out within the "Milk Zero" Project: Effetti della gestione delle produzioni nell'azienda zootecnica sulla presenza di contaminanti nei prodotti lattiero-caseari: Studio di una filiera a residuo zero in Piemonte. Bando di ricerca: PSR 2014–2020 - 16.2.1 - Az. 2 - Progetti pilota - Piattaforma tecnologica bioeconomia.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.chroma.2023.464010.

References

[1] F. Stilo, C. Bicchi, A.M. Jimenez-Carvelo, L. Cuadros-Rodriguez, S.E. Reichenbach, C. Cordero, Chromatographic fingerprinting by comprehensive two-dimensional chromatography: fundamentals and tools, TrAC Trends Anal. Chem. 134 (2021) 116133, doi:10.1016/j.trac.2020.116133.
[2] S.E. Reichenbach, X. Tian, C. Cordero, Q. Tao, Features for non-targeted cross-sample analysis with comprehensive two-dimensional chromatography, J. Chromatogr. A 1226 (2012) 140–148, doi:10.1016/j.chroma.2011.07.046.
[3] F. Stilo, C. Bicchi, A. Robbat, S.E. Reichenbach, C. Cordero, Untargeted approaches in food-omics: the potential of comprehensive two-dimensional gas chromatography/mass spectrometry, TrAC Trends Anal. Chem. 135 (2021) 116162, doi:10.1016/j.trac.2020.116162.
[4] B.A. Parsons, L.C. Marney, W.C. Siegler, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-based fisher ratio analysis of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry (GC × GC-TOFMS) data using a null distribution approach, Anal. Chem. 87 (2015) 3812–3819, doi:10.1021/ac504472s.
[5] L.C. Marney, W.Christopher Siegler, B.A. Parsons, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-based Fisher-ratio software for improved feature selection analysis of comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry data, Talanta 115 (2013) 887–895, doi:10.1016/j.talanta.2013.06.038.
[6] P.E. Sudol, G.S. Ochoa, R.E. Synovec, Investigation of the limit of discovery using tile-based Fisher ratio analysis with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry, J. Chromatogr. A. (2021), doi:10.1016/j.chroma.2021.462092.
[7] B.V. Hollingsworth, S.E. Reichenbach, Q. Tao, A. Visvanathan, Comparative visualization for comprehensive two-dimensional gas chromatography, J. Chromatogr. A 1105 (2006) 51–58, doi:10.1016/j.chroma.2005.11.074.
[8] What is computer vision?, Website: https://research.ibm.com/topics/computer-vision. Accessed April 21, 2023.
[9] F. Magagna, L. Valverde-Som, C. Ruíz-Samblás, L. Cuadros-Rodríguez, S.E. Reichenbach, C. Bicchi, C. Cordero, Combined untargeted and targeted fingerprinting with comprehensive two-dimensional chromatography for volatiles and ripening indicators in olive oil, Anal. Chim. Acta 936 (2016) 245–258, doi:10.1016/j.aca.2016.07.005.
[10] S.E. Reichenbach, C.A. Zini, K.P. Nicolli, J.E. Welke, C. Cordero, Q. Tao, Benchmarking machine learning methods for comprehensive chemical fingerprinting and pattern recognition, J. Chromatogr. A 1595 (2019) 158–167, doi:10.1016/j.chroma.2019.02.027.
[11] S. Mori, H. Saito, An overview of augmented visualization: observing the real world as desired, APSIPA Trans. Signal Inf. Process. 7 (2018), doi:10.1017/ATSIP.2018.13.
[12] L. Cuadros-Rodríguez, C. Ruiz-Samblás, L. Valverde-Som, E. Pérez-Castaño, A. González-Casado, Chromatographic fingerprinting: an innovative approach for food "identitation" and food authentication – A tutorial, Anal. Chim. Acta 909 (2016) 9–23, doi:10.1016/j.aca.2015.12.042.
[13] Y. Wang, J. O'Reilly, Y. Chen, J. Pawliszyn, Equilibrium in-fibre standardisation technique for solid-phase microextraction, J. Chromatogr. A 1072 (2005) 13–17, doi:10.1016/j.chroma.2004.12.084.
[14] C. Cordero, C. Cagliero, E. Liberto, L. Nicolotti, P. Rubiolo, B. Sgorbini, C. Bicchi, High concentration capacity sample preparation techniques to improve the informative potential of two-dimensional comprehensive gas chromatography–mass spectrometry: application to sensomics, J. Chromatogr. A 1318 (2013) 1–11, doi:10.1016/j.chroma.2013.09.065.
[15] L. Nicolotti, C. Cordero, C. Cagliero, E. Liberto, B. Sgorbini, P. Rubiolo, C. Bicchi, Quantitative fingerprinting by headspace-Two-dimensional comprehensive gas chromatography-mass spectrometry of solid matrices: some challenging aspects of the exhaustive assessment of food volatiles, Anal. Chim. Acta 798 (2013) 115–125, doi:10.1016/j.aca.2013.08.052.
[16] D. Bressanello, E. Liberto, M. Collino, F. Chiazza, R. Mastrocola, S.E. Reichenbach, C. Bicchi, C. Cordero, Combined untargeted and targeted fingerprinting by comprehensive two-dimensional gas chromatography: revealing fructose-induced changes in mice urinary metabolic signatures, Anal. Bioanal. Chem. 410 (2018) 2723–2737, doi:10.1007/s00216-018-0950-9.
[17] F. Stilo, E. Liberto, S.E. Reichenbach, Q. Tao, C. Bicchi, C. Cordero, Untargeted and targeted fingerprinting of extra virgin olive oil volatiles by comprehensive two-dimensional gas chromatography with mass spectrometry: challenges in long-term studies, J. Agric. Food Chem. 67 (2019) 5289–5302, doi:10.1021/acs.jafc.9b01661.
[18] S. Squara, F. Manig, T. Henle, M. Hellwig, A. Caratti, C. Bicchi, S.E. Reichenbach, Q. Tao, M. Collino, C. Cordero, Extending the breadth of saliva metabolome

fingerprinting by smart template strategies and effective pattern realignment on comprehensive two-dimensional gas chromatographic data, Anal. Bioanal. Chem. (2023), doi:10.1007/s00216-023-04516-x.
[19] NIST/EPA/NIH Mass Spectral Library with Search Program: (Data Version: NIST 08, Software Version 2.0f) doi:10.18434/T4H594.
[20] S.E. Reichenbach, X. Tian, A.A. Boateng, C.A. Mullen, C. Cordero, Q. Tao, Reliable peak selection for multisample analysis with comprehensive two-dimensional chromatography, Anal. Chem. 85 (2013) 4974–4981, doi:10.1021/ac303773v.
[21] GC Image User Guide is available at the web link: http://cse.unl.edu/~reich/gcimage/index.html. Accessed April 21, 2023.
[22] D. Ballabio, A MATLAB toolbox for principal component analysis and unsupervised exploration of data structure, Chemom. Intell. Lab. Syst. 149 (2015) 1–9, doi:10.1016/j.chemolab.2015.10.003.
[23] D. Ballabio, V. Consonni, Classification tools in chemistry. part 1: linear models. PLS-DA, Anal. Methods 5 (2013) 3790–3798, doi:10.1039/c3ay40582f.
[24] F. Stilo, E. Liberto, N. Spigolon, G. Genova, G. Rosso, M. Fontana, S.E. Reichenbach, C. Bicchi, C. Cordero, An effective chromatographic fingerprinting workflow based on comprehensive two-dimensional gas chromatography – Mass spectrometry to establish volatiles patterns discriminative of spoiled hazelnuts (Corylus avellana L.), Food Chem. 340 (2021) 128135, doi:10.1016/j.foodchem.2020.128135.
[25] S.E. Reichenbach, P.W. Carr, D.R. Stoll, Q. Tao, Smart Templates for peak pattern matching with comprehensive two-dimensional liquid chromatography, J. Chromatogr. A 1216 (2009) 3458–3466, doi:10.1016/j.chroma.2008.09.058.
[26] C. Cordero, E. Liberto, C. Bicchi, P. Rubiolo, S.E. Reichenbach, X. Tian, Q. Tao, Targeted and non-targeted approaches for complex natural sample profiling by GC×GC-qMS, J. Chromatogr. Sci. 48 (2010) 251–261, doi:10.1093/chromsci/48.4.251.
[27] H.-.G. Schmarr, J. Bernhardt, Profiling analysis of volatile compounds from fruits using comprehensive two-dimensional gas chromatography and image processing techniques, J. Chromatogr. A. 1217 (2010) 565–574, doi:10.1016/j.chroma.2009.11.063.
[28] S.E. Reichenbach, D.W. Rempe, Q. Tao, D. Bressanello, E. Liberto, C. Bicchi, S. Balducci, C. Cordero, Alignment for comprehensive two-dimensional gas chromatography with dual secondary columns and detectors, Anal. Chem. 87 (2015) 10056–10063, doi:10.1021/acs.analchem.5b02718.
[29] D.W. Rempe, S.E. Reichenbach, Q. Tao, C. Cordero, W.E. Rathbun, C.A. Zini, Effectiveness of global, low-degree polynomial transformations for GCxGC data alignment, Anal. Chem. 88 (2016) 10028–10035, doi:10.1021/acs.analchem.6b02254.
[30] K.M. Pierce, L.F. Wood, B.W. Wright, R.E. Synovec, A comprehensive two-dimensional retention time alignment algorithm to enhance chemometric analysis of comprehensive two-dimensional separation data, Anal. Chem. 77 (2005) 7735–7743, doi:10.1021/ac0511142.
[31] D. Zhang, X. Huang, F.E. Regnier, M. Zhang, Two-dimensional correlation optimized warping algorithm for aligning GC×GC−MS data, Anal. Chem. 80 (2008) 2664–2671, doi:10.1021/ac7024317.
[32] J. Gros, D. Nabi, P. Dimitriou-Christidis, R. Rutler, J.S. Arey, Robust algorithm for aligning two-dimensional chromatograms, Anal. Chem. 84 (2012) 9033–9040, doi:10.1021/ac301367s.
[33] Y. Zushi, J. Gros, Q. Tao, S.E. Reichenbach, S. Hashimoto, J.S. Arey, Pixel-by-pixel correction of retention time shifts in chromatograms from comprehensive two-dimensional gas chromatography coupled to high resolution time-of-flight mass spectrometry, J. Chromatogr. A 1508 (2017) 121–129, doi:10.1016/j.chroma.2017.05.065.
[34] W.P.H. De Boer, J. Lankelma, Two-dimensional semi-parametric alignment of chromatograms, J. Chromatogr. A 1345 (2014) 193–199, doi:10.1016/j.chroma.2014.04.034.
[35] C. Couprie, L. Duval, M. Moreaud, S. Hénon, M. Tebib, V. Souchon, BARCHAN: blob Alignment for Robust CHromatographic ANalysis, J. Chromatogr. A 1484 (2017) 65–72, doi:10.1016/j.chroma.2017.01.003.
[36] S.E. Reichenbach, M. Ni, D. Zhang, E.B. Ledford, Image background removal in comprehensive two-dimensional gas chromatography, J. Chromatogr. A 985 (2003) 47–56, doi:10.1016/S0021-9673(02)01498-X.
[37] H. Maarse, Volatile Compounds in Foods and Beverages, Routledge, 2017, doi:10.1201/9780203734285.
[38] G. Reineccius, Flavor Chemistry and Technology, Second edition, CRC Press, Boca Raton, FL, 2005.
[39] T.F. O'Callaghan, H. Faulkner, S. McAuliffe, M.G. O'Sullivan, D. Hennessy, P. Dillon, K.N. Kilcawley, C. Stanton, R.P. Ross, Quality characteristics, chemical composition, and sensory properties of butter from cows on pasture versus indoor feeding systems, J. Dairy Sci. 99 (2016) 9441–9460, doi:10.3168/jds.2016-11271.
[40] S. Mallia, F. Escher, H. Schlichtherle-Cerny, Aroma-active compounds of butter: a review, Eur. Food Res. Technol. 226 (2008) 315–325, doi:10.1007/s00217-006-0555-y.
[41] S.R. Lee, C. Macku, T. Shibamoto, Isolation and identification of headspace volatiles formed in heated butter, J. Agric. Food Chem. 39 (1991) 1972–1975, doi:10.1021/jf00011a017.
[42] P. Schieberle, K. Gassenmeier, H. Guth, A. Sen, W. Grosch, Character impact odour compounds of different kinds of butter, LWT Food Sci. Technol. 26 (1993) 347–356, doi:10.1006/fstl.1993.1070.
[43] D.G. Peterson, G.A. Reineccius, Characterization of the volatile compounds that constitute fresh sweet cream butter aroma, Flavour Fragr. J. 18 (2003) 215–220, doi:10.1002/ffj.1192.
[44] V.M. Dillon, Natural anti-microbial systems | preservative effects during storage, in: Encyclopedia of Food Microbiology, Elsevier, 2014, pp. 941–947.
[45] M. Preininger, W. Grosch, Evaluation of key odorants of the neutral volatiles of emmentaler cheese by the calculation of odour activity values, Leb. Und Technol. 27 (1994) 237–244, doi:10.1006/fstl.1994.1048.