Fused Adjacency Matrices to enhance information extraction: the beer benchmark

# Accepted Manuscript


Fused Adjacency Matrices to enhance information extraction: the beer benchmark

Nicola Cavallini, Francesco Savorani, Rasmus Bro, Marina Cocchi

PII: S0003-2670(19)30197-7

DOI: https://doi.org/10.1016/j.aca.2019.02.023

Reference: ACA 236587

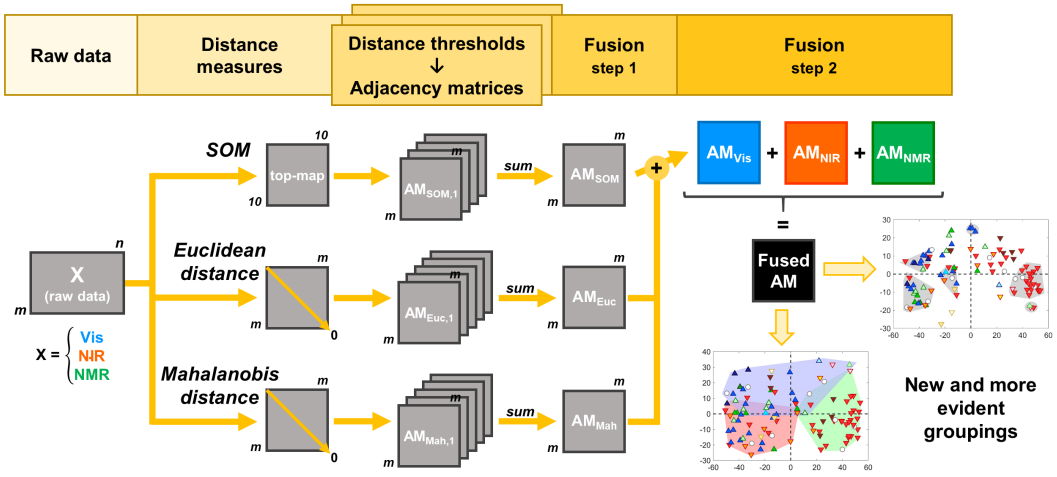
To appear in: *Analytica Chimica Acta*

Received Date: 11 May 2018

Revised Date: 31 January 2019

Accepted Date: 4 February 2019

Please cite this article as: N. Cavallini, F. Savorani, R. Bro, M. Cocchi, Fused Adjacency Matrices to enhance information extraction: the beer benchmark, *Analytica Chimica Acta*, https://doi.org/10.1016/j.aca.2019.02.023.



**ACA-18-1339R1 – underlined revision**

**Article title:** Fused Adjacency Matrices to enhance information extraction: the beer benchmark

**Authors and affiliations:** Nicola Cavallini[1,2], Francesco Savorani[3], Rasmus Bro[2], Marina Cocchi[1]

[1] Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio Emilia Via Campi 103 – 41125 Modena (MO) Italy

[2] Chemometrics and Analytical Technology, Department of Food Science, Faculty of Science, University of Copenhagen, Rolighedsvej 26, 1958 Frederiksberg C, Denmark

[3] Department of Applied Science and Technology, Politecnico di Torino, Corso Duca degli Abruzzi, 24 - 10129 Torino (TO) Italy

**Corresponding author:** Marina Cocchi (marina.cocchi@unimore.it)

Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio Emilia Via Campi 103 – 41125 Modena (MO) Italy

**Abstract**

Multivariate exploratory data analysis allows revealing patterns and extracting information from complex multivariate data sets. However, highly complex data may not show evident groupings or trends in the principal component space, e.g. because the variation of the variables are not grouped but rather continuous. In these cases, classical exploratory methods may not provide satisfactory results when the aim is to find distinct groupings in the data.

To enhance information extraction in such situations, we propose a novel approach inspired by the concept of combining weak classifiers, but in the unsupervised context. The approach is based on the fusion of several adjacency matrices obtained by different distance measures on data from different analytical platforms. This paper is intended to present and discuss the potential of the approach through a benchmark data set of beer samples. The beer data were acquired using three spectroscopic techniques: Visible, near-Infrared and Nuclear Magnetic Resonance.

The results of fusing the three data sets via the proposed approach are compared with those from the single data blocks (Visible, NIR and NMR) and from a standard mid-level data fusion methodology. It is shown that, with the suggested approach, groupings related to beer style and other features are efficiently recovered, and generally more evident.

**Keywords**

Data fusion, Adjacency Matrix, Clustering, Data visualization, Spectroscopy, Beer

**Abbreviations**

| | |
|---|---|
| AM | Adjacency Matrix |
| MCR | Multivariate Curve Resolution |
| OPTICS | Ordering Points to Identify the Clustering Structure |
| PC | Principal Component |
| PCA | Principal Component Analysis |

| | |
|---|---|
| RD | Reachability Distance |
| RP | Reachability Plot |
| SOM | Kohonen's Self-Organizing Map |
| Vis | Visible |

## 1. Introduction

Exploratory multivariate data analysis (EMDA, [1]) offers very powerful tools for looking into complex data. Using EMDA it is possible, for example, to reveal underlying structures and discover groups of similar samples and visualizing such patterns in an accessible and simple way.

Principal Component Analysis (PCA, [1–3]) is probably the most common EMDA approach, together with some variants (Maximum Likelihood PCA [4], Projection Pursuit PCA [5,6]) but other linear methods such as Independent Component Analysis **(**ICA**,** [7,8]**)** and Multidimensional Scaling **(**MDS, [1,9]**)** are also quite popular. Non-linear mapping methods like Kohonen's Self-Organizing Maps (SOMs, [10,11]) are considered complementary to methods like PCA [12], because of their ability to account for non-linear phenomena. All these techniques are called "projection" methods, since they are based on projecting the original high-dimensional data to a space of lower dimensions, which makes it easier to model, plot and visualize the data. Another, different way of recovering structures and groups of samples from data is represented by the clustering methods [13,14]. Dissimilarity (or similarity) is at the core of clustering, and it is often assessed using a distance measure, based on which linkage/grouping criteria are defined.

Despite the large variety of EMDA methods available, there are still cases in which it is difficult to obtain satisfactory results. Highly complex data may not show simple groupings and/or trends in the principal component space and may be so complex that normal visualizations are only shedding limited light on the underlying characteristics.

In this perspective, we propose what we define as a Fused Adjacency Matrix approach. The overall idea of the approach is to combine multiple "weak sources" of information that when combined will

yield more discriminatory information. This "combination" concept comes from the field of supervised learning, and more specifically from methods like Random Forest [15] or Weak Learning Algorithm [16], in which multiple weak classifiers are combined to make stronger class assignments [17,18]. Another strategy also used in the supervised context is to combine the results obtained by an ensemble of different classification methods [19,20]. In this context, several fusion rules were proposed [19–21] to combine the different classifiers/classifier outcomes. More recently, a fusion strategy for non-optimized classifiers was proposed, i.e. by considering a window of tuning parameters values for each classifier in the fusion process [22].

Our new approach shares both the ideas of combining outcomes from different methods and considering windows of parameters values, and it applies to the unsupervised framework with the aim of performing exploratory analysis. The approach consists of two steps, each one based on the fusion of adjacency matrices (AMs). In the first step different distance thresholds and metrics are used to compute several AMs, which are then fused using a sum rule, to obtain just a single matrix as an output. Once having performed this first step on different blocks of data (e.g. acquired by different analytical platforms) the resulting output matrices are then combined into the Fused Adjacency Matrix ($AM_{Fus}$ in Figure 1, step 2). This second step accomplishes the fusion of data sets obtained by several analytical techniques [23]. The proposed approach is intended as an unsupervised exploratory tool to better highlight grouping structure, but it can also be seen as a method for mid-level data fusion of clustering models.

The Fused Adjacency Matrix approach is presented using, as a benchmark, a real case dataset of analysis of beer samples. This dataset consists of three data blocks obtained from different spectroscopic techniques: Visible (Vis), Near Infra-Red (NIR) and Nuclear Magnetic Resonance (NMR), the latter as interval-resolved data. This data set represents a challenging benchmark to show the approach's potential, due to its potential richness in analytical information acquired, associated with its weak grouping structure and limited *a priori* knowledge (rather general such as beer style, alcohol content and colour). The beer samples were collected from supermarket, and the

general purpose was that of evidencing both peculiar beers and groups of similar samples, in other words mining all the possible similarities/peculiarities, just based on the chemical fingerprint acquired. Beer has been the object of several studies, mostly focused on a specific beer type aiming either at gathering the composition [24–27] or controlling the brewing process [28,29]. To achieve these aims very different analytical techniques have been applied: NMR [24,25,27,30–32], LC-MS [30,33,34], GC-MS [35,36], vibrational (NIR and IR) [24,26,28,37] and UV-Visible [38] spectroscopies. The benchmark beer dataset consists of three data blocks obtained from different spectroscopic techniques: Visible (Vis), Near Infra-Red (NIR) and Nuclear Magnetic Resonance (NMR), the latter as interval-resolved data.

The paper is organized as follows: Section 2 outlines how the data were obtained and what kind of data analysis tools were employed; a description of the Fused Adjacency Matrix approach is given in Section 2.2.5 and depicted in Figure 1; Section 3 reports the main results of the single datasets (Vis, NIR, NMR), the mid-level data fusion [39,40] and the Fused Adjacency Matrix approaches; more detailed comparisons and a summary are reported in Section 3.6, while comparisons among the different fusion steps are reported in Section 3.7 by means of Procrustes Analysis; finally, an example of how to link back the Fused Adjacency Matrix results to the original data is given in Section 3.8 using NMR as an example.

## 2. Materials and methods

Detailed information about each beer sample, such as beer styles, names, brands and production sites are given in Table S1, in the Supplementary Materials. The number of samples by yeast family and beer style are reported in Table S2.

### 2.1. Experimental

#### 2.1.1. Sampling and sample preparation

One hundred beer products were purchased from local stores. All were rather pale in colour and clear in the sense that there were no clearly visible particles suspended in the liquid. They differ by brand, location of production, percentage of alcohol by volume (ABV), colour and beer style. To make the interpretation of plots more straightforward, it was decided to gather some beer styles under the same "miscellaneous" label. In Figures 2–7, legend entries "Ales misc." and "Lagers misc." represent the following styles (in parentheses is reported the number of samples for each sub-style):

- miscellaneous Ales: ale (1), amber (1), Belgian (1), brown (1), English (1), red (1);
- miscellaneous Lagers: amber (2), amber/strong (1), Czech (4).

A collection of 2 mL eppendorfs was directly prepared from the original commercial containers (cans or glass bottles). Three eppendorfs for each beer sample were prepared and kept frozen at –20°C. The initial steps of thawing and degassing [24] were common across all the different spectroscopic techniques, and were performed as follows: 1) 10 minutes thawing in water bath at room temperature; 2) 20 minutes of ultrasonic bath in water at room temperature. Since all the specimens were clear (i.e. no suspended particles), filtration was not required. The degassing procedure is highly recommended by literature studies [24,25,27] and it is aimed at reducing measurement interferences due to bubble formation both on the NIR sample vessel and within the NMR tubes.

### 2.1.2. Vis-NIR data acquisition and preprocesing

Visible (Vis) and Near-Infrared (NIR) spectra were acquired together using a NIRS FOSS DS2500 spectrometer, in the range 400–2500 nm (0.5 nm resolution). A cup with a round quartz window was equipped with a 0.2 mm-gap golden reflector to operate in transflectance mode. Each spectrum was obtained by taking the average over 16 scans acquired at different positions of the cup's window. No additional steps to the preparation procedure described in Section 2.1.1 were necessary prior to recording the Vis-NIR spectra. The specimens were prepared in batches of 25 samples and

placed right after processing inside a thermally insulated styrofoam box, equipped with ice chips and a lid. This setup was made to keep the specimens in stable conditions while running the experiments.

For each sample three replicates were acquired but the order of acquisition was randomized both with respect to samples and replicates. A control sample for each batch was also prepared under the same conditions as the other specimens. A pack of six canned beers was purchased from a local store and kept in a fridge at 4°C. Right before preparing a batch, the eppendorfs were filled with fresh beer. This allowed checking for time drifts among different batches, since they were analysed at different time points.

Similarity among replicates was assessed by performing a Principal Component Analysis on the data centered with respect to replicates, i.e. subtracting from each sample the average of its replicates: the first principal component explained 88.33% of the total variance, and the anomalous spectra were identified as the ones far exceeding the scores confidence limits. Six outliers were identified and by looking at the raw spectra it was found that all of them were affected by scattering effects. After removing these outliers, each sample had at least two replicates. A new dataset consisting of 100 spectra was then obtained by taking the replicates' average.

The Standard Normal Variate (SNV) correction was separately performed on the Vis and the NIR datasets [41,42]. Mean centering was finally applied prior to data analysis.

### 2.1.3. $^{1}$H-NMR data acquisition and preprocessing

All the $^{1}$H-NMR profiles were acquired on a Bruker Avance III 600 spectrometer (Bruker Biospin Gmbh, Rheinstetten, Germany) operating at Larmor frequency of 600.13 MHz for protons, equipped with a double tuned cryoprobe (TCl) set for 5 mm sample tubes and a cooled autosampler (SampleJet, at 5°C). Spectra were acquired from all the beer specimens with TOPSPIN 2.1 (Bruker Biospin Gmbh, Rheinstetten, Germany), using the NOESYGPPR1D sequence [27,32]. Presaturation of the water signal (4.77 ppm, [24,25,27,30–32,43–45]) was employed, while the

ethanol signals were not suppressed [27,31,32]. All the experiments were performed at 298 K with a fixed receiver gain. Each Free Induction Decay (FID) was collected using a total of 64 scans plus 4 dummy scans. Prior to Fourier transformation the FIDs were zero-filled to 64k points and a 0.3 Hz Lorentzian line broadening was applied. The spectra were in some cases automatically and in some other manually baseline- and phase-corrected using the TOPSPIN processing tools, depending on the results of the automatic correction assessed by a trained NMR user. For all spectra, the ppm scale was referenced to the TSP peak at 0.00 ppm. The spectral window was 20.5 ppm.

After thawing and degassing, the specimens were kept at 5°C. Preparation of the NMR tubes was executed in batches of twelve samples, which were collected from the fridge and placed within a thermally insulated styrofoam box equipped with a ground of ice chips and closed with a lid. The newly prepared tubes were placed into the autosampler rack, which was also stored within the thermal box.

All the specimens were prepared to contain 10% $D_2O$, 0,02% of sodium-3-(trimethylsilyl)propionate-$d_4$ (TSP-$d_4$) as a chemical shift reference [24,25,27,30–32,43,44] and 20% phosphate buffer (pH = 3.55). The required volume for the NMR tubes was 600 µL, and it was obtained by mixing: 420 µL of beer specimen, 60 µL of $D_2O$ and 120 µL of phosphate buffer in $H_2O$. Spectra were acquired in random order with respect to samples and replicates.

Duarte et al. [43] studied the composition of ale and lager beers, and reported pH values within the 3.7–4.4 interval. The addition of a phosphate buffer (pH = 3.55) was aimed to obtain a set of specimens with more homogeneous pH values, so that the signal's horizontal shifts across spectra, due to the different protonation forms of compounds such as organic acids [31,32], could be reduced.

The NMR spectra were imported into Matlab and the signals aligned using *icoshift* [46,47]. Sixty-four spectral features were resolved by means of Multivariate Curve Resolution (MCR, [48]). MCR was applied to resolve the NMR spectra, by building MCR models on spectral intervals carefully selected one at a time rather than trying to make one overall model [49].

NMR data carry different information in different spectral regions. As a consequence, NMR spectra are usually roughly split into three regions [43,49]: the aliphatic/organic acids region (0–3 ppm), the carbohydrates region (3–5 ppm) and the aromatic region (6–9 ppm). These regions mainly differ because of involved metabolites/molecules, baseline noise, and signal's average intensity [49]. By using an interval-based approach it is possible to effectively handle those differences and to obtain meaningful chemical quantifications from each region. Interpretability and model performances are also generally improved.

One MCR model was built for each manually defined interval, using non-negativity constraint on both profiles and concentrations. For each model, the components representing chemical information were retained, whereas components describing baseline variations or noise were excluded. Sixty-four resolved components were eventually selected, and their relative concentrations were then merged to create a new dataset (NMR features). Twenty-one of these features were tentatively assigned based on literature assignments, while the remaining features were labelled as "unassigned". All exploratory analyses were performed on the NMR features dataset after autoscaling the 64 features.

### 2.2. Data Analysis

This section is organized as follows: first, we provide a brief recall of the different unsupervised data reduction techniques used for exploratory analysis and compression (feature extraction), then the clustering techniques employed in both exploratory and the proposed new approach, and finally the adopted data fusion strategies. The novel proposed approach is described at the end of the section.

The raw Vis/NIR data and the NMR features data (section 2.2.1) will be made available for download at http://www.models.life.ku.dk/datasets

#### 2.2.1. Data reduction

Multivariate Curve resolution (MCR, [48]) was applied to reduce the NMR spectra by features extraction, as explained in Section 2.1.3. MCR was also tested on the Vis and NIR datasets. Both the whole and interval-based approaches led to unclear results, probably because of the strong overlap and broadness of the pure signals; this may hinder meaningful curve resolution outcome. For these reasons, no compression other than Principal Component Analysis (PCA [3]) was performed on the Vis and NIR datasets.

PCA was also used for exploratory purposes: in Figures 2 and 3 it was applied to the preprocessed Visible and NIR spectral datasets, in Figure 5 to the autoscaled mid-level fused dataset and in Figure 6 to the Fused Adjacency Matrix ($AM_{Fus}$), preprocessed as described in Section 2.2.5.

### 2.2.2. Kohonen's Self-Organizing Maps (SOM)

In order to account for more complex structure in sample space and possible non-linearities, the Kohonen's Self-Organizing Maps (SOM [10,11]) were employed. SOM is a type of artificial neural network that is particularly suitable for modelling non-linear boundaries between samples belonging to different groups. Its aim is to obtain a low-dimensional representation of the high-dimensional input space. The high-dimensional space is mapped using a set of representative coordinates, which are distributed unevenly over the space, based on data structure and sample density. These coordinates are called nodes (or neurons) and are organized on a "top-map", typically a 2D grid whose geometry may vary. During the learning phase, the SOM network iteratively rearranges the samples over the top-map, assigning them to the most similar node [10]. At the same time the nodes get updated, based on the samples that were assigned to them. Since this is an unsupervised method, there is not a target arrangement of samples, therefore the network must adapt itself (hence the name "self-organizing" maps) according to the data structure. The top-map can be used as an exploratory tool for the identification of clusters [10], since it allows to assess similarity between samples in a simple and direct way, by comparing their position on the top-map.

SOM mapping preserves the topology, and this means that distances and proximity relations between samples are preserved [10]. As a result of this, all the nodes that are at the same topological distance from a given node define a "neighbourhood": a representation of nearest, second- and third-nearest neighbourhoods is given on the top-map in Figure 1.

In our work, a simple two-dimensional, 10-by-10 squared grid of nodes was used [11]. The network was trained for 10000 epochs, with rectangular neighbourhoods and a gaussian function for modulating the distance based-learning.

### 2.2.3. Ordering Points to Identify the Clustering Structure (OPTICS)

OPTICS [50–52] is a density-based clustering method aimed at revealing the data clustering structure. This method consists of an iterative procedure that only needs an initial input parameter, namely *k*, which is the minimal number of objects forming a cluster. Daszykowski and Walczak [52] suggested a rule of thumb for selecting *k*:

(1) $$k = \text{integer}\left(\frac{m}{25}\right)$$

where *m* is the number of samples.

OPTICS is based on the concept of Reachability Distance (RD). RD is a similarity measure [52], which is basically an Euclidean distance that describes how distant/similar is an object from the one processed at the preceding step. The graphical output of OPTICS is called Reachability Plot (RP), and it is obtained by plotting the RDs as vertical bars arranged along the x-axis according to the processing sequence.

At each iteration, the OPTICS algorithm selects one object and compares it with all the objects that have not been processed yet. This is done by computing all the pairwise Euclidean distances. Then, the next object to be processed is selected among the *k* nearest neighbours: the distance at which this next object is found becomes its RD, which is stored unchanged until the end of the procedure. The final output is therefore a set of RD values, which can be plotted as bars in the RP. A cluster is formed by objects that happen to be very close to each other, so it can be expected that these objects

would have, on average, a similar number of neighbours at similar distances, i.e. they would have similar neighbourhoods. These short distances among neighbours result in very similar RD values. When a cluster has been processed, then the next object would likely belong to another cluster: the next RD value in the processing sequence is therefore going to be larger than the values preceding it, which are related to previous cluster. This "jump" from one cluster to another is graphically recognizable in the RP because it corresponds to a very high bar. Clusters therefore appear as hollows created by groups of samples sharing similarly low RDs.

It is important to consider that the RP does not explicitly cluster the objects [52], but it rather allows deducing the number of clusters in the data.

### 2.2.4. Mid-level data fusion

Data fusion methods are strategies for combining different sources of complementary information, e.g. data blocks obtained from the analysis of the same set of samples by means of different analytical techniques. Data fusion strategies are generally grouped into three levels: low-, mid- and high-level methods [23,40,53]. Mid-level data fusion is accomplished by combining relevant features extracted from each data block.

In the present study, a mid-level data fusion dataset was obtained by creating a matrix augmented in the variables' direction. Seventy-seven features were merged: 7 PCA scores from the Vis dataset and 6 PCA scores from the NIR dataset were merged with the 64 NMR features. To represent the three different blocks evenly, autoscaling followed by block-scaling was performed.

### 2.2.5. Fused Adjacency Matrix approach

The Fused Adjacency Matrix approach is a two-step procedure: in the first step, information is extracted by processing single data blocks (in the present work Vis, NIR and NMR), and in the second step the extracted pieces of information are fused together. These two steps are marked in the lower part of Figure 1.

The approach is based on the concept of combining different weak sources of information [15–18] as it is done, for instance, in the classification context by the Random Forest algorithm (RF, [15]). In RF the results of several weak classifiers are merged by counting how many times a sample was assigned to one of the defined categories; then the sample is assigned to the category to which it was more often assigned.

In our unsupervised case, we convert the distance information into several adjacency matrices, which represent the weak sources of information. Adjacency matrices (AMs) are squared binary symmetric matrices ($m \times m$) in which a one is present when the adjacency condition is fulfilled by the pair of samples under exam, and a zero is present when this condition is not fulfilled. In other words, these matrices carry the information about whether two samples are close enough to each other (they are "adjacent") as compared to, for instance, a distance threshold (the adjacency condition). Merging these AMs using a sum rule [19] will result in a new squared symmetric matrix in which, those pairs of samples that were consistently found adjacent will be characterized by high values, while those pairs of samples which were consistently found far apart will have low values or, even better, values close to zero. This is the overall idea of the proposed approach.

In our approach, for a given data block ($X$ in Figure 1, on the left side), fourteen different AMs are obtained. Ten are derived by using Euclidean and Mahalanobis distances (Equation 1), and four by using SOM as a "clustering" method (Equation 2). Due to the number of implemented thresholds, the contribution of each distance measure to form the $\mathbf{AM_X}$ was comparable; however, the use of a weighted sum can be advised in the more general case.

(2) $$\mathbf{X} \rightarrow \mathbf{D_{Euc/Mah}} \rightarrow thr = [0.05, 0.1, 0.2, 0.3, 0.4] \rightarrow \mathbf{AM_{Euc/Mah}} = \sum_{thr=1}^{5} \mathbf{AM_{thr}}$$

(3) $$\mathbf{X} \rightarrow \text{SOM} \rightarrow \text{topmap} \rightarrow g = [0, 1, 2, 3] \rightarrow \mathbf{AM_{SOM}} = \sum_{g=0}^{3} \mathbf{AM}_{g,\mathbf{neigh}}$$

The Euclidean and Mahalanobis distance matrices are both normalized between zero and one, and the same window [22] of five threshold values (0.05 - 0.1 - 0.2 - 0.3 - 0.4) is applied to both the **D** matrices. SOM does not provide a distance matrix, but instead a grid of nodes (the top-map), on which the samples are arranged. In this case, the adjacency condition to be checked is whether the

two considered samples belong to the same *g* topological neighbourhood or to a closer one. We defined four topological rectangular [54] neighbourhoods (*g* = 0, 1, 2, 3), including the "zero*th* level", which corresponds to a single node. Since different SOM runs generally produce slightly different outputs, the average over ten runs was taken to make the resulting adjacency matrix $\mathbf{AM_{SOM}}$ more robust.

(4) $$\mathbf{AM_X} = \mathbf{AM_{Euc}} + \mathbf{AM_{Mah}} + \mathbf{AM_{SOM}}$$ (X = Vis, NIR, NMR)

(5) $$\mathbf{AM_{Fus}} = \sum_X \mathbf{AM_X} = \mathbf{AM_{Vis}} + \mathbf{AM_{NIR}} + \mathbf{AM_{NMR}}$$

Figure 1 provides a graphical representation of the whole Fused Adjacency Matrix approach. For a given data block **X**, its corresponding output is the matrix $\mathbf{AM_X}$ (Equation 3). When more than one **X** data blocks are available (like in the benchmark case presented in this work, where X = Vis, NIR, NMR), the resulting $\mathbf{AM_X}$ matrices can be combined using, again, a sum rule ([22], equation 4). The result is the Fused Adjacency Matrix $\mathbf{AM_{Fus}}$, depicted in black in Figure 1. In this work, the values in $\mathbf{AM_{Fus}}$ vary between zero and 42, as a result of summing a total of 42 AMs which have ones on their diagonal. Prior to analysis, the Fused Adjacency Matrix $\mathbf{AM_{Fus}}$ was double centered [55] so that:

(6) $$\mathbf{AM_{Fus,cent}} = \mathbf{AM_{Fus}} - \overline{\mathbf{AM}}_{\mathbf{Fus},m} - \overline{\mathbf{AM}}_{\mathbf{Fus},n} + \overline{\mathbf{AM}}_{\mathbf{Fus},mn}$$

which corresponds to remove the column mean $\overline{\mathbf{AM}}_{\mathbf{Fus},n}$ and the row mean $\overline{\mathbf{AM}}_{\mathbf{Fus},m}$ (which are exactly the same because $\mathbf{AM_{Fus}}$ is symmetric), and finally adding back the overall mean $\overline{\mathbf{AM}}_{\mathbf{Fus},mn}$, similarly to the way distance matrices are usually preprocessed [56].

### 2.3. Software

The whole data analysis process was carried out on MATLAB 2016a (Mathworks, MA, USA). PCA analysis was performed by using the PLS Toolbox 8.1.1 (Eigenvector Research Inc. WA, USA). NMR spectral alignment was operated using *i*coshift ([46,47], http://www.models.life.ku.dk/icoshift, last access 31/01/2019). NMR interval-resolution was operated by means of the MCR-ALS GUI by Joaquim Jaumot, Anna de Juan and Romà Tauler.

([57], https://mcrals.wordpress.com/, last access 31/01/2019). The OPTICS algorithm was written by Michal Daszykowski and it can be found at http://chemometria.us.edu.pl/download/OPTICS.M (last access 31/01/2019). Kohonen's Self-Organizing Maps were computed by using a homemade routine by Federico Marini (Università La Sapienza, Roma). The Fused Adjacency Matrix was computed by using in-house written MATLAB routines, which will be made available for download at http://www.models.life.ku.dk/algorithms.

## 3. Results and discussion

The results are organized in the following sections: first, results referring to each single spectral dataset (Sections from 3.1 to 3.3) are presented, then results from mid-level data fusion are discussed in Section 3.4 and, eventually results from the Fused Adjacency Matrix approach are reported in Section 3.5; more detailed comparisons among the different results are reported in Section 3.6 and summarized in Table 1. The different fusion steps were also inspected by means of Procrustes Analysis, and the results are reported in Section 3.7 Finally, an example of how to link the Fused Adjacency Matrix to the original NMR variables is given in Section 3.8.

It is important to clarify that the results regarding the proposed novel approach are only those reported in Section 3.5 The results for the Visible, NIR and NMR data were obtained working on the preprocessed spectral data (resolved features, in the case of NMR), so no AMs were involved in the single-data block analyses.

### 3.1. Visible dataset

The visible spectra, after preprocessing, were analysed by PCA and OPTICS. Figure 2 reports the results, namely the OPTICS reachability plot (RP) in Figure 2a, and the PC1-PC2 score plot in Figures 2b and 2c, colored according to beer style (b) and colour intensity (c).

Two main groups were identified by OPTICS. The first one, the Ales group, is mainly composed by ale-style samples and it is less homogeneous compared to the second, the Lagers group, which is

largely composed by lager-style samples. The two groups also have different density: the Lagers group results denser than the Ales group, and this can be seen in both the RP (Fig.2a) and the score plot (Fig.2b). The colour scale employed in Figure 2c describes the beer colour intensity, that is defined as the absorption of the sample at 430 nm, taken as reference wavelength [58]. A colour intensity gradient is recognizable along PC1 (Fig.2c). The sample distribution along PC2 is, on the contrary, much less clear. Some of the mid-coloured samples are spread along PC2, and the four samples with the strongest absorption have negative scores on this component. These four samples belong to very different beer styles but look rather grouped in the PC1-PC2 score plot. This is not reflected by the RP, where the samples show increasingly higher distances. Actually, by inspecting the score plots of higher PCs (not shown) these non-grouped samples are always found at extreme positions with respect to the rest of the samples. Since OPTICS operates on the full spectra, the increasing RD trend is due to the piece of information that is not included in the PC1-PC2 score plot.

### 3.2. NIR dataset

The information that could be extracted from the NIR dataset is rather limited, and this can be seen by inspecting the RP (Fig.3a) and the PC1 score plot (Fig.3b), both obtained from the NIR preprocessed spectra.

A clear alcohol content (% alcohol by volume, ABV%) gradient is recognizable along PC1, as shown in Figure 3b. Ethanol content is therefore efficiently represented by PC1, whose corresponding loadings (not shown) are characterized by two intense ethanol bands within the region 2200–2400 nm [37].

Two main clusters of samples were identified by inspecting the RP (Fig.3a), a small one which contains a mix of beer types (“mixed group”) and the Lagers group. The Light beer samples appear rather grouped, as it is indicated by the shaded light blue rectangular area in Figures 3a and 3b. The samples located at the right end of the plot can be considered as non-grouped. This was also found

in PCA, where the two identified clusters have reduced variability along PC1 with respect to the non-grouped samples (Fig.3b). The non-grouped set is much more scattered, as it has both higher bars in the RP (Fig.3a) and a large variability range along PC1 (Fig.3b).

### 3.3. NMR dataset

A data representation from the field of Sensomics [59,60], was used for inspecting the NMR features and the results are shown in Figure 4. The heatmap [60] in the central part of the figure represents the data values. The columns of the heatmap represent the samples while the rows represent the variables (concentrations of MCR-resolved features in the different samples). Rows and columns were reordered according to the sequences obtained by running OPTICS first in the samples' direction (RP on top) and then also in the variables' direction (RP on the left side). This allows highlighting both groups of samples and variables, making it easier to relate the most influent groups of variables to each group of samples [60].

To obtain clearer groupings in the variables' direction, correlation among the NMR features was used, instead of distance, to calculate the reachability distance for the RP plot. Three main groups of variables can be identified (Figure 4 variables' RP, on the left side): the first group mainly contains amino acids, together with uridine and gallate; the second group is composed of yet unassigned variables, and the third group is partially related to maltose and to two unassigned variables.

The samples' RP shows a cluster that can be identified as the Lagers group. The rest of the plot is rather uninformative from a group-spotting point of view, since its largest part consists of a sequence of increasing RDs (non-grouped set). Interestingly, the Light beer samples constitute a recognizable sub-group which, as expected, has generally low values for all the variables. Also, a small group can be spotted at the centre of the RP plot (group D in Figure 4), and it is characterized by medium-low values in amino acids and medium values for the second group of variables. The non-grouped set contains very different beer styles. The samples belonging to this group generally have higher amino acids content, but also maltose (third group of variables).

### 3.4. Mid-level data fusion

The PCA and OPTICS results obtained from the preprocessed mid-level fused dataset are shown in Figure 5. The OPTICS results resemble those of the NMR features dataset: a slightly defined Lagers group at the beginning of the RP, followed by a tail of slowly increasing RDs forming a non-grouped set (Fig.5a). However, the sample distribution obtained by PCA (score plot in Fig.5b) is mainly determined by few variables, according to the loadings plot (Fig.5c). Features related to ABV ("Scores PC1–NIR") and colour ("Scores PC1–Vis", "Scores PC2–Vis") are the most influential.

All the Light beer samples are located at negative PC1 and positive PC2 scores, while two of the strongest samples lie far away in the opposite direction. This defines an ABV direction (light blue arrow in Figure 5b). Even though the Light beer samples seem to be rather grouped in PCA, they are not found grouped in the RP. Again, an explanation for this discrepancy can be found in the different amount of information described by the RP (the whole preprocessed data) and the first two PCs shown in Figure 5b, which only account for 29.63% of the total variance of the mid-level fused dataset. Almost perpendicularly to the ABV direction, the variable "Scores PC1–Vis" (Fig.5c) tends to separate the most coloured samples (Fig.5b, highlighted in orange), and helps to separate along PC1 the Lagers from the Ales, which usually have more intense colours.

### 3.5. Fused Adjacency Matrix

The results obtained by OPTICS and PCA on the Fused Adjacency Matrix preprocessed as explained in Section 2.2.5 are discussed here and shown in Figure 6.

Two clusters of samples and a non-grouped set can be identified in the RP (Fig.6a). These three groups have a correspondence in the PC3-PC1 score plot of the same matrix (Fig.6b) The non-grouped set is more scattered in PCA (blue patch in Figure 6b), and it contains the strongest one and three of the five Light beer samples. The Ales and Lagers groups are much more defined compared

to the results found with the single techniques and the mid-level data fusion approach. It is also interesting to notice the sample distribution within the Lagers group, where the "simple" lager samples (in red in Figure 6b) are very grouped on the right side, which is in an opposite position compared to the Ales group.

PC1 is related to the colour, and when combined with PC4 the samples adopt an arch-like distribution (Fig.6c). The PC1-PC4 score plot not only shows the colour trend, but also suggests new groups of samples, which are highlighted in grey in Figure 6c. To gather which characteristic features are shared within these sub-groups the sub-group average NIR spectra (Fig.S1a) and NMR resolved features (Fig.S1b) were compared. Most of the groups have some distinctive regions, e.g. sub-groups 6 and 7 have higher content of amino acids content, while the three close IPAs (sub-group 4) have high values in NMR for maltose and a set of features not yet completely identified, among which ethanal, isopentanol and higher alcohols were tentatively assigned.

Based on our current knowledge, it is not possible to fully explain these groupings, however work is in progress analysing a database of consumer preferences obtained from the website ratebeer.com[1] to assess if some of the grouping may be related to such information. Preliminary results show that PC1 of the Fused Adjacency Matrix seems to have a strong inverse relationship ($R^2$ = –0.973) with the overall score computed by the website from the users' evaluations (Fig.S1c).

[1]https://www.ratebeer.com/ (last access 31/01/2019)

### 3.6. Beer features comparison summary

In this section, more detailed comparisons among the results obtained by the different data blocks and data fusion approaches are reported. Table 1 is organized as a summary of these comparisons. Some overall samples' sets and beer features were tracked along the single data blocks.

#### 3.6.1. Lagers group

The Lagers group was identifiable in all representations of the data, and it appears to be rather stable. The Vis and **AM$_{Fus}$** datasets showed the best results in terms of samples grouping, which is probably reflected by their similarity, as highlighted by Procrustes Analysis (Section 3.7).

An interesting group of lager-style samples is the HI samples set, which includes beer products from the same brand, Hite. This set of samples is organized in couples of replicates: "Pale Lager" (HI.1-2, HI.3-4), "Dry Finish" (HI.6-7), "Golden" (HI.8-9) and "Fresh" (HI.10-11-12-13), where the second replicate underwent thermal treatment to simulate ageing. Only sample HI.5 does not have a replicate and it is also a different beer product ("MAX"). The HI samples were generally found in the Lagers group, with some exceptions: HI.1 and HI.5 in NIR (Fig.3a); HI.8-9 and HI.5 in NMR (Fig.4). No fixed order related to thermal treatment was found, neither with OPTICS nor with PCA, in any dataset. Moreover, no consistent order of the replicates was found neither in the spectral datasets, nor in the mid-level fused dataset, even though in the NMR case some of the HI samples were found gathered in two sub-groups: group B (HI.10-11 and HI.12-13) and group C (HI.4-3, HI.6-7) in Figure 4. Group B has higher content of some amino acids, acetate, uridine and an unassigned variable between the two last ones. On the contrary, this piece of information clearly emerged by analysis of **AM$_{Fus}$** dataset. In fact, the HI samples were found very well grouped together in the RP (HI in Figure 6a), forming a rather ordered sequence of couples of HI replicates; couple HI.3-4 was not found among the other HI samples, but some positions further in the sequence of the RP (Fig.6a).

Another interesting set of samples is represented by the EU beers. They belong to the same brand and three of them are the same product (EU.1-2-3, "Brüger Premium Pils"), while EU.4 ("Servus") is different. However, sample EU.2, differently from the other three EU samples, did not undergo thermal treatment. These samples were not found grouped in the Vis and NIR cases, while in NMR, mid-level data fusion and **AM$_{Fus}$** the EU group was recovered in the RPs, albeit to different extents. In the NMR case, the samples are ordered (group A in Figure 4) as EU.1, EU.3 ("Brüger" treated), then EU.2 ("Brüger" non-treated) and finally EU.4 ("Servus" treated). In the case of mid-level data

fusion, a similar situation was found, but EU.4 was found further in the RP. Interestingly, in the $\mathbf{AM_{Fus}}$ case, the three thermal treated samples (EU.1, EU.3 and EU.4) were found grouped together (group A in Figure 6a), while EU.2 one was found further in the OPTICS sequence, suggesting that, only by this approach, a clearer difference based on the treatment was recovered.

Three "unclassified" samples (LE.1, OE.4, KR.1) were consistently found in the Lagers group. These products are described as "summer beers", therefore their presence in the Lagers groups is not unforeseen: this product type is intended to be refreshing and easy-to-drink, and it usually is lighter in aromas and alcohol content. For these reasons it can be expected to find these summer beers more similar to the lagers than the ales.

### 3.6.2. Light samples set

The Light samples set includes five beers of different styles (KR.2, Classic light / LE.2, IPA light / FB.2, Lager light / TO.4, Lager light / NO.2, Light Ale). These beers are labelled as "light" and they are produced with the aim of obtaining a lower content of ethanol and flavours.

The NIR and the NMR datasets gave the best results in terms of grouping the Light samples set. In the NIR case the Light samples were found grouped both in the RP and the PCA scores (light blue patches in Figure 3). They lie at extreme positive values along PC1, which is a component that describes ethanol content. A confirmation of the generally lower content in flavours was found from the NMR results: all the Light samples share a similar pattern of very low values along all the variables of the dataset (Light sub-group in Figure 4).

The Light samples set was found rather grouped in the data fusion cases (Figures 5b and 6b), but only in PCA. In the Vis case, the Light samples are neither grouped in RP or PCA but belong to the Lagers group: lighter beers are usually less processed/fermented, so they tend to develop less intense colour.

### 3.6.3. ABV trend

No ABV trend was evident in the Vis case. This is naturally present in the NIR case (Fig.3b), since PC1 describes the ethanol content. The trend is also present in the mid-level data fusion case, since variable PC1 from NIR is highly influential (Fig.5c). No clear ABV trend was found in the RP for the NMR case, even if it was found in PCA, which is reported in the Supplementary Materials as Figure S2a.

The **$AM_{Fus}$** case is rather different. The ABV trend is present in PC1-PC3 (score plot reported in Figure S3, in the Supplementary Materials), but in a transformed way. The strongest and the lightest beers all lie in the top part of the plot and they all belong to the non-grouped set (as in Figure 6b). These samples represent the extremes in ABV, so their position is probably due to the fact that the approach is just able to detect their dissimilarity from the bulk of "ABV-average" samples.

### 3.6.4. Lagers Strong set

The Lagers Strong set includes six beers (ordered by increasing ABV, MA.3, SI.9, MA.5, MA.6, MA.2, FB.3) and it is interesting to track their position because of their style: lagers strong are beers brewed with lager yeasts, but more alcohol is obtained during the brewing process.

The Lagers Strong set was generally found split into two groups: four "low-ABV" and two "high-ABV" samples. The low-ABV samples (MA.3, SI.9, MA.5, MA.6) were found in the Lagers group in the cases of Vis, mid-level data fusion and **$AM_{Fus}$**, while the NIR and NMR cases provided two different situations. In the NIR case, the three lowest ABV samples were found in the mixed group, closer to the Lagers than the three highest ABV samples (Fig.3a). On the contrary, in the NMR case, the Lager Strong samples are all in the Lagers group and do not follow any ABV order (Fig.4). Both the data fusion approaches, in RP by OPTICS (Fig.5a and Fig.6a) is clearly highlighted that the four low-ABV samples are more similar to the lagers (they belong to the Lagers group) but are also located closer to each other within the RP sequence. However, the separation between high- and low-ABV samples is much better appreciable in the PCA of the **$AM_{Fus}$** (Fig.6b) than in the mid-level data fusion score plot (Fig.5b). In **$AM_{Fus}$**, moving along PC1 from the Lagers

group towards the Ales group, the four low-ABV samples are found, while the two high-ABV samples are much more distant, and closer to the strongest samples in the dataset. On the contrary, the same samples in the mid-level data fusion score plot (Fig.5b) are located in the same area.

### 3.6.5. Colour trend

The colour trend naturally originates from the Vis dataset (Fig.2c). No trace of it was found neither in the NIR nor the NMR cases. Both the data fusion methods were able to recover this piece of information, even though the $\mathbf{AM_{Fus}}$ (Fig.6c) provides a clearer trend than the mid-level data fusion (Fig.5b).

### 3.6.6. Summary Remarks

The trends and groupings described above generally correspond to the main known traits of the beer styles under examination. While the single spectral data blocks can primarily provide one aspect each, both the data fusion approaches were able to collect and keep most pieces of information. The Fused Adjacency Matrix, however, could capture finer structures in the main groups, for instance the very well-ordered HITE group, with the replicates of each product found in a sequence by OPTICS, or the EU set, where the treated samples were found grouped together and the non-treated one was found much further away. Trends like colour intensity and lager/ales distinction were recovered more clearly by the Fused Adjacency Matrix, while others like ABV content and the Light samples set were slightly better retrieved by the mid-level data fusion approach.

It is also very promising that the Fused Adjacency Matrix approach can highlight small sub-groups (Fig.6c) which may be worth further investigation of their chemical/sensory characteristics. A deeper characterization of these sub-groups may, for instance, provide new inspiration in beer production, helping to define intersections between established and more general styles.

*Table 1 to be inserted about here*

585 **3.7. Comparisons by means of Procrustes Analysis**

586 In Sections from 3.1 to 3.6 we have graphically inspected and compared the information gathered
587 by the different data blocks as depicted in the principal components space, with the aim of
588 highlighting similarities and differences among them. This way of visually exploring the data easily
589 allows spotting trends and peculiarities, but subjectivity and limited availability of metadata (i.e.
590 additional information such as the beer style or the ABV content) can sometimes be a drawback.
591 A more objective evaluation of how similar/different are the results obtained from the different data
592 blocks by comparing their PCA spaces can be obtained by means of Procrustes Analysis (PA,
593 [61,62]). Like in our beer benchmark case, the same set of objects can be described by two distinct
594 sets of PC scores, obtained for instance from two different analytical sources. The aim of PA is to
595 obtain the closest match between these two PC spaces by applying operations such as scaling,
596 rotation, reflection and translation. The similarity of the two spaces is expressed using a
597 dissimilarity parameter *d*, ranging from zero to one [62].
598 In this work, the PCA spaces obtained from the different blocks (i.e. each single analytical platform,
599 the mid-level fused data set and the $\mathbf{AM_{Fus}}$ data set, referred to as inter-block comparison) are
600 compared by PA analysis. Also, the data obtained from the different steps of the procedure, going
601 from the raw data to the AMs for each single data set (which will be named $\mathbf{AM_X}$, with the suffix X
602 being Vis, NIR and NMR, in turn) have been compared by PA. The latter case is referred to as
603 intra-block comparisons. An overview of the results is given hereinafter, while the visual
604 representation is reported in Figure S4, in the Supplementary Materials.
605 Inter-block comparisons were made, in pairs, using the PC scores of the Visible spectra (7 PCs), the
606 NIR spectra (6 PCs), the NMR features (6 PCs), the mid-level fused data (5 PCs) and the Fused
607 Adjacency Matrix ($\mathbf{AM_{Fus}}$, 7 PCs). The same number of principal components as that considered to
608 build the mid-level fused dataset were used in PA, to keep it constant, and the results are shown in
609 Figure S4a, where the dissimilarity value between each pair of data sets is reported. $\mathbf{AM_{Fus}}$ is
610 substantially different (dissimilarity higher than 0.5) from the mid-level fused data, which suggests

that these two datasets carry different information. $\mathbf{AM_{Fus}}$ was also found rather different from the other datasets: this is a desirable situation, since we are dealing with a data fusion approach. A too strong resemblance with any single source dataset would have meant that the fusion process was giving too much importance to that source, while a too loose similarity would have meant that the information was either too reduced or not captured by the approach.

The effect of the different fusion steps was also assessed. These intra-block comparisons were made for each data block individually (using the same number of PCs as specified above), and the results are shown in Figure S4b. One interesting point is the transition from the distance information to its correspondent $\mathbf{AM_X}$. The Euclidean distance $\mathbf{D_{Euc}}$ resulted consistently similar to the Euclidean $\mathbf{AM_{Euc}}$ meaning that the "coded" AM version of the data is keeping a large part of the original distance information. The same was observed with the Mahalanobis distance, albeit for the NMR case the similarity between $\mathbf{D_{Mah}}$ and $\mathbf{AM_{Mah}}$ was found lower (Fig.S4b). By inspecting the corresponding score plot it appears that this difference is due to a limited number of samples which have extreme values on the second component in PCA of $\mathbf{D_{Mah}}$ and are not in $\mathbf{AM_{Mah}}$ (adjacency being assigned on interval values is less sensitive to extreme values). Another interesting relation is between the Euclidean and SOM AMs: the matrices $\mathbf{AM_{Euc}}$ and $\mathbf{AM_{SOM}}$ are very similar, either because the samples pattern in the beer data can be well described by a linear model or because the Euclidean distance (which is a non-linear transform) is sufficient to model the non-linearity present in the data pattern. These two AMs also represent the two major contributions to the single-data block $\mathbf{AM_X}$. The Mahalanobis distance was consistently found rather different from $\mathbf{AM_X}$ and the other distance measures. This is probably because higher PCs bring in rather different information with respect to the first ones, as in order to avoid singularities we have calculated the Mahalanobis distance on PCA-compressed data and thus it corresponds to Euclidean distances on the autoscaled PCs. However, a systematic different behavior of the Mahalanobis distance with respect to other metrics (including Euclidean) has been previously observed in a study considering several data sets [63].

637

638 **3.8. Link to the original variables**

639 One of the major issues when dealing with adjacency matrices is that the link with the original
640 variables is lost. When an adjacency matrix is built, the "adjacency condition" for each pair of
641 samples is evaluated, therefore the focus is on how distant the two samples are: the original
642 variables are only used to compute the distances.

643 A way for linking back the Fused Adjacency Matrix results to the original variables is presented in
644 Figure 7 using the NMR features dataset as an example. By using the same representation used in
645 Figure 4, the samples were reordered using the RP sequence obtained from the Fused Adjacency
646 Matrix. Therefore, the heatmaps of the two figures only differ in the order of their columns. Such a
647 new column sorting allows a direct comparison between the observed sample clusters and the
648 chemical features linked to specific class of compounds, as detailed in the following section.

649 The Ales group in Figure 7 shows medium-high values in correspondence of the amino acids. The
650 non-grouped set also has some samples with comparable values for the amino acids, but the Ales
651 group has a more uniform composition. The amino acids region also represents the main difference
652 between the Ales and the Lagers groups. This is in accordance with the results obtained by *Duarte*
653 *et al.* [24], who suggested that the aromatic region could be used to distinguish between ales and
654 lagers.

655 Two sub-groups can be noticed within the Ales group (A and B in Figure 7). The first sub-group
656 (A) is mixed, and consists of seven ales, four lagers and one unclassified beer. These samples have
657 medium values for variables from 3 to 11, which include compounds such as tryptophan, gallate,
658 phenylalanine, uridine and two signals from proline. Their amino acid content is on the other hand
659 much lower if compared to the other samples belonging to the Ales group. The second sub-group (B
660 in Figure 7) consists of five ales and two lagers. This sub-group is characterized by high values
661 related to the first 20 variables, which include all the identified amino acids together with gallate
662 and uridine.

The Lagers group generally has medium-low values, especially in the case of the second group of variables and the amino acids group. Several sub-groups can be identified within the Lagers group (C, D, E, F and G in Figure 7). A couple of samples at the beginning of the group (C) have almost identical patterns, especially for the amino acids content. These two samples are the same beer product, but the second one underwent thermal treatment. Some differences can be spotted along the two patterns, and the second sample always has higher values at these points. A second sub-group (D) consists of four lager samples of the same brand, which are among the poorest in amino acids content. Their patterns look very similar to sub-group E, which contains two beers of the previous brand, two more lagers and one lager strong. Sub-groups F and G also have similar patterns, but the samples in F tend to have higher values in amino acids, but lower values for the variables in the upper part of the map. At the boundary between the Lagers group and the non-grouped set, a sub-group of four samples (H) can be found. This small group is characterized by high values in amino acids and medium values for the maltose group.

This visualization approach is very efficient when dealing with data such as extracted features, while in the case of continuous data (e.g. spectra, chromatograms) reordering the original variables would make the visual interpretation very difficult. An example with the Vis and NIR cases is given in Supplementary Material, Figure S5a and S5b respectively, without having performed variables reordering. In the case of Vis (Fig.S5a) different intensity of the absorption bands between the two main Ales and Lagers group can be observed, while for the NIR case (Fig.S5b) the pattern is not so clear to interpret and differences in absorption intensity, for most of the spectral regions, are highlighted only for the non-grouped set.

## 4. Conclusions

The Fused Adjacency Matrix approach can recover coherent information from different datasets with highly complex structures, highlighting groups and trends in a way comparable to and in some

cases superior to the mid-level fusion approach. Differences and similarities among the different approaches were shown, and the most important findings are organized and reported in Table 1.

As it should be expected from a data fusion approach, the Fused Adjacency Matrix is able to retain the information from the original datasets, and to reveal other features arising from the combination of the fused sources. Possible new sample clusters were also highlighted, but their interpretation is not straightforward: this is for sure an aspect that deserves deeper investigation.

Further research about the Fused Adjacency Matrix approach should be directed mainly in two directions. Firstly, the approach should be tested on other datasets, ideally of very different provenience, nature and complexity. Secondly, the approach itself should also be improved from a structural point of view. For instance, the issue of linking back to the original variables may be addressed, with the aim of enhancing the interpretability of the results. Another aspect that may be investigated is the influence on the whole process of the different thresholds and neighbourhoods. This influence may be assessed by folding the single AMs (i.e. the matrices at the steps prior to the summing and averaging operations in Figure 1) in a three-way array and analysed it by means of PARAFAC or Tucker modelling.

Finally, the obtained results and new groupings may be used to investigate beer from the gastronomic point of view, with particular focus on sensory and consumer evaluations. Assessing the link between the objective world of analytical chemistry and the subjective world of consumer experience may produce great value for both the industry and the beer lovers.

## Conflict of interest

There is no conflict of interest.

## Acknowledgements

Helena da Silva Friis is greatly acknowledged for contributing to sampling and part of the experimental work.

714

715 **References**


716 [1] M. Li Vigni, C. Durante, M. Cocchi, Exploratory Data Analysis, in: Data Handl. Sci.
717 Technol., Elsevier, 2013: pp. 55–126. doi:10.1016/B978-0-444-59528-7.00003-X.

718 [2] I.T. Jolliffe, Principal component analysis, 2nd ed., Springer, 2002.

719 [3] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods. 6 (2014) 2812–2831.
720 doi:10.1039/C3AY41907J.

721 [4] P.D. Wentzell, Other Topics in Soft-Modeling: Maximum Likelihood-Based Soft-Modeling
722 Methods, in: Compr. Chemom., Elsevier, 2009: pp. 507–558. doi:10.1016/B978-044452701-
723 1.00057-0.

724 [5] J.H. Friedman, J.W. Tukey, A Projection Pursuit Algorithm for Exploratory Data Analysis,
725 IEEE Trans. Comput. C-23 (1974) 881–890. doi:10.1109/T-C.1974.224051.

726 [6] J.F.Q. Pereira, C.S. Silva, A. Braz, M.F. Pimentel, R.S. Honorato, C. Pasquini, P.D.
727 Wentzell, Projection pursuit and PCA associated with near and middle infrared hyperspectral
728 images to investigate forensic cases of fraudulent documents, Microchem. J. 130 (2017) 412–
729 419. doi:10.1016/j.microc.2016.10.024.

730 [7] F. Westad, M. Kermit, Independent Component Analysis, in: Compr. Chemom., Elsevier,
731 2009: pp. 227–248. doi:10.1016/B978-044452701-1.00045-4.

732 [8] A. Hyvärinen, Independent component analysis: recent advances., Philos. Trans. A. Math.
733 Phys. Eng. Sci. 371 (2013) 20110534. doi:10.1098/rsta.2011.0534.

734 [9] M.C. Hout, M.H. Papesh, S.D. Goldinger, Multidimensional scaling, Wiley Interdiscip. Rev.
735 Cogn. Sci. 4 (2013) 93–103. doi:10.1002/wcs.1203.

736 [10] F. Marini, R. Bucci, A.L. Magrì, A.D. Magrì, Artificial neural networks in chemometrics:
737 History, examples and perspectives, Microchem. J. 88 (2008) 178–185.
738 doi:10.1016/j.microc.2007.11.008.

739 [11] T. Kohonen, Essentials of the self-organizing map, Neural Networks. 37 (2013) 52–65.

doi:10.1016/J.NEUNET.2012.09.018.

[12] K. Varmuza, P. Filzmoser, Introduction to multivariate statistical analysis in chemometrics, CRC Press, 2009.

[13] I. Lee, J. Yang, Common Clustering Algorithms, in: Compr. Chemom., Elsevier, 2009: pp. 577–618. doi:10.1016/B978-044452701-1.00064-8.

[14] R. Gelbarda, O. Goldmanb, I. Spiegler, Investigating diversity of clustering methods: An empirical comparison, Data Knowl. Eng. 63 (2007) 155–166. doi:10.1016/J.DATAK.2007.01.002.

[15] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32. doi:10.1023/A:1010933404324.

[16] Yan-Yong Xu, Xian-Zhong Zhou, Zhong-Wei Guo, Weak learning algorithm for multi-label multiclass text categorization, in: Proceedings. Int. Conf. Mach. Learn. Cybern., IEEE, 2002: pp. 890–894. doi:10.1109/ICMLC.2002.1174511.

[17] P. Latinne, O. Debeir, C. Decaestecker, Combining Different Methods and Numbers of Weak Decision Trees, Pattern Anal. Appl. 5 (2002) 201–209. doi:10.1007/s100440200018.

[18] Chuanyi Ji, Sheng Ma, Combinations of weak classifiers, IEEE Trans. Neural Networks. 8 (1997) 32–42. doi:10.1109/72.554189.

[19] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 226–239. doi:10.1109/34.667881.

[20] L.I. Kuncheva, Combining pattern classifiers - methods and algorithms, 2014.

[21] A.-O. Boudraa, A. Bentabet, F. Salzenstein, Dempster-Shafer's Basic Probability Assignment Based on Fuzzy Membership Functions, ELCVIA Electron. Lett. Comput. Vis. Image Anal. 4 (2004) 1. doi:10.5565/rev/elcvia.68.

[22] B. Brownfield, T. Lemos, J.H. Kalivas, Consensus Classification Using Non-Optimized Classifiers, Anal. Chem. 90 (2018) 4429–4437. doi:10.1021/acs.analchem.7b04399.

[23] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food and beverage authentication and quality assessment – A review, Anal. Chim. Acta.

766 891 (2015) 1–14. doi:10.1016/j.aca.2015.04.042.

767 [24] I.F. Duarte, A. Barros, C. Almeida, M. Spraul, A.M. Gil, Multivariate Analysis of NMR and
768 FTIR Data as a Potential Tool for the Quality Control of Beer, J. Agric. Food Chem. 52
769 (2004) 1031–1038. doi:10.1021/jf030659z.

770 [25] C. Almeida, I.F. Duarte, A. Barros, J. Rodrigues, M. Spraul, A.M. Gil, Composition of Beer
771 by 1H NMR Spectroscopy: Effects of Brewing Site and Date of Production, J. Agric. Food
772 Chem. 54 (2006) 700–706. doi:10.1021/JF0526947.

773 [26] D.W. Lachenmeier, Rapid quality control of spirit drinks and beer using multivariate data
774 analysis of Fourier transform infrared spectra, Food Chem. 101 (2007) 825–832.
775 doi:10.1016/j.foodchem.2005.12.032.

776 [27] J.A. Rodrigues, A.S. Barros, B. Carvalho, T. Brandão, A.M. Gil, Probing beer aging
777 chemistry by nuclear magnetic resonance and multivariate analysis, Anal. Chim. Acta. 702
778 (2011) 178–187. doi:10.1016/j.aca.2011.06.042.

779 [28] S. Grassi, J.M. Amigo, C.B. Lyndgaard, R. Foschino, E. Casiraghi, Assessment of the sugars
780 and ethanol development in beer fermentation with FT-IR and multivariate curve resolution
781 models, Food Res. Int. 62 (2014) 602–608. doi:10.1016/J.FOODRES.2014.03.058.

782 [29] V. Giovenzana, R. Beghi, R. Guidetti, Rapid evaluation of craft beer quality during
783 fermentation process by vis/NIR spectroscopy, J. Food Eng. 142 (2014) 80–86.
784 doi:10.1016/J.JFOODENG.2014.06.017.

785 [30] I.F. Duarte, M. Godejohann, U. Braumann, M. Spraul, A.M. Gil, Application of NMR
786 Spectroscopy and LC-NMR/MS to the Identification of Carbohydrates in Beer, J. Agric.
787 Food Chem. 51 (2003) 4847–4852. doi:10.1021/JF030097J.

788 [31] L.I. Nord, P. Vaag, J.Ø. Duus, Quantification of Organic and Amino Acids in Beer by 1H
789 NMR Spectroscopy, Anal. Chem. 76 (2004) 4790–4798. doi:10.1021/ac0496852.

790 [32] J.A. Rodrigues, G.L. Erny, A.S. Barros, V.I. Esteves, T. Brandão, A.A. Ferreira, E. Cabrita,
791 A.M. Gil, Quantification of organic acids in beer by nuclear magnetic resonance (NMR)-

based methods, Anal. Chim. Acta. 674 (2010) 166–175. doi:10.1016/j.aca.2010.06.029.

[33] O. Oladokun, A. Tarrega, S. James, K. Smart, J. Hort, D. Cook, The impact of hop bitter acid and polyphenol profiles on the perceived bitterness of beer, Food Chem. 205 (2016) 212–220. doi:10.1016/j.foodchem.2016.03.023.

[34] C. Andrés-Iglesias, C.A. Blanco, J. Blanco, O. Montero, Mass spectrometry-based metabolomics approach to determine differential metabolites between regular and non-alcohol beers, Food Chem. 157 (2014) 205–212. doi:10.1016/j.foodchem.2014.01.123.

[35] S. Rossi, V. Sileoni, G. Perretti, O. Marconi, Characterization of the volatile profiles of beer using headspace solid-phase microextraction and gas chromatography-mass spectrometry, J. Sci. Food Agric. 94 (2014) 919–928. doi:10.1002/jsfa.6336.

[36] E. Bravi, O. Marconi, V. Sileoni, G. Perretti, Determination of free fatty acids in beer, Food Chem. 215 (2017) 341–346. doi:10.1016/J.FOODCHEM.2016.07.153.

[37] S. Engelhard, H.-G. Löhmannsröben, F. Schael, Quantifying Ethanol Content of Beer Using Interpretive Near-Infrared Spectroscopy, Appl. Spectrosc. 58 (2004) 1205–1209. doi:10.1366/0003702042336000.

[38] O. Klein, A. Roth, F. Dornuf, O. Schöller, W. Mäntele, The Good Vibrations of Beer. The Use of Infrared and UV/Vis Spectroscopy and Chemometry for the Quantitative Analysis of Beverages, Zeitschrift Für Naturforsch. B. 67 (2012) 1005–1015. doi:10.5560/znb.2012-0166.

[39] A. Biancolillo, R. Bucci, A.L. Magrì, A.D. Magrì, F. Marini, Data-fusion for multiplatform characterization of an italian craft beer aimed at its authentication, Anal. Chim. Acta. 820 (2014) 23–31. doi:10.1016/j.aca.2014.02.024.

[40] M. Silvestri, A. Elia, D. Bertelli, E. Salvatore, C. Durante, M. Li Vigni, A. Marchetti, M. Cocchi, A mid level data fusion strategy for the Varietal Classification of Lambrusco PDO wines, Chemom. Intell. Lab. Syst. 137 (2014) 181–189. doi:10.1016/j.chemolab.2014.06.012.

[41] D.-W. Sun, Å. Rinnan, L. Nørgaard, F. van den Berg, J. Thygesen, R. Bro, S.B. Engelsen, Data Pre-processing, in: Da-Wen Sun (Ed.), Infrared Spectrosc. Food Qual. Anal. Control, Elsevier, 2009: pp. 29–50. doi:10.1016/B978-0-12-374136-3.00002-X.

[42] L. Vera, L. Aceña, J. Guasch, R. Boqué, M. Mestres, O. Busto, Discrimination and sensory description of beers through data fusion, Talanta. (2011). doi:10.1016/j.talanta.2011.09.052.

[43] I. Duarte, A. Barros, P.S. Belton, R. Righelato, M. Spraul, E. Humpfer, A.M. Gil, High-Resolution Nuclear Magnetic Resonance Spectroscopy and Multivariate Analysis for the Characterization of Beer, J. Agric. Food Chem. 50 (2002) 2475–2481. doi:10.1021/jf011345j.

[44] A.M. Gil, I.F. Duarte, M. Godejohann, U. Braumann, M. Maraschin, M. Spraul, Characterization of the aromatic composition of some liquid foods by nuclear magnetic resonance spectrometry and liquid chromatography with nuclear magnetic resonance and mass spectrometric detection, Anal. Chim. Acta. 488 (2003) 35–51. doi:10.1016/S0003-2670(03)00579-8.

[45] D.W. Lachenmeier, W. Frank, E. Humpfer, H. Schäfer, S. Keller, M. Mörtter, M. Spraul, Quality control of beer using high-resolution nuclear magnetic resonance spectroscopy and multivariate analysis, Eur. Food Res. Technol. 220 (2005) 215–221. doi:10.1007/s00217-004-1070-7.

[46] F. Savorani, G. Tomasi, S.B. Engelsen, icoshift: A versatile tool for the rapid alignment of 1D NMR spectra, J. Magn. Reson. 202 (2010) 190–202. doi:10.1016/j.jmr.2009.11.012.

[47] F. Savorani, G. Tomasi, S.B. Engelsen, Alignment of 1D NMR Data using the iCoshift Tool: A Tutorial, in: J. van Duynhoven, P.S. Belton, Webb. G.A., H. van As (Eds.), Magn. Reson. Food Sci. Food Thought, Royal Society of Chemistry, 2013: pp. 14–24. doi:10.1039/9781849737531-00014.

[48] A. de Juan, R. Tauler, Multivariate Curve Resolution (MCR) from 2000: Progress in Concepts and Applications, Crit. Rev. Anal. Chem. 36 (2006) 163–176.

844 doi:10.1080/10408340600970005.

845 [49] F. Savorani, M.A. Rasmussen, Å. Rinnan, S.B. Engelsen, Interval-Based Chemometric
846 Methods in NMR Foodomics, in: Cyril Ruckebusch (Ed.), Data Handl. Sci. Technol.,
847 Elsevier, 2013: pp. 449–486. doi:10.1016/B978-0-444-59528-7.00012-0.

848 [50] M. Ankerst, M.M. Breunig, H.-P. Kriegel, J. Sander, OPTICS: Ordering Points To Identify
849 the Clustering Structure, in: Proc. 1999 ACM SIGMOD Int. Conf. Manag. Data - SIGMOD
850 '99, ACM Press, New York, New York, USA, 1999: pp. 49–60. doi:10.1145/304182.304187.

851 [51] M. Daszykowski, B. Walczak, D.L. Massart, Looking for Natural Patterns in Analytical
852 Data. 2. Tracing Local Density with OPTICS, J. Chem. Inf. Model. 42 (2002) 500–507.
853 doi:10.1021/CI010384S.

854 [52] M. Daszykowski, B. Walczak, Density-Based Clustering Methods, in: Compr. Chemom.,
855 2009: pp. 635–654. doi:10.1016/B978-044452701-1.00067-3.

856 [53] M. Bevilacqua, R. Bucci, A.D. Magrì, A.L. Magrì, F. Marini, Data Fusion for Food
857 Authentication. Combining near and Mid Infrared to Trace the Origin of Extra Virgin Olive
858 Oils, NIR News. 24 (2013) 12–15. doi:10.1255/nirn.1355.

859 [54] F. Marini, Artificial neural networks in foodstuff analyses: Trends and perspectives A
860 review, Anal. Chim. Acta. 635 (2009) 121–131. doi:10.1016/J.ACA.2009.01.009.

861 [55] R. Vitale, O.E. de Noord, A. Ferrer, A kernel-based approach for fault diagnosis in batch
862 processes, J. Chemom. 28 (2014) S697–S707. doi:10.1002/cem.2629.

863 [56] B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear Component Analysis as a Kernel
864 Eigenvalue Problem, Neural Comput. 10 (1998) 1299–1319.
865 doi:10.1162/089976698300017467.

866 [57] J. Jaumot, A. de Juan, R. Tauler, MCR-ALS GUI 2.0: New features and applications,
867 Chemom. Intell. Lab. Syst. 140 (2015) 1–12. doi:10.1016/j.chemolab.2014.10.003.

868 [58] Analytica-EBC; European Brewery Convention: Analytica-EBC; Fachverlag Hans Carl,
869 (2008). http://analytica-ebc.com/index.php?mod=contents&scat=16.

[59] D. Intelmann, G. Haseleu, A. Dunkel, A. Lagemann, A. Stephan, T. Hofmann, Comprehensive Sensomics Analysis of Hop-Derived Bitter Compounds during Storage of Beer, J. Agric. Food Chem. 59 (2011) 1939–1953. doi:10.1021/jf104392y.

[60] I. Stanimirova, C. Boucon, B. Walczak, Relating gas chromatographic profiles to sensory measurements describing the end products of the Maillard reaction, Talanta. 83 (2011) 1239–1246. doi:10.1016/J.TALANTA.2010.09.018.

[61] J.M. Andrade, M.P. Gómez-Carracedo, W. Krzanowski, M. Kubista, Procrustes rotation in analytical chemistry, a tutorial, Chemom. Intell. Lab. Syst. 72 (2004) 123–132. doi:10.1016/J.CHEMOLAB.2004.01.007.

[62] P.D. Wentzell, S. Hou, C.S. Silva, C.C. Wicks, M.F. Pimentel, Procrustes rotation as a diagnostic tool for projection pursuit analysis, Anal. Chim. Acta. 877 (2015) 51–63. doi:10.1016/j.aca.2015.03.006.

[63] R. Todeschini, D. Ballabio, V. Consonni, Distances and Other Dissimilarity Measures in Chemometrics, Encycl. Anal. Chem. Appl. Theory Instrum. (2015) 1–34. doi:10.1002/9780470027318.a9438.

**List of Tables**



**List of Figures**

in the legend; (c) PC1 vs PC2 score plot coloured according to beer colour intensity: one intensity value for each spectrum is calculated by taking the average of intensity values in the interval 430±5 nm. The background patches in (b) highlight the OPTICS groups defined in (a).

**Figure 3.** NIR spectra dataset: (a) Reachability Plot, bars are colored by beer style, as detailed in the legend; and (b) PCA score plot colored by ABV content. Samples in both in (a) and (b) were reordered according to OPTICS order.

**Figure 4.** Heatmap of NMR features with Reachability Plots: variable's RP on the left side ($k$ = 3), samples' RP on top ($k$ = 5). OPTICS in the variables' direction was performed on the correlation matrix, instead of the variables themselves. In the central part of the figure it is shown the heatmap obtained by reordering both the samples and the variables according to the respective OPTICS sequences. The dataset was normalized between zero and one to enhance its visual representation and interpretability.

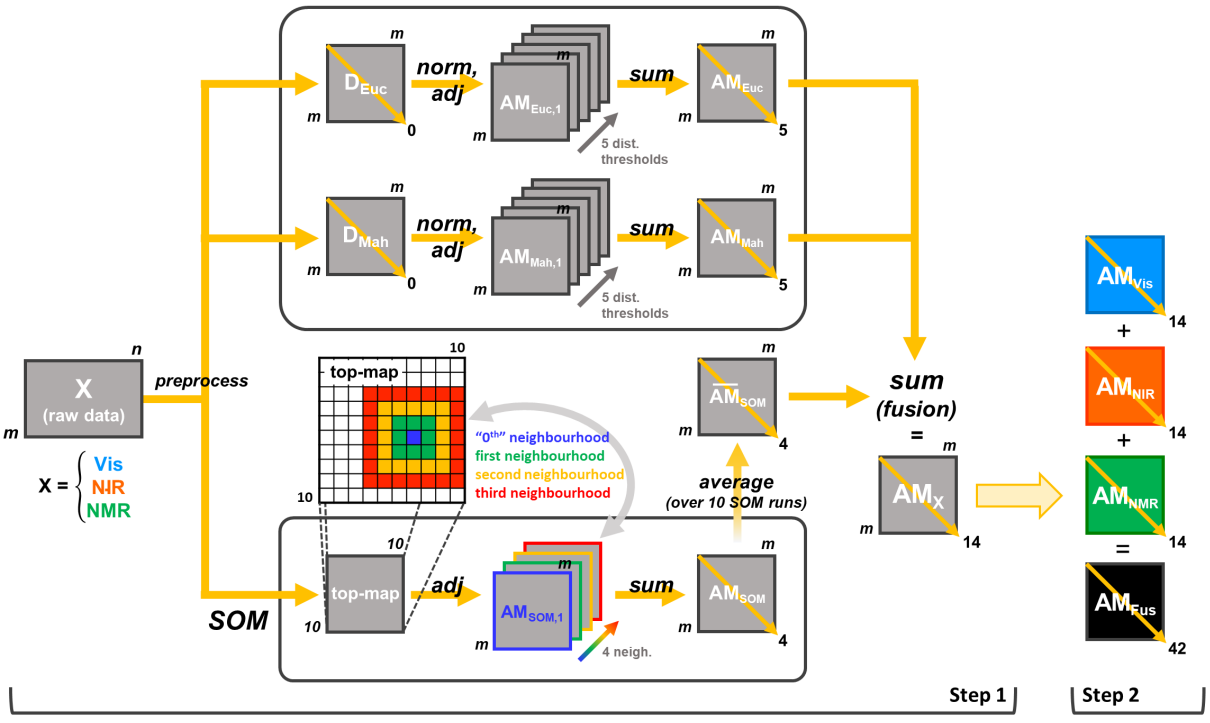
**Figure 5.** Mid-level fused dataset: (a) Reachability Plot, (b) PC2 vs PC1 score plot, (c) PC2 vs PC1 loadings plot; colours and symbols explained in the legend on the plot. The area highlighted in orange corresponds to the most coloured beer samples.

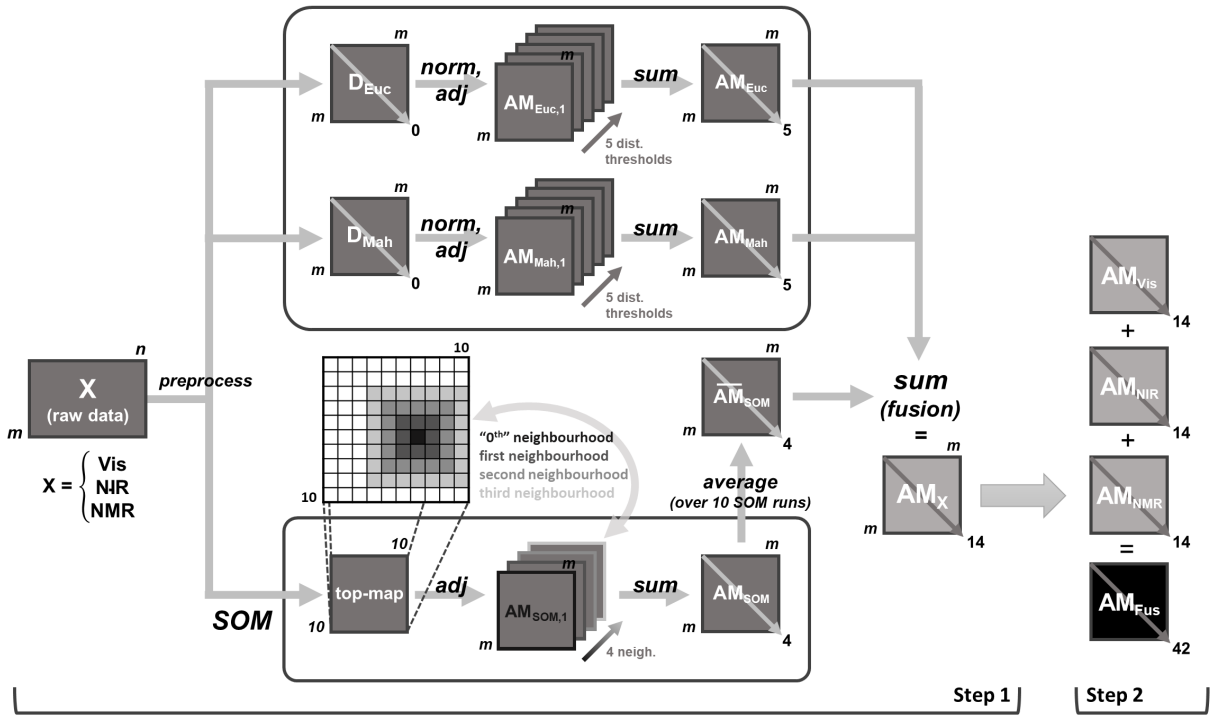
**Figure 6.** Fused Adjacency Matrix: (a) Reachability Plot; (b) PC3 vs PC1 score plot, colours and symbols explained in the legend on the plot; the background patches in (b) highlight the OPTICS groups defined in (a). (c) PC4 vs PC1 score plot, colours and symbols explained in the legend on the plot; the curved arrow in (c) describes the beer colour intensity trend; the red background patches in (c) highlight possible new groups.

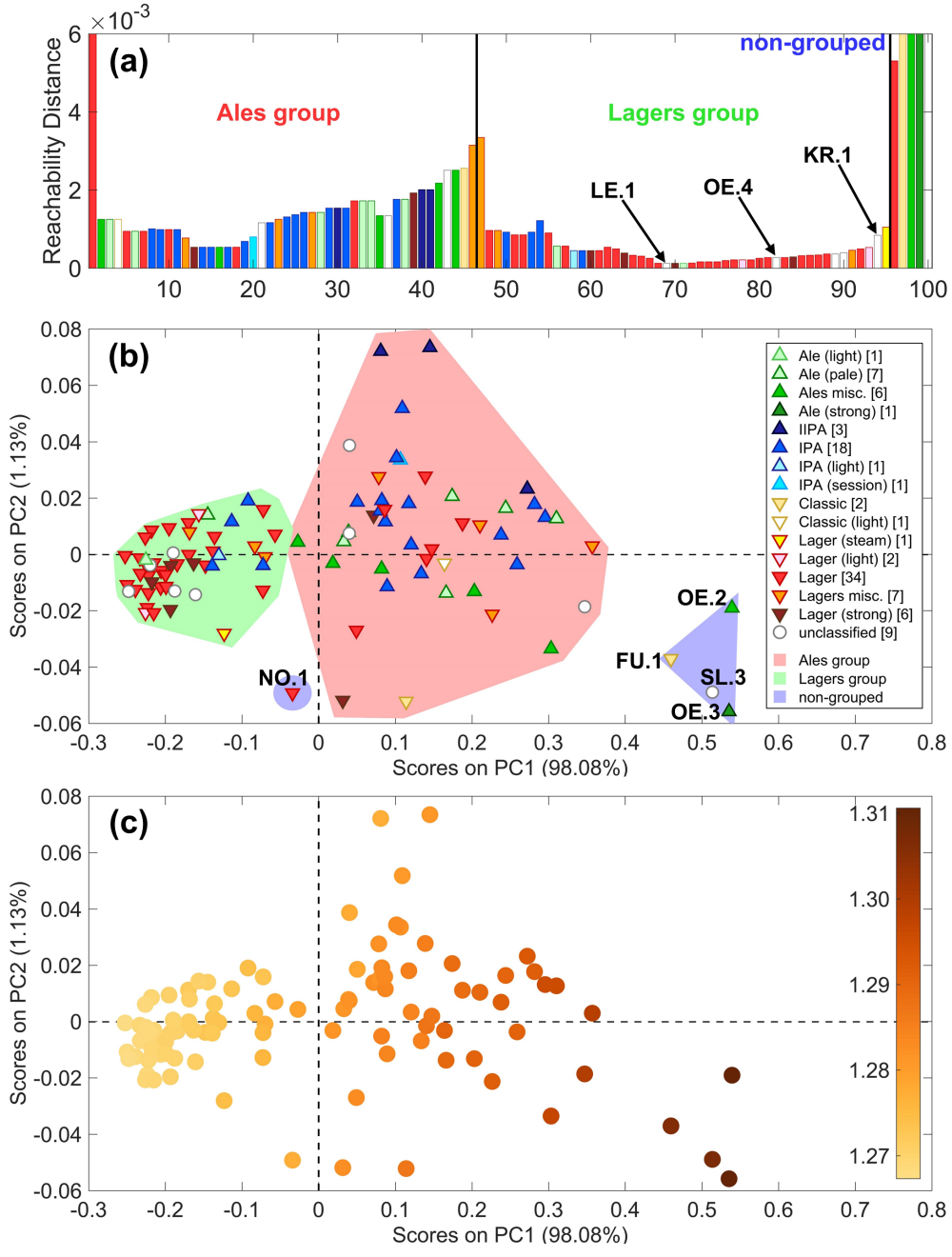
**Figure 7.** Heatmap of NMR features with Reachability Plots: variables' RP on the left side (OPTICS performed as described in the caption of Figure 4), samples' RP on top ($k$ = 5). The samples are reordered according to the OPTICS sequence obtained from the Fused Adjacency Matrix (as in Figure 6). The dataset was normalized between zero and one to enhance its visual representation and interpretability.

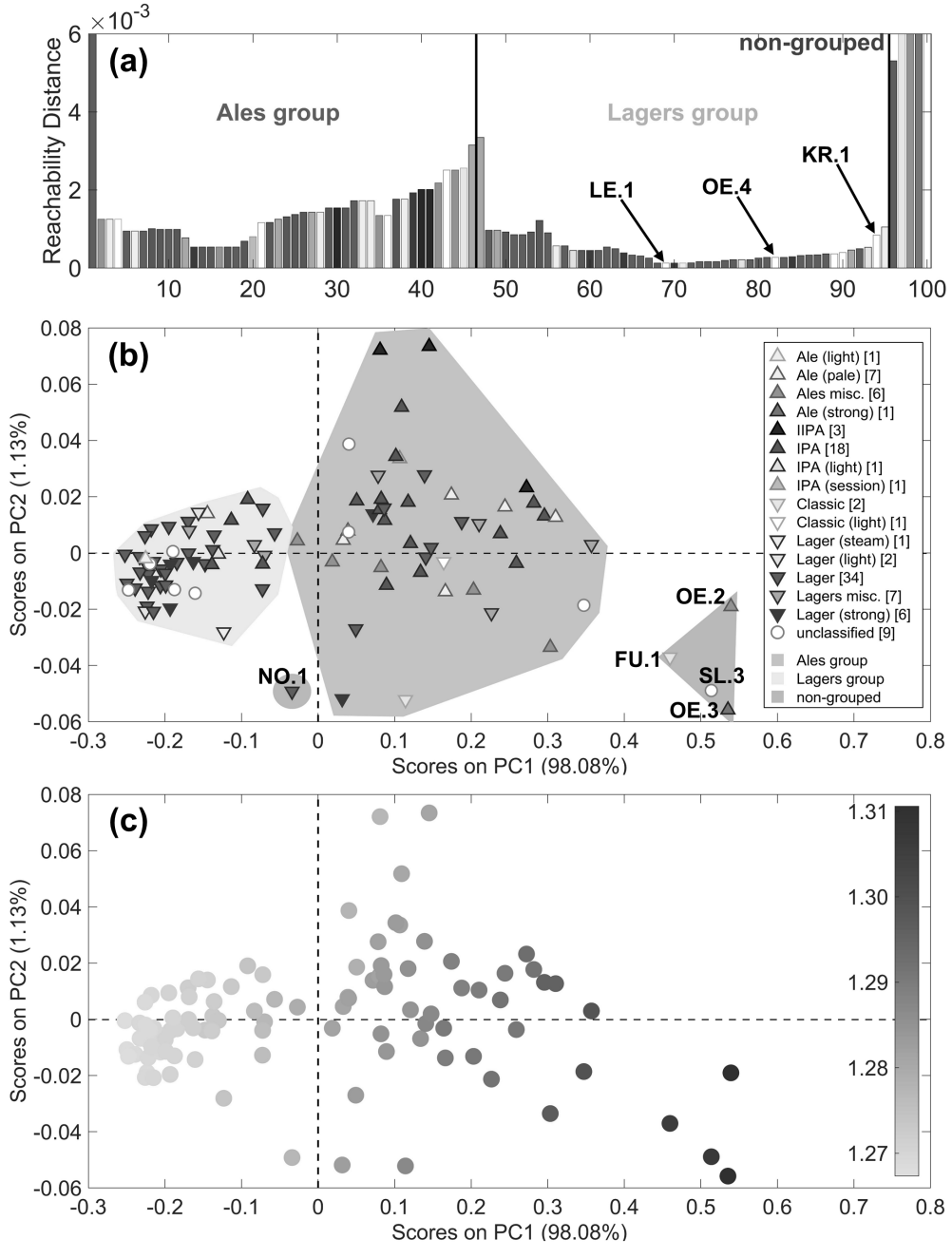
**Table 1** Comparison summary (*ordered by increasing ABV)

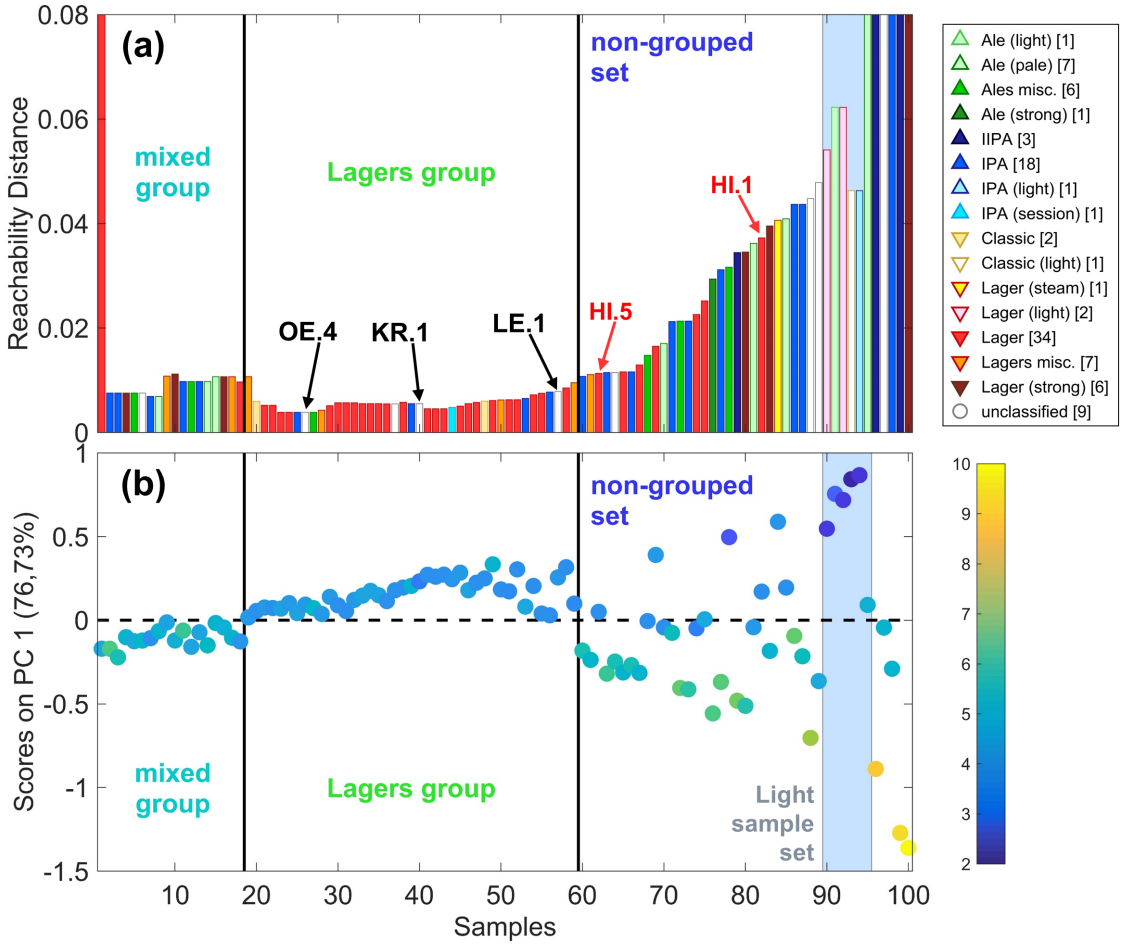
| | **Visible** | **NIR** | **NMR** (Fig.4) | **Mid-level data fusion** | **Fused Adjacency Matrix** $AM_{Fus}$ |
|---|---|---|---|---|---|
| **Lagers group** | Dense cluster in RP. (Fig.2a)<br>Grouped in PCA. (negative scores, Fig.2b) | Slightly defined in RP. (Fig.3a)<br>At positive PC1 scores, close to zero. (Fig.3b) | Slightly defined in RP.<br>Medium to low variable values in general.<br>Some sub-groups; contains the Light samples set as a sub-group. | Slightly defined in RP. (Fig.5a)<br>At negative PC1 scores. (Fig.5b) | Defined cluster in RP. (Fig.6a)<br>HI samples grouped and well-ordered together in RP. (Fig.6a)<br>Grouped in PCA. (Fig.6b) |
| **Unclassified ○ fresh/summer beers in the Lagers group**<br>(most frequent ones: LE.1, OE.4, KR.1) | LE.1, OE.4, WI.2, SK.4, KR.1 (Fig.2a) | OE.4, UG.3, KR.1, LE.1 (Fig.3a) | LE.1, OE.4<br>KR.1 is in the non-grouped set. | LE.1, OE.4, KR.1, TY.3 (Fig.5a) | LE.1, OE.4, KR.1, WI.2 (Fig.6b) |
| **Light samples set**<br>(KR.2, LE.2, FB.2, TO.4, NO.2)* | All in the Lagers group. (Fig.2b)<br>Generally lighter colours. (Fig.2c) | Quite grouped in RP. (Fig.3a)<br>All extreme on PC1. (Fig.3b) | Grouped in RP.<br>Included in the Lagers group.<br>Low values in general. | Not grouped in RP. (Fig.5a)<br>Grouped in PCA. (Fig.5b) | Not grouped in RP. (Fig.6a)<br>Grouped in PCA. (Fig.6b) |
| **Lager Strong**<br>four low-ABV: MA.3, SI.9, MA.5, MA.6<br>two high-ABV: MA.2, FB.3 | Four low-ABV in the Lagers group, low-colour. (Fig.2a-b)<br>Two high-ABV in the non-grouped set, mid-colour. (Fig.2a-b) | Three in the mixed group. (Fig.3a / SI.9, MA.5, MA.3)<br>Three in the non-grouped set. (Fig.3a / *MA.6*, MA.2, FB.3) | All in the Lagers group. | Four low-ABV in the Lagers group. (Fig.5a)<br>Two high-ABV quite far in the non-grouped set. (Fig.5a) | Four low-ABV close to the Lagers group in PCA. (Fig.6b)<br>Two high-ABV close to the Ales. (Fig.6b) |
| **ABV trend** | Not found. | Very well described by PC1. (Fig.3b) | Found in PCA (Fig.S2a); probably reflecting the sugar content. | Found in PC1-PC2 score plot. (Fig.5b) | Found in a transformed way. (Fig.S3) |
| **Colour trend** | Clearly found along PC1. (Fig.2c) | Not found. | Not found. | In PCA the stronger colored samples lie at positive PC1 and PC2 scores. (Fig.5b) | Nicely represented by PC1 and PC4. (Fig.6c) |

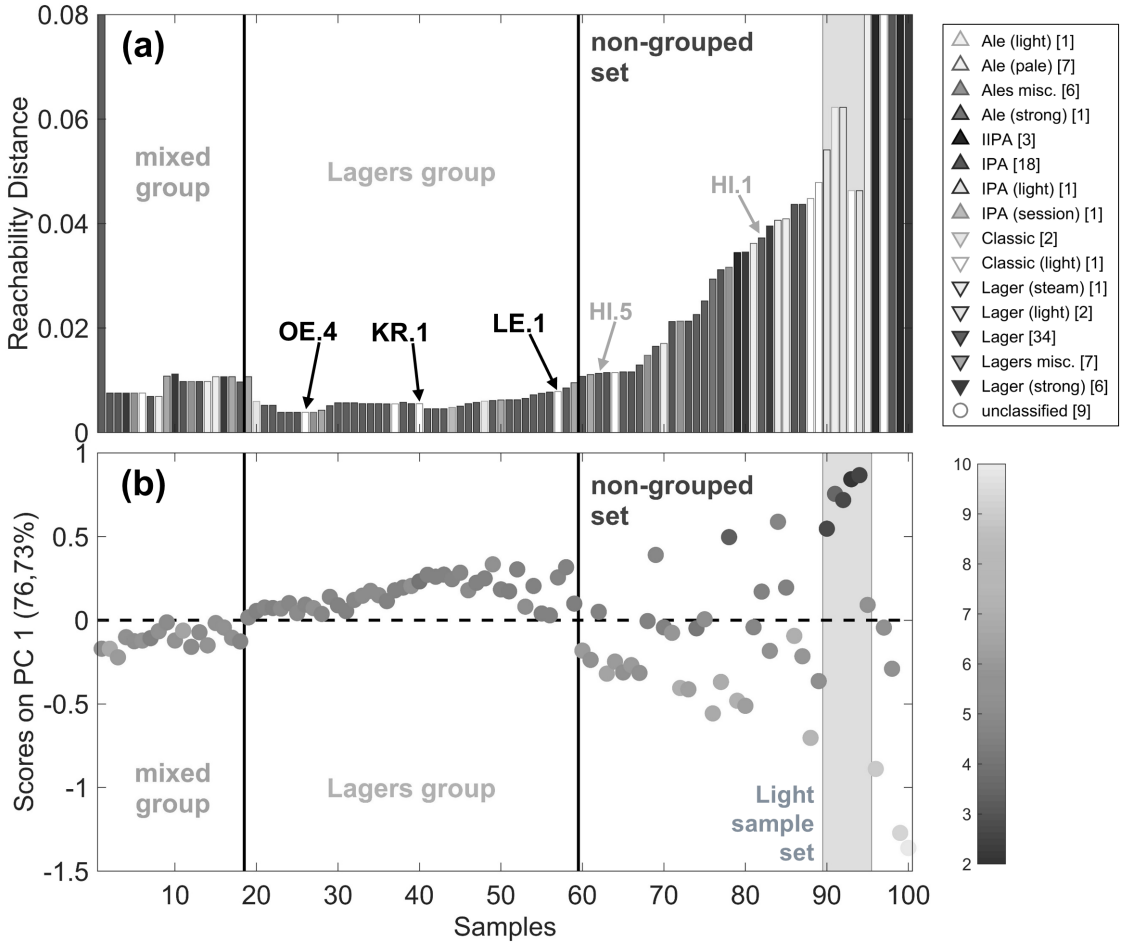


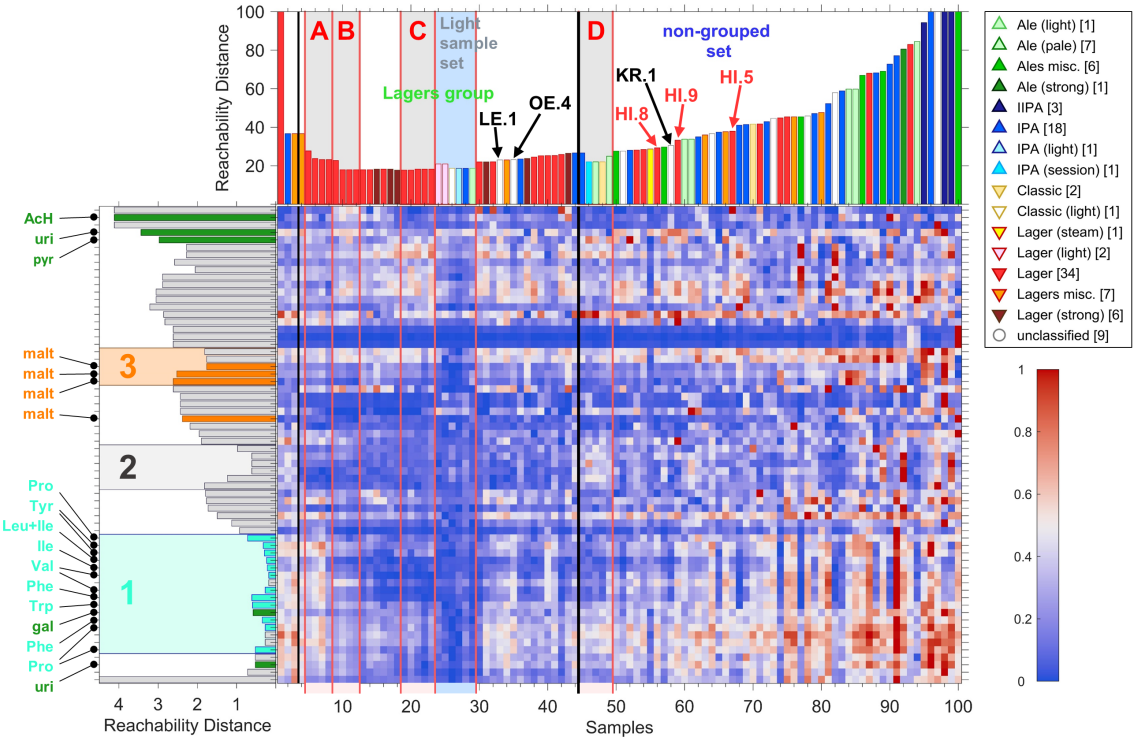


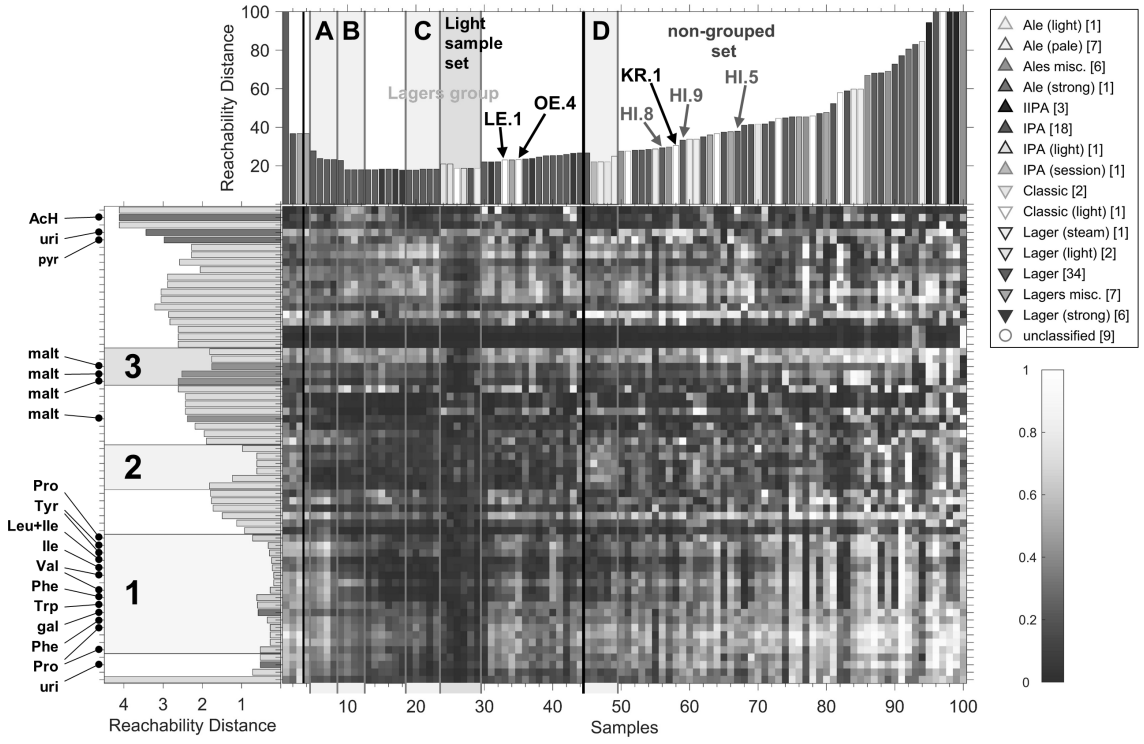

(a)
×10-3
Reachability Distance
Ales group
Lagers group
non-grouped
LE.1
OE.4
KR.1
(b)
Scores on PC2 (1.13%)
Scores on PC1 (98.08%)
Ale (light) [1]
Ale (pale) [7]
Ales misc. [6]
Ale (strong) [1]
IIPA [3]
IPA [18]
IPA (light) [1]
IPA (session) [1]
Classic [2]
Classic (light) [1]
Lager (steam) [1]
Lager (light) [2]
Lager [34]
Lagers misc. [7]
Lager (strong) [6]
unclassified [9]
Ales group
Lagers group
non-grouped
NO.1
FU.1
OE.2
SL.3
OE.3
(c)
Scores on PC2 (1.13%)
Scores on PC1 (98.08%)

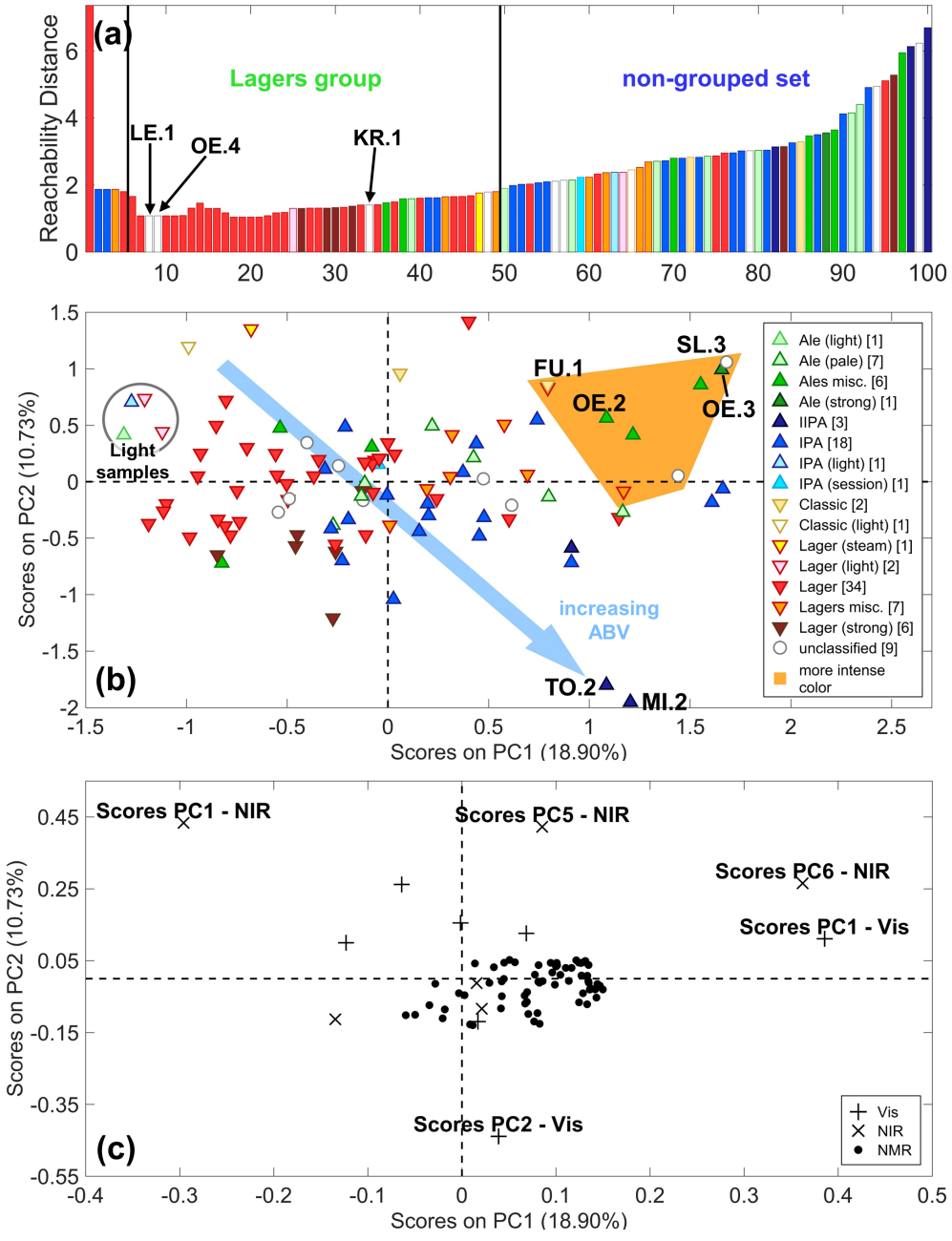

(a)
×10^-3
Reachability Distance
Ales group
Lagers group
non-grouped
LE.1
OE.4
KR.1
(b)
Scores on PC2 (1.13%)
Scores on PC1 (98.08%)
Ale (light) [1]
Ale (pale) [7]
Ales misc. [6]
Ale (strong) [1]
IIPA [3]
IPA [18]
IPA (light) [1]
IPA (session) [1]
Classic [2]
Classic (light) [1]
Lager (steam) [1]
Lager (light) [2]
Lager [34]
Lagers misc. [7]
Lager (strong) [6]
unclassified [9]
Ales group
Lagers group
non-grouped
NO.1
FU.1
OE.2
SL.3
OE.3
(c)
Scores on PC2 (1.13%)
Scores on PC1 (98.08%)

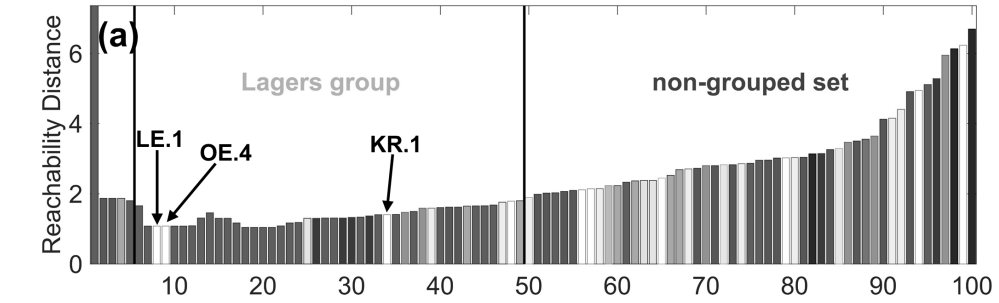

(a)
mixed group
Lagers group
non-grouped set
HI.1
HI.5
OE.4
KR.1
LE.1
Reachability Distance
0.08
0.06
0.04
0.02
0
Ale (light) [1]
Ale (pale) [7]
Ales misc. [6]
Ale (strong) [1]
IIPA [3]
IPA [18]
IPA (light) [1]
IPA (session) [1]
Classic [2]
Classic (light) [1]
Lager (steam) [1]
Lager (light) [2]
Lager [34]
Lagers misc. [7]
Lager (strong) [6]
unclassified [9]
(b)
non-grouped set
mixed group
Lagers group
Light sample set
Scores on PC 1 (76,73%)
1
0.5
0
-0.5
-1
-1.5
10
9
8
7
6
5
4
3
2
10
20
30
40
50
60
70
80
90
100
Samples

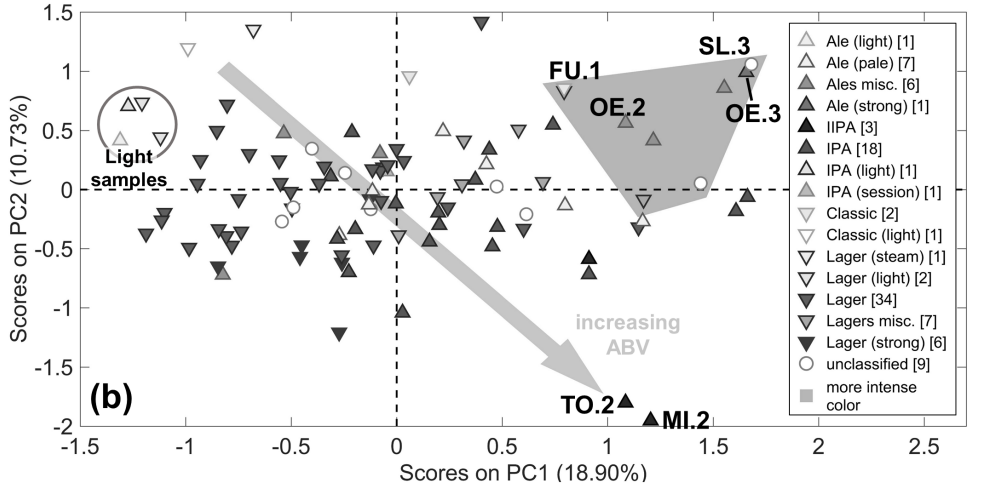

(a)
mixed group
Lagers group
non-grouped set
HI.1
HI.5
OE.4
KR.1
LE.1
Reachability Distance
0.08
0.06
0.04
0.02
0
(b)
non-grouped set
mixed group
Lagers group
Light sample set
Scores on PC 1 (76,73%)
1
0.5
0
-0.5
-1
-1.5
10
20
30
40
50
60
70
80
90
100
Samples
Ale (light) [1]
Ale (pale) [7]
Ales misc. [6]
Ale (strong) [1]
IIPA [3]
IPA [18]
IPA (light) [1]
IPA (session) [1]
Classic [2]
Classic (light) [1]
Lager (steam) [1]
Lager (light) [2]
Lager [34]
Lagers misc. [7]
Lager (strong) [6]
unclassified [9]
10
9
8
7
6
5
4
3
2

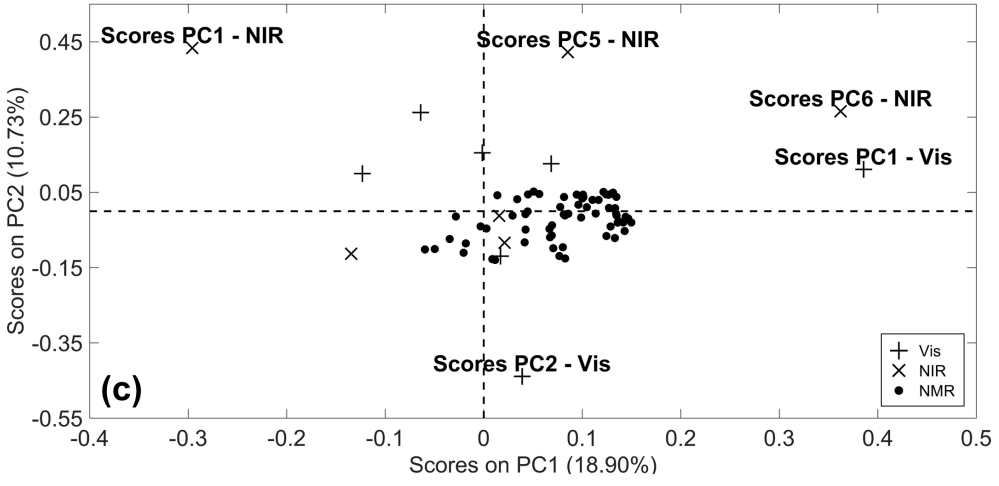


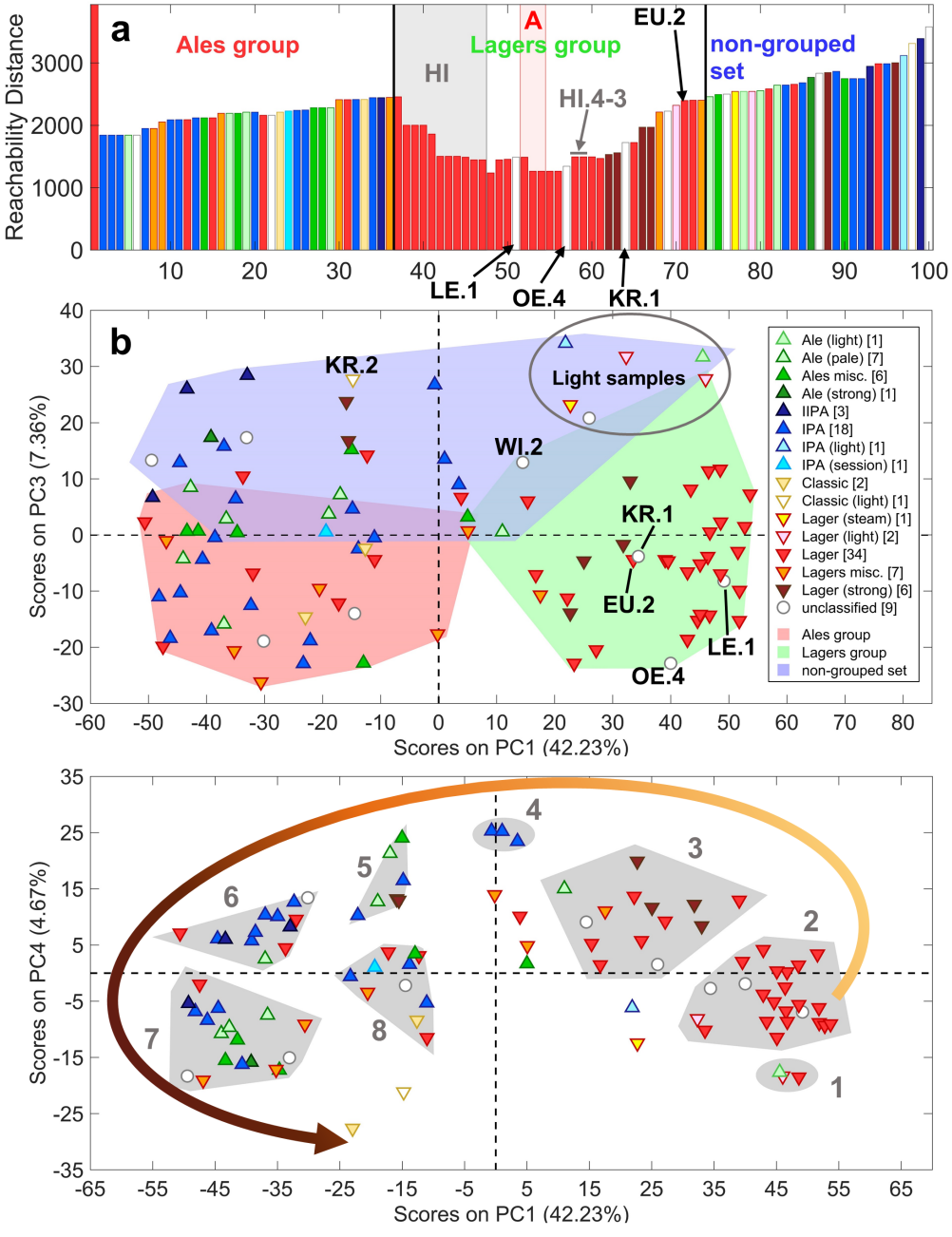


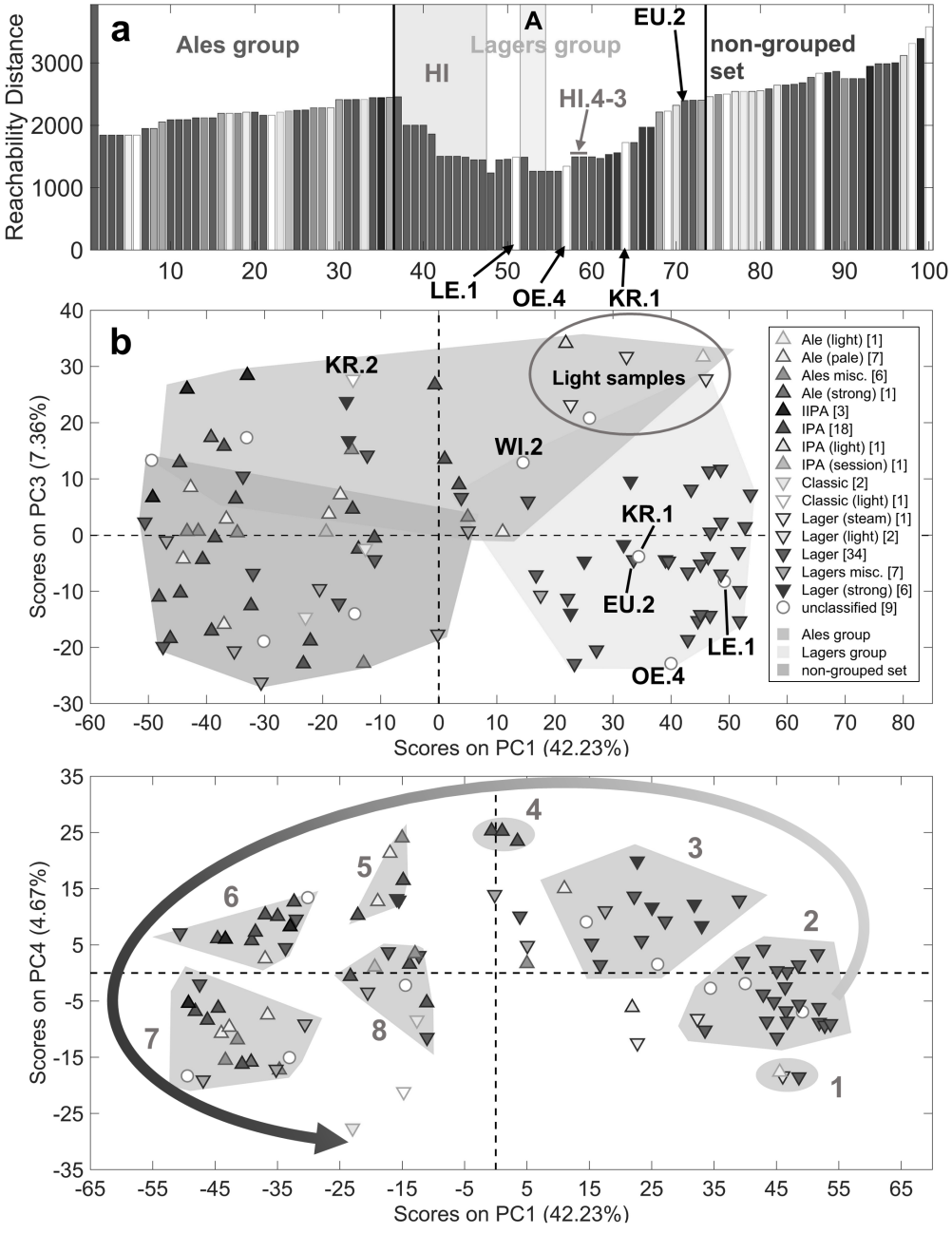

(a)
Reachability Distance
Lagers group
non-grouped set
LE.1
OE.4
KR.1
(b)
Scores on PC2 (10.73%)
Scores on PC1 (18.90%)
Light samples
FU.1
SL.3
OE.2
OE.3
increasing ABV
TO.2
MI.2
Ale (light) [1]
Ale (pale) [7]
Ales misc. [6]
Ale (strong) [1]
IIPA [3]
IPA [18]
IPA (light) [1]
IPA (session) [1]
Classic [2]
Classic (light) [1]
Lager (steam) [1]
Lager (light) [2]
Lager [34]
Lagers misc. [7]
Lager (strong) [6]
unclassified [9]
more intense color
(c)
Scores PC1 - NIR
Scores PC5 - NIR
Scores PC6 - NIR
Scores PC1 - Vis
Scores PC2 - Vis
Vis
NIR
NMR

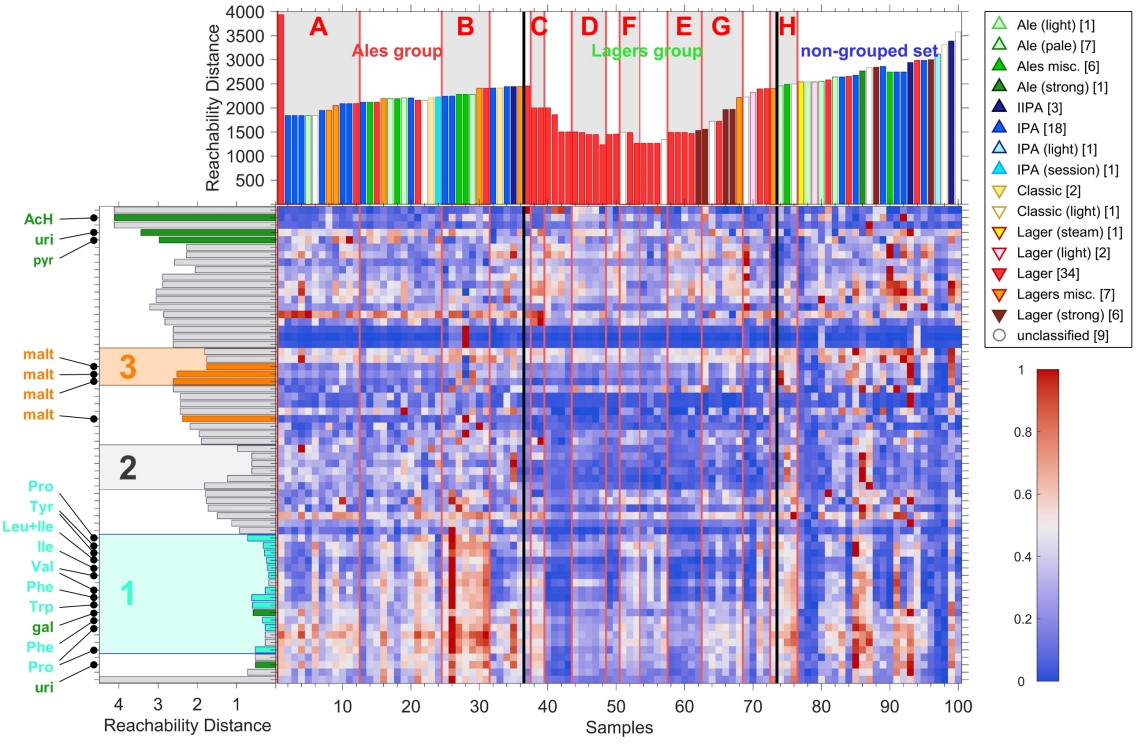

(a)
Reachability Distance
Lagers group
non-grouped set
LE.1
OE.4
KR.1
0
2
4
6
10
20
30
40
50
60
70
80
90
100

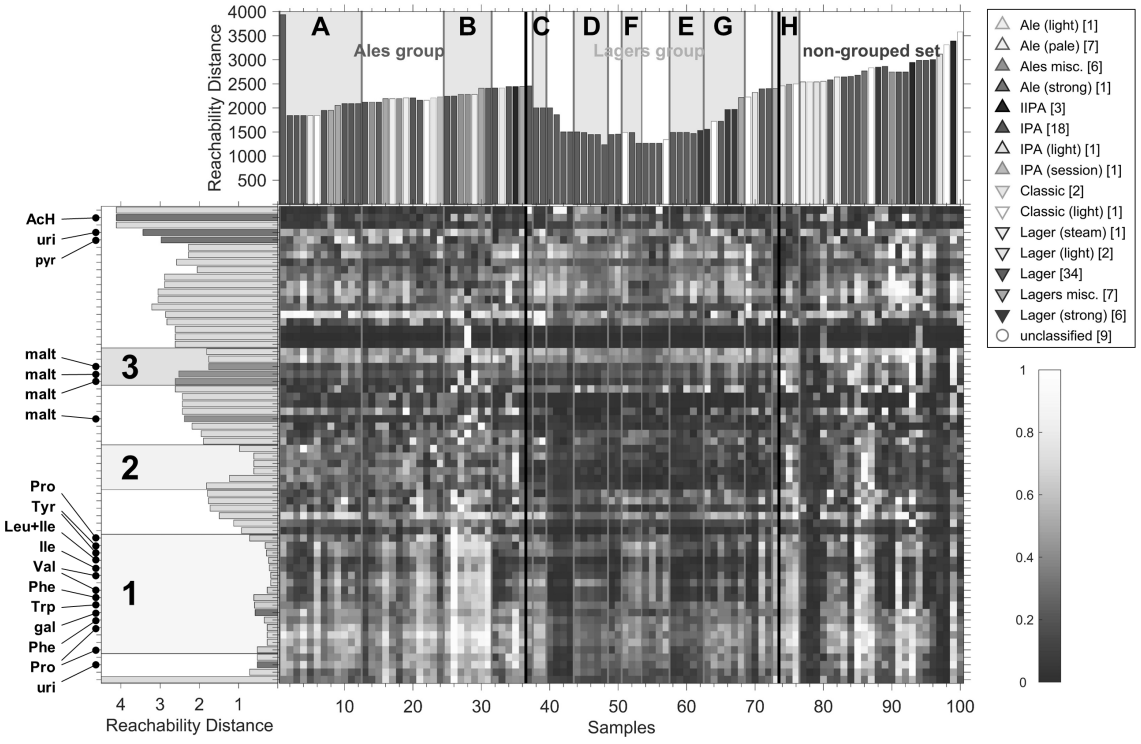

(b)
Scores on PC2 (10.73%)
Scores on PC1 (18.90%)
Light samples
FU.1
SL.3
OE.2
OE.3
increasing ABV
TO.2
MI.2
Ale (light) [1]
Ale (pale) [7]
Ales misc. [6]
Ale (strong) [1]
IIPA [3]
IPA [18]
IPA (light) [1]
IPA (session) [1]
Classic [2]
Classic (light) [1]
Lager (steam) [1]
Lager (light) [2]
Lager [34]
Lagers misc. [7]
Lager (strong) [6]
unclassified [9]
more intense color


(c)
Scores PC1 - NIR
Scores PC5 - NIR
Scores PC6 - NIR
Scores PC1 - Vis
Scores PC2 - Vis
Scores on PC2 (10.73%)
Scores on PC1 (18.90%)
Vis
NIR
NMR




a
Ales group
HI
A
Lagers group
HI.4-3
EU.2
non-grouped set
Reachability Distance
LE.1
OE.4
KR.1
b
KR.2
Light samples
WI.2
KR.1
EU.2
LE.1
OE.4
Ale (light) [1]
Ale (pale) [7]
Ales misc. [6]
Ale (strong) [1]
IIPA [3]
IPA [18]
IPA (light) [1]
IPA (session) [1]
Classic [2]
Classic (light) [1]
Lager (steam) [1]
Lager (light) [2]
Lager [34]
Lagers misc. [7]
Lager (strong) [6]
unclassified [9]
Ales group
Lagers group
non-grouped set
Scores on PC3 (7.36%)
Scores on PC1 (42.23%)
Scores on PC4 (4.67%)
Scores on PC1 (42.23%)





1. A new approach to enhance information extraction from highly complex datasets is proposed.
2. The approach is based on the fusion of adjacency matrices obtained from different clustering strategies.
3. Information extracted from different data blocks is fused, so the approach can also be a method for high-level data fusion.
4. Visible, NIR and NMR data of beer samples are used as a benchmark for testing the approach.
5. The approach can highlight groups in a better way than the single-block and mid-level data-fusion approaches.

**Declaration of interests**

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

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