

Fused Adjacency Matrices to enhance information extraction: the beer benchmark

Original

Fused Adjacency Matrices to enhance information extraction: the beer benchmark / Cavallini, Nicola; Savorani, Francesco; Bro, Rasmus; Cocchi, Marina. - In: ANALYTICA CHIMICA ACTA. - ISSN 0003-2670. - ELETTRONICO. - 1061:(2019), pp. 70-83. [10.1016/j.aca.2019.02.023]

Availability:

This version is available at: 11583/2815371 since: 2020-04-22T17:48:48Z

Publisher:

Elsevier

Published

DOI:10.1016/j.aca.2019.02.023

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

Elsevier postprint/Author's Accepted Manuscript

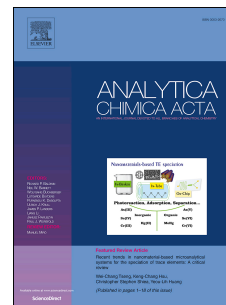
© 2019. This manuscript version is made available under the CC-BY-NC-ND 4.0 license
<http://creativecommons.org/licenses/by-nc-nd/4.0/>. The final authenticated version is available online at:
<http://dx.doi.org/10.1016/j.aca.2019.02.023>

(Article begins on next page)

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*

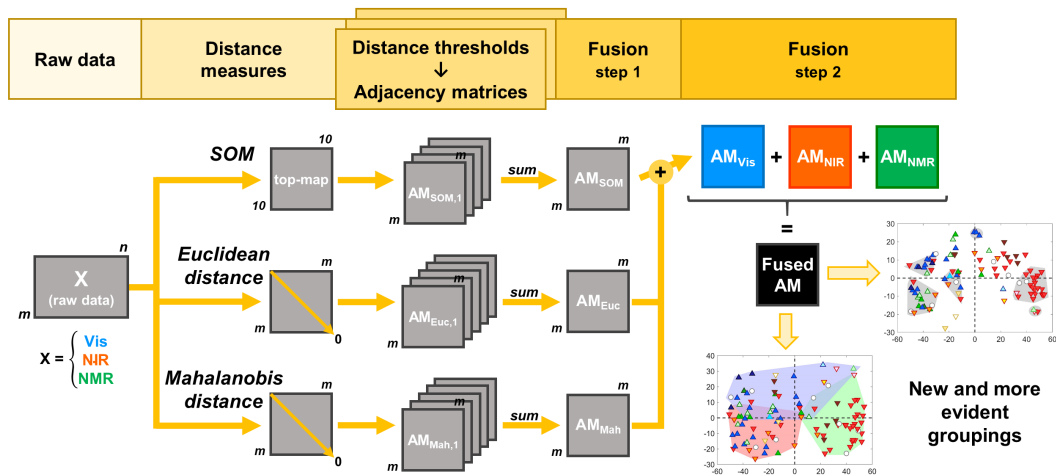
Received Date: 11 May 2018

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>.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



ACA-18-1339R1 – underlined revision1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

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

Authors and affiliations: Nicola Cavallini^{1,2}, Francesco Savorani³, Rasmus Bro², Marina Cocchi¹

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 |

| | | |
|----|-----|-------------------------------|
| 42 | RD | Reachability Distance |
| 43 | RP | Reachability Plot |
| 44 | SOM | Kohonen's Self-Organizing Map |
| 45 | Vis | Visible |

46

47 **1. Introduction**

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

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

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

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

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

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

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

94 general purpose was that of evidencing both peculiar beers and groups of similar samples, in other
95 words mining all the possible similarities/peculiarities, just based on the chemical fingerprint
96 acquired. Beer has been the object of several studies, mostly focused on a specific beer type aiming
97 either at gathering the composition [24–27] or controlling the brewing process [28,29]. To achieve
98 these aims very different analytical techniques have been applied: NMR [24,25,27,30–32], LC-MS
99 [30,33,34], GC-MS [35,36], vibrational (NIR and IR) [24,26,28,37] and UV-Visible [38]
100 spectroscopies. The benchmark beer dataset consists of three data blocks obtained from different
101 spectroscopic techniques: Visible (Vis), Near Infra-Red (NIR) and Nuclear Magnetic Resonance
102 (NMR), the latter as interval-resolved data.

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

111

112 **2. Materials and methods**

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

116

117 **2.1. Experimental**

118 **2.1.1. Sampling and sample preparation**

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

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

128 A collection of 2 mL eppendorfs was directly prepared from the original commercial containers
129 (cans or glass bottles). Three eppendorfs for each beer sample were prepared and kept frozen at –
130 20°C. The initial steps of thawing and degassing [24] were common across all the different
131 spectroscopic techniques, and were performed as follows: 1) 10 minutes thawing in water bath at
132 room temperature; 2) 20 minutes of ultrasonic bath in water at room temperature. Since all the
133 specimens were clear (i.e. no suspended particles), filtration was not required. The degassing
134 procedure is highly recommended by literature studies [24,25,27] and it is aimed at reducing
135 measurement interferences due to bubble formation both on the NIR sample vessel and within the
136 NMR tubes.

137

138 **2.1.2. Vis-NIR data acquisition and preprocessing**

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

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

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

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

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

163

164 **2.1.3. ¹H-NMR data acquisition and preprocessing**

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

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

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

183 All the specimens were prepared to contain 10% D₂O, 0,02% of sodium-3-
184 (trimethylsilyl)propionate-*d*₄ (TSP-*d*₄) as a chemical shift reference [24,25,27,30–32,43,44] and
185 20% phosphate buffer (pH = 3.55). The required volume for the NMR tubes was 600 μL, and it was
186 obtained by mixing: 420 μL of beer specimen, 60 μL of D₂O and 120 μL of phosphate buffer in
187 H₂O. Spectra were acquired in random order with respect to samples and replicates.

188 Duarte et al. [43] studied the composition of ale and lager beers, and reported pH values within the
189 3.7–4.4 interval. The addition of a phosphate buffer (pH = 3.55) was aimed to obtain a set of
190 specimens with more homogeneous pH values, so that the signal's horizontal shifts across spectra,
191 due to the different protonation forms of compounds such as organic acids [31,32], could be
192 reduced.

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

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

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

212

213 **2.2. Data Analysis**

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

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

221

222 **2.2.1. Data reduction**

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

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

232

233 **2.2.2. Kohonen's Self-Organizing Maps (SOM)**

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

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

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

255

256 **2.2.3. Ordering Points to Identify the Clustering Structure (OPTICS)**

257 OPTICS [50–52] is a density-based clustering method aimed at revealing the data clustering
258 structure. This method consists of an iterative procedure that only needs an initial input parameter,
259 namely k , which is the minimal number of objects forming a cluster. Daszykowski and Walczak
260 [52] suggested a rule of thumb for selecting k :

$$261 \quad (1) \quad k = \text{integer} \left(\frac{m}{25} \right)$$

262 where m is the number of samples.

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

268 At each iteration, the OPTICS algorithm selects one object and compares it with all the objects that
269 have not been processed yet. This is done by computing all the pairwise Euclidean distances. Then,
270 the next object to be processed is selected among the k nearest neighbours: the distance at which
271 this next object is found becomes its RD, which is stored unchanged until the end of the procedure.

272 The final output is therefore a set of RD values, which can be plotted as bars in the RP. A cluster is
273 formed by objects that happen to be very close to each other, so it can be expected that these objects

274 would have, on average, a similar number of neighbours at similar distances, i.e. they would have
275 similar neighbourhoods. These short distances among neighbours result in very similar RD values.
276 When a cluster has been processed, then the next object would likely belong to another cluster: the
277 next RD value in the processing sequence is therefore going to be larger than the values preceding
278 it, which are related to previous cluster. This “jump” from one cluster to another is graphically
279 recognizable in the RP because it corresponds to a very high bar. Clusters therefore appear as
280 hollows created by groups of samples sharing similarly low RDs.
281 It is important to consider that the RP does not explicitly cluster the objects [52], but it rather allows
282 deducing the number of clusters in the data.

283

284 **2.2.4. Mid-level data fusion**

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

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

294

295 **2.2.5. Fused Adjacency Matrix approach**

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

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

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

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

$$320 \quad (2) \quad \mathbf{X} \rightarrow \mathbf{D}_{\text{Euc/Mah}} \rightarrow thr = [0.05, 0.1, 0.2, 0.3, 0.4] \rightarrow \mathbf{AM}_{\text{Euc/Mah}} = \sum_{thr=1}^5 \mathbf{AM}_{thr}$$

$$321 \quad (3) \quad \mathbf{X} \rightarrow \text{SOM} \rightarrow \text{topmap} \rightarrow g = [0, 1, 2, 3] \rightarrow \mathbf{AM}_{\text{SOM}} = \sum_{g=0}^3 \mathbf{AM}_{g, \text{neigh}}$$

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

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

$$331 \quad (4) \quad \mathbf{AM}_X = \mathbf{AM}_{\text{Euc}} + \mathbf{AM}_{\text{Mah}} + \mathbf{AM}_{\text{SOM}} \quad (X = \text{Vis, NIR, NMR})$$

$$332 \quad (5) \quad \mathbf{AM}_{\text{Fus}} = \sum_X \mathbf{AM}_X = \mathbf{AM}_{\text{Vis}} + \mathbf{AM}_{\text{NIR}} + \mathbf{AM}_{\text{NMR}}$$

333 Figure 1 provides a graphical representation of the whole Fused Adjacency Matrix approach. For a
 334 given data block \mathbf{X} , its corresponding output is the matrix \mathbf{AM}_X (Equation 3). When more than one
 335 \mathbf{X} data blocks are available (like in the benchmark case presented in this work, where $X = \text{Vis, NIR,}$
 336 NMR), the resulting \mathbf{AM}_X matrices can be combined using, again, a sum rule ([22], equation 4).
 337 The result is the Fused Adjacency Matrix \mathbf{AM}_{Fus} , depicted in black in Figure 1. In this work, the
 338 values in \mathbf{AM}_{Fus} vary between zero and 42, as a result of summing a total of 42 AMs which have
 339 ones on their diagonal. Prior to analysis, the Fused Adjacency Matrix \mathbf{AM}_{Fus} was double centered
 340 [55] so that:

$$341 \quad (6) \quad \mathbf{AM}_{\text{Fus,cent}} = \mathbf{AM}_{\text{Fus}} - \overline{\mathbf{AM}}_{\text{Fus},m} - \overline{\mathbf{AM}}_{\text{Fus},n} + \overline{\mathbf{AM}}_{\text{Fus},mn}$$

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

346 2.3. Software

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

352 ([57], <https://mcrals.wordpress.com/>, last access 31/01/2019). The OPTICS algorithm was written
353 by Michal Daszykowski and it can be found at <http://chemometria.us.edu.pl/download/OPTICS.M>
354 (last access 31/01/2019). Kohonen's Self-Organizing Maps were computed by using a homemade
355 routine by Federico Marini (Università La Sapienza, Roma). The Fused Adjacency Matrix was
356 computed by using in-house written MATLAB routines, which will be made available for download
357 at <http://www.models.life.ku.dk/algorithms>.

358

359 **3. Results and discussion**

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

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

371

372 **3.1. Visible dataset**

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

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

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

391

392 **3.2. NIR dataset**

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

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

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

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

407

408 3.3. NMR dataset

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

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

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

430

431 3.4. Mid-level data fusion

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

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

448

449 3.5. Fused Adjacency Matrix

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

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

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

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

468 Based on our current knowledge, it is not possible to fully explain these groupings, however work is
469 in progress analysing a database of consumer preferences obtained from the website ratebeer.com¹
470 to assess if some of the grouping may be related to such information. Preliminary results show that
471 PC1 of the Fused Adjacency Matrix seems to have a strong inverse relationship ($R^2 = -0.973$) with
472 the overall score computed by the website from the users' evaluations (Fig.S1c).

473 ¹<https://www.ratebeer.com/> (last access 31/01/2019)

474

475 **3.6. Beer features comparison summary**

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

479

480 **3.6.1. Lagers group**

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

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

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

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

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

516

517 **3.6.2. Light samples set**

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

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

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

531

532 **3.6.3. ABV trend**

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

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

543

544 **3.6.4. Lagers Strong set**

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

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

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

562

563 **3.6.5. Colour trend**

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

568

569 **3.6.6. Summary Remarks**

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

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

583

Table 1 to be inserted about here

584

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

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

616 The effect of the different fusion steps was also assessed. These intra-block comparisons were made
617 for each data block individually (using the same number of PCs as specified above), and the results
618 are shown in Figure S4b. One interesting point is the transition from the distance information to its
619 correspondent \mathbf{AM}_X . The Euclidean distance \mathbf{D}_{Euc} resulted consistently similar to the Euclidean
620 \mathbf{AM}_{Euc} meaning that the “coded” AM version of the data is keeping a large part of the original
621 distance information. The same was observed with the Mahalanobis distance, albeit for the NMR
622 case the similarity between \mathbf{D}_{Mah} and \mathbf{AM}_{Mah} was found lower (Fig.S4b). By inspecting the
623 corresponding score plot it appears that this difference is due to a limited number of samples which
624 have extreme values on the second component in PCA of \mathbf{D}_{Mah} and are not in \mathbf{AM}_{Mah} (adjacency
625 being assigned on interval values is less sensitive to extreme values). Another interesting relation is
626 between the Euclidean and SOM AMs: the matrices \mathbf{AM}_{Euc} and \mathbf{AM}_{SOM} are very similar, either
627 because the samples pattern in the beer data can be well described by a linear model or because the
628 Euclidean distance (which is a non-linear transform) is sufficient to model the non-linearity present
629 in the data pattern. These two AMs also represent the two major contributions to the single-data
630 block \mathbf{AM}_X . The Mahalanobis distance was consistently found rather different from \mathbf{AM}_X and the
631 other distance measures. This is probably because higher PCs bring in rather different information
632 with respect to the first ones, as in order to avoid singularities we have calculated the Mahalanobis
633 distance on PCA-compressed data and thus it corresponds to Euclidean distances on the autoscaled
634 PCs. However, a systematic different behavior of the Mahalanobis distance with respect to other
635 metrics (including Euclidean) has been previously observed in a study considering several data sets
636 [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.

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

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

684

685 **4. Conclusions**

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

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

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

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

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

707

708 **Conflict of interest**

709 There is no conflict of interest.

710

711 **Acknowledgements**

712 Helena da Silva Friis is greatly acknowledged for contributing to sampling and part of the
713 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.

- 740 doi:10.1016/J.NEUNET.2012.09.018.
- 741 [12] K. Varmuza, P. Filzmoser, Introduction to multivariate statistical analysis in chemometrics,
742 CRC Press, 2009.
- 743 [13] I. Lee, J. Yang, Common Clustering Algorithms, in: Compr. Chemom., Elsevier, 2009: pp.
744 577–618. doi:10.1016/B978-044452701-1.00064-8.
- 745 [14] R. Gelbarda, O. Goldmanb, I. Spiegler, Investigating diversity of clustering methods: An
746 empirical comparison, *Data Knowl. Eng.* 63 (2007) 155–166.
747 doi:10.1016/J.DATAK.2007.01.002.
- 748 [15] L. Breiman, Random Forests, *Mach. Learn.* 45 (2001) 5–32. doi:10.1023/A:1010933404324.
- 749 [16] Yan-Yong Xu, Xian-Zhong Zhou, Zhong-Wei Guo, Weak learning algorithm for multi-label
750 multiclass text categorization, in: Proceedings. Int. Conf. Mach. Learn. Cybern., IEEE, 2002:
751 pp. 890–894. doi:10.1109/ICMLC.2002.1174511.
- 752 [17] P. Latinne, O. Debeir, C. Decaestecker, Combining Different Methods and Numbers of Weak
753 Decision Trees, *Pattern Anal. Appl.* 5 (2002) 201–209. doi:10.1007/s100440200018.
- 754 [18] Chuanyi Ji, Sheng Ma, Combinations of weak classifiers, *IEEE Trans. Neural Networks.* 8
755 (1997) 32–42. doi:10.1109/72.554189.
- 756 [19] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, *IEEE Trans. Pattern*
757 *Anal. Mach. Intell.* 20 (1998) 226–239. doi:10.1109/34.667881.
- 758 [20] L.I. Kuncheva, *Combining pattern classifiers - methods and algorithms*, 2014.
- 759 [21] A.-O. Boudraa, A. Bentabet, F. Salzenstein, Dempster-Shafer's Basic Probability
760 Assignment Based on Fuzzy Membership Functions, *ELCVIA Electron. Lett. Comput. Vis.*
761 *Image Anal.* 4 (2004) 1. doi:10.5565/rev/elcvia.68.
- 762 [22] B. Brownfield, T. Lemos, J.H. Kalivas, Consensus Classification Using Non-Optimized
763 Classifiers, *Anal. Chem.* 90 (2018) 4429–4437. doi:10.1021/acs.analchem.7b04399.
- 764 [23] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies
765 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)-

- 792 based methods, *Anal. Chim. Acta.* 674 (2010) 166–175. doi:10.1016/j.aca.2010.06.029.
- 793 [33] O. Oladokun, A. Tarrega, S. James, K. Smart, J. Hort, D. Cook, The impact of hop bitter acid
794 and polyphenol profiles on the perceived bitterness of beer, *Food Chem.* 205 (2016) 212–
795 220. doi:10.1016/j.foodchem.2016.03.023.
- 796 [34] C. Andrés-Iglesias, C.A. Blanco, J. Blanco, O. Montero, Mass spectrometry-based
797 metabolomics approach to determine differential metabolites between regular and non-
798 alcohol beers, *Food Chem.* 157 (2014) 205–212. doi:10.1016/j.foodchem.2014.01.123.
- 799 [35] S. Rossi, V. Sileoni, G. Perretti, O. Marconi, Characterization of the volatile profiles of beer
800 using headspace solid-phase microextraction and gas chromatography-mass spectrometry, *J.*
801 *Sci. Food Agric.* 94 (2014) 919–928. doi:10.1002/jsfa.6336.
- 802 [36] E. Bravi, O. Marconi, V. Sileoni, G. Perretti, Determination of free fatty acids in beer, *Food*
803 *Chem.* 215 (2017) 341–346. doi:10.1016/J.FOODCHEM.2016.07.153.
- 804 [37] S. Engelhard, H.-G. Löhmannsröben, F. Schael, Quantifying Ethanol Content of Beer Using
805 Interpretive Near-Infrared Spectroscopy, *Appl. Spectrosc.* 58 (2004) 1205–1209.
806 doi:10.1366/0003702042336000.
- 807 [38] O. Klein, A. Roth, F. Dornuf, O. Schöller, W. Mäntele, The Good Vibrations of Beer. The
808 Use of Infrared and UV/Vis Spectroscopy and Chemometry for the Quantitative Analysis of
809 Beverages, *Zeitschrift Für Naturforsch. B.* 67 (2012) 1005–1015. doi:10.5560/znb.2012-
810 0166.
- 811 [39] A. Biancolillo, R. Bucci, A.L. Magrì, A.D. Magrì, F. Marini, Data-fusion for multiplatform
812 characterization of an italian craft beer aimed at its authentication, *Anal. Chim. Acta.* 820
813 (2014) 23–31. doi:10.1016/j.aca.2014.02.024.
- 814 [40] M. Silvestri, A. Elia, D. Bertelli, E. Salvatore, C. Durante, M. Li Vigni, A. Marchetti, M.
815 Cocchi, A mid level data fusion strategy for the Varietal Classification of Lambrusco PDO
816 wines, *Chemom. Intell. Lab. Syst.* 137 (2014) 181–189.
817 doi:10.1016/j.chemolab.2014.06.012.

- 818 [41] D.-W. Sun, Å. Rinnan, L. Nørgaard, F. van den Berg, J. Thygesen, R. Bro, S.B. Engelsen,
819 Data Pre-processing, in: Da-Wen Sun (Ed.), *Infrared Spectrosc. Food Qual. Anal. Control*,
820 Elsevier, 2009: pp. 29–50. doi:10.1016/B978-0-12-374136-3.00002-X.
- 821 [42] L. Vera, L. Aceña, J. Guasch, R. Boqué, M. Mestres, O. Busto, Discrimination and sensory
822 description of beers through data fusion, *Talanta*. (2011). doi:10.1016/j.talanta.2011.09.052.
- 823 [43] I. Duarte, A. Barros, P.S. Belton, R. Righelato, M. Spraul, E. Humpfer, A.M. Gil, High-
824 Resolution Nuclear Magnetic Resonance Spectroscopy and Multivariate Analysis for the
825 Characterization of Beer, *J. Agric. Food Chem.* 50 (2002) 2475–2481.
826 doi:10.1021/jf011345j.
- 827 [44] A.M. Gil, I.F. Duarte, M. Godejohann, U. Braumann, M. Maraschin, M. Spraul,
828 Characterization of the aromatic composition of some liquid foods by nuclear magnetic
829 resonance spectrometry and liquid chromatography with nuclear magnetic resonance and
830 mass spectrometric detection, *Anal. Chim. Acta.* 488 (2003) 35–51. doi:10.1016/S0003-
831 2670(03)00579-8.
- 832 [45] D.W. Lachenmeier, W. Frank, E. Humpfer, H. Schäfer, S. Keller, M. Mörtter, M. Spraul,
833 Quality control of beer using high-resolution nuclear magnetic resonance spectroscopy and
834 multivariate analysis, *Eur. Food Res. Technol.* 220 (2005) 215–221. doi:10.1007/s00217-
835 004-1070-7.
- 836 [46] F. Savorani, G. Tomasi, S.B. Engelsen, icoshift: A versatile tool for the rapid alignment of
837 1D NMR spectra, *J. Magn. Reson.* 202 (2010) 190–202. doi:10.1016/j.jmr.2009.11.012.
- 838 [47] F. Savorani, G. Tomasi, S.B. Engelsen, Alignment of 1D NMR Data using the iCoshift Tool:
839 A Tutorial, in: J. van Duynhoven, P.S. Belton, Webb. G.A., H. van As (Eds.), *Magn. Reson.*
840 *Food Sci. Food Thought*, Royal Society of Chemistry, 2013: pp. 14–24.
841 doi:10.1039/9781849737531-00014.
- 842 [48] A. de Juan, R. Tauler, Multivariate Curve Resolution (MCR) from 2000: Progress in
843 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>.

- 870 [59] D. Intelmann, G. Haseleu, A. Dunkel, A. Lagemann, A. Stephan, T. Hofmann,
871 Comprehensive Sensomics Analysis of Hop-Derived Bitter Compounds during Storage of
872 Beer, *J. Agric. Food Chem.* 59 (2011) 1939–1953. doi:10.1021/jf104392y.
- 873 [60] I. Stanimirova, C. Boucon, B. Walczak, Relating gas chromatographic profiles to sensory
874 measurements describing the end products of the Maillard reaction, *Talanta*. 83 (2011) 1239–
875 1246. doi:10.1016/J.TALANTA.2010.09.018.
- 876 [61] J.M. Andrade, M.P. Gómez-Carracedo, W. Krzanowski, M. Kubista, Procrustes rotation in
877 analytical chemistry, a tutorial, *Chemom. Intell. Lab. Syst.* 72 (2004) 123–132.
878 doi:10.1016/J.CHEMOLAB.2004.01.007.
- 879 [62] P.D. Wentzell, S. Hou, C.S. Silva, C.C. Wicks, M.F. Pimentel, Procrustes rotation as a
880 diagnostic tool for projection pursuit analysis, *Anal. Chim. Acta.* 877 (2015) 51–63.
881 doi:10.1016/j.aca.2015.03.006.
- 882 [63] R. Todeschini, D. Ballabio, V. Consonni, Distances and Other Dissimilarity Measures in
883 Chemometrics, *Encycl. Anal. Chem. Appl. Theory Instrum.* (2015) 1–34.
884 doi:10.1002/9780470027318.a9438.

885

886 List of Tables

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

888

889 List of Figures

890 **Figure 1.** Graphical representation of how the Fused Adjacency Matrix \mathbf{AM}_{Fus} is obtained. In the
891 top box, the adjacency matrices are obtained from Euclidean and Mahalanobis distances, while in
892 the lower box they are obtained using SOM.

893

894 **Figure 2.** Visible spectra dataset: (a) Reachability Plot; (b) PC1 vs PC2 score plot, different
895 symbols refer to top (\blacktriangle) and bottom (\blacktriangledown) fermentation, while colours are by beer style, as detailed

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

899

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

903

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

910

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

914

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

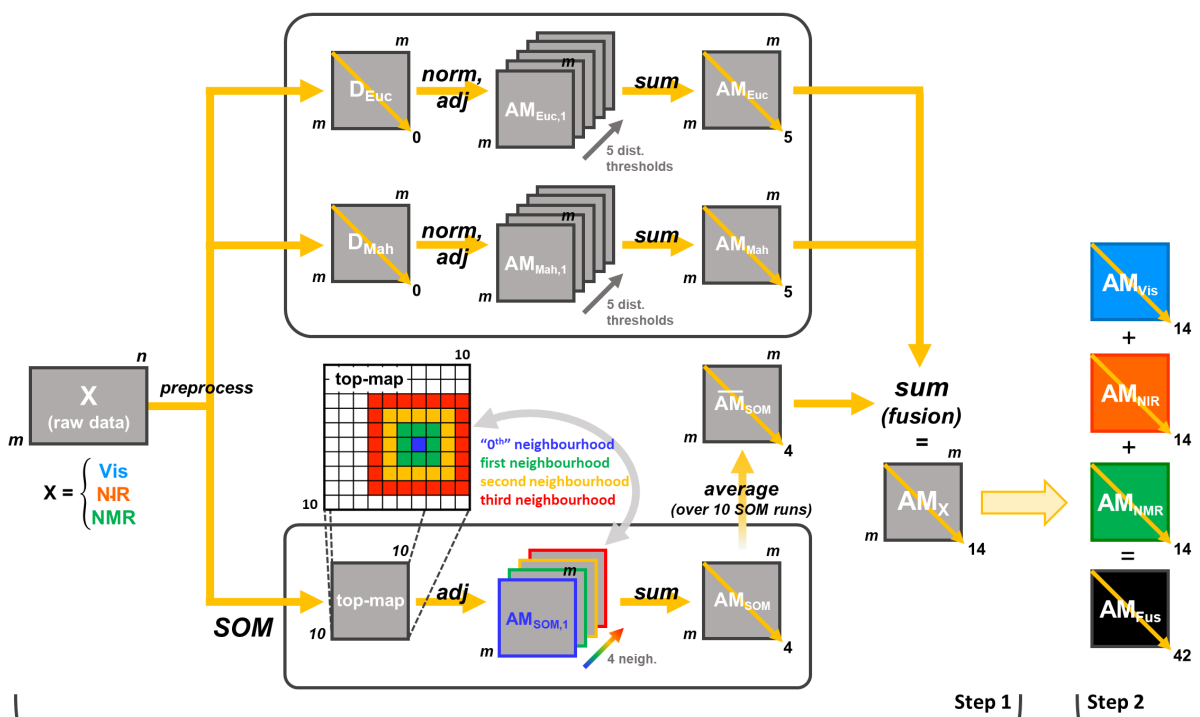
920

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

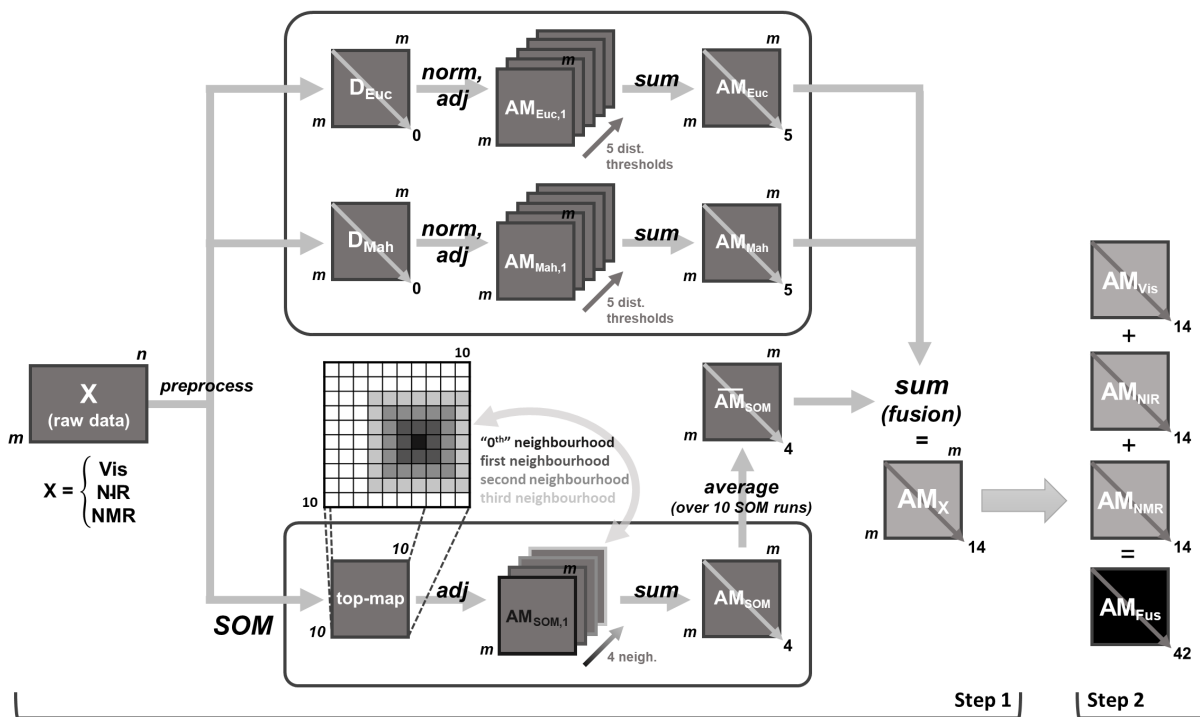
ACCEPTED MANUSCRIPT

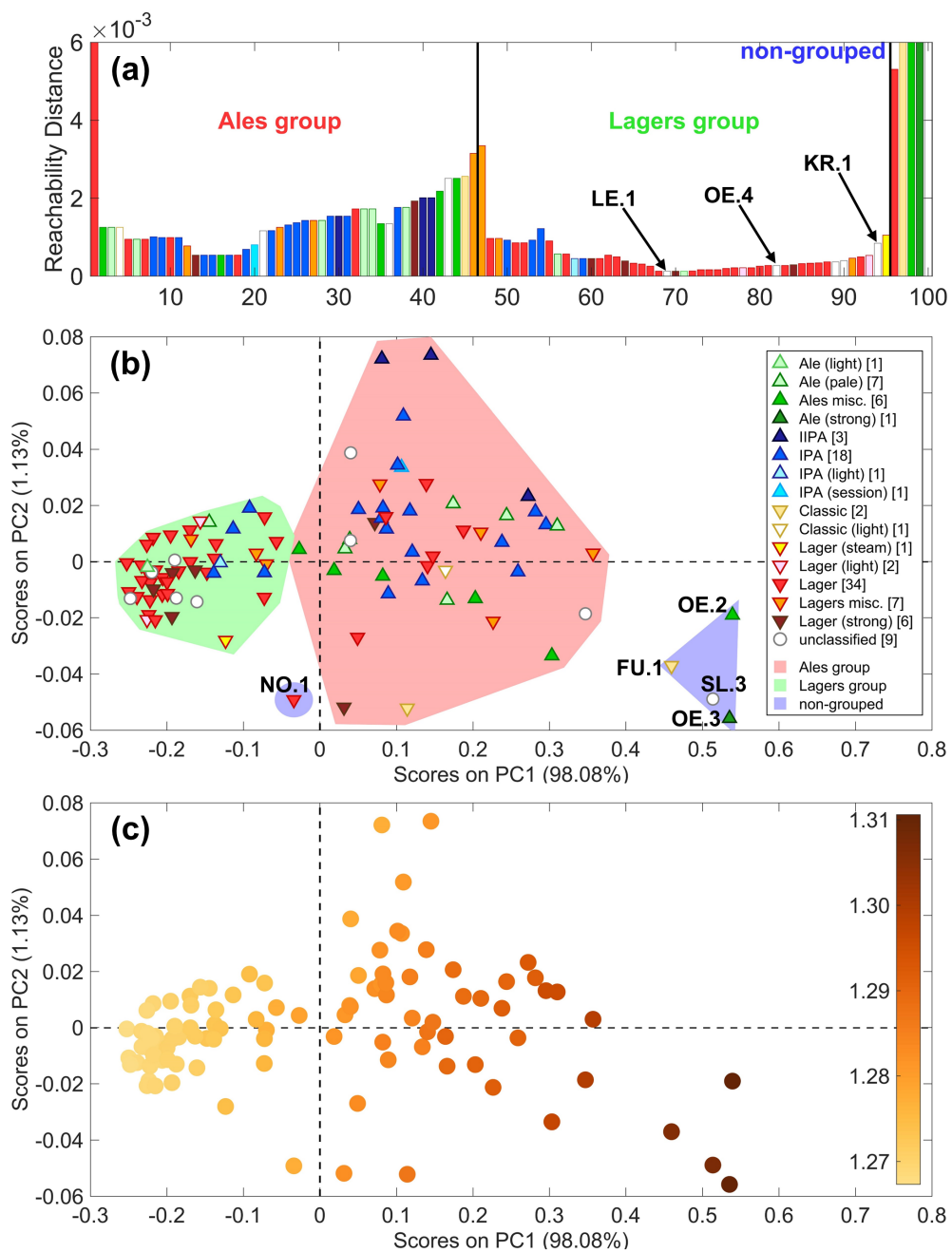
Table 1 Comparison summary (*ordered by increasing ABV)

| | Visible | NIR | NMR (Fig.4) | Mid-level data fusion | Fused Adjacency Matrix AM_{Fus} |
|---|---|--|---|---|---|
| Lagers group | Dense cluster in RP. (Fig.2a) Grouped in PCA. (negative scores, Fig.2b) | Slightly defined in RP. (Fig.3a) At positive PC1 scores, close to zero. (Fig.3b) | Slightly defined in RP. Medium to low variable values in general. Some sub-groups; contains the Light samples set as a sub-group. | Slightly defined in RP. (Fig.5a) At negative PC1 scores. (Fig.5b) | Defined cluster in RP. (Fig.6a) HI samples grouped and well-ordered together in RP. (Fig.6a) Grouped in PCA. (Fig.6b) |
| Unclassified ◦ fresh/summer beers in the Lagers group (most frequent ones: <u>LE.1</u> , <u>OE.4</u> , <u>KR.1</u>) | <u>LE.1</u> , <u>OE.4</u> , WI.2, SK.4, <u>KR.1</u> (Fig.2a) | <u>OE.4</u> , UG.3, <u>KR.1</u> , <u>LE.1</u> (Fig.3a) | <u>LE.1</u> , <u>OE.4</u> <u>KR.1</u> is in the non-grouped set. | <u>LE.1</u> , <u>OE.4</u> , <u>KR.1</u> , TY.3 (Fig.5a) | <u>LE.1</u> , <u>OE.4</u> , <u>KR.1</u> , WI.2 (Fig.6b) |
| Light samples set (KR.2, LE.2, FB.2, TO.4, NO.2)* | All in the Lagers group. (Fig.2b) Generally lighter colours. (Fig.2c) | Quite grouped in RP. (Fig.3a) All extreme on PC1. (Fig.3b) | Grouped in RP. Included in the Lagers group. Low values in general. | Not grouped in RP. (Fig.5a) Grouped in PCA. (Fig.5b) | Not grouped in RP. (Fig.6a) Grouped in PCA. (Fig.6b) |
| Lager Strong four low-ABV: MA.3, SI.9, MA.5, MA.6 two high-ABV: MA.2, FB.3 | Four low-ABV in the Lagers group, low-colour. (Fig.2a-b) 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) 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) Two high-ABV quite far in the non-grouped set. (Fig.5a) | Four low-ABV close to the Lagers group in PCA. (Fig.6b) 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) |

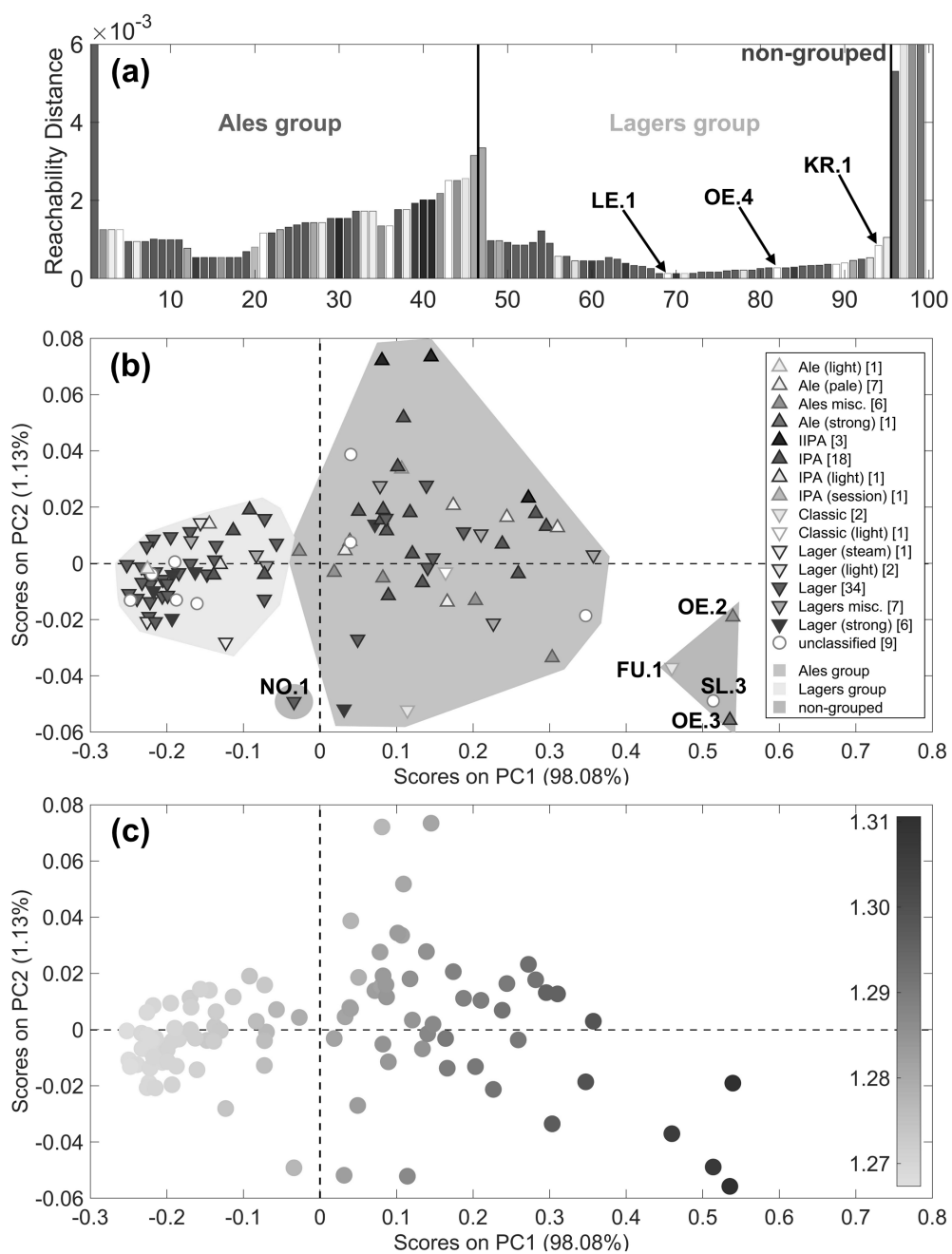


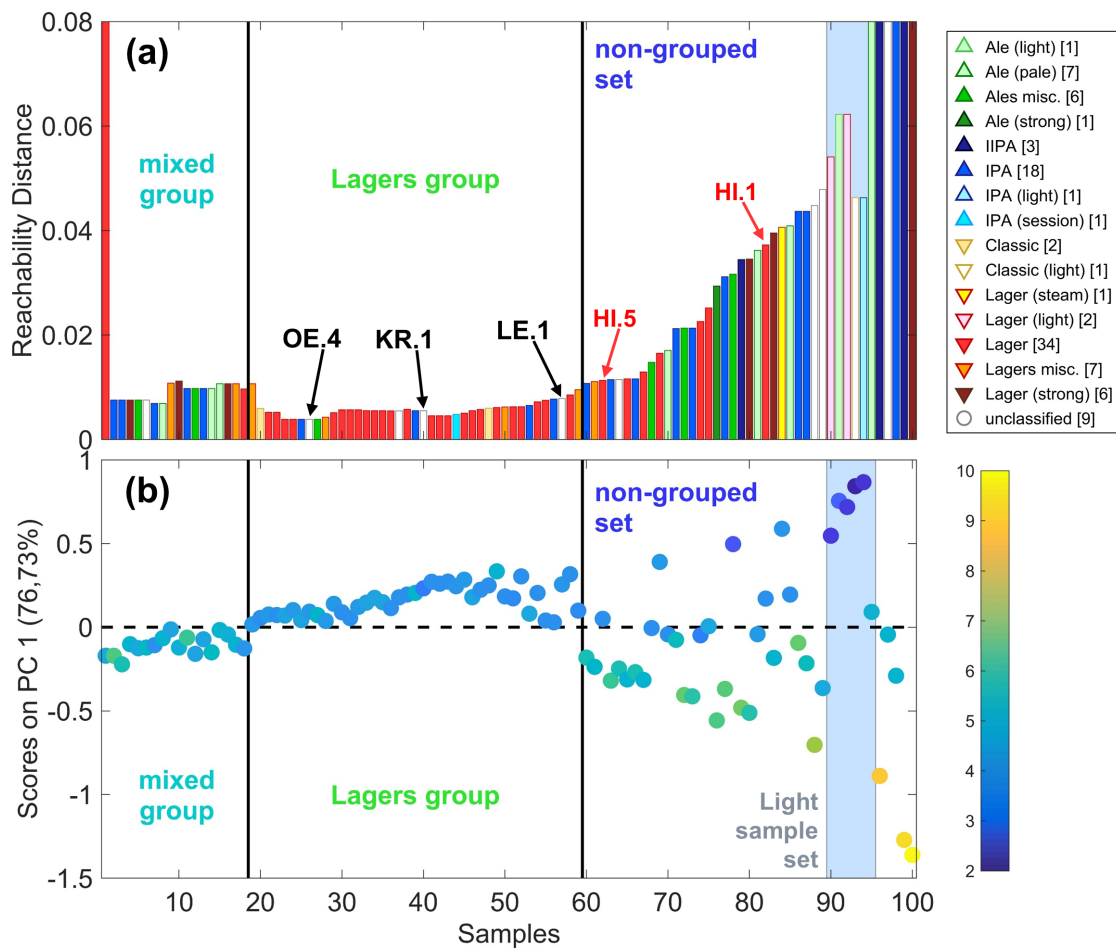
ACCEPTED MANUSCRIPT



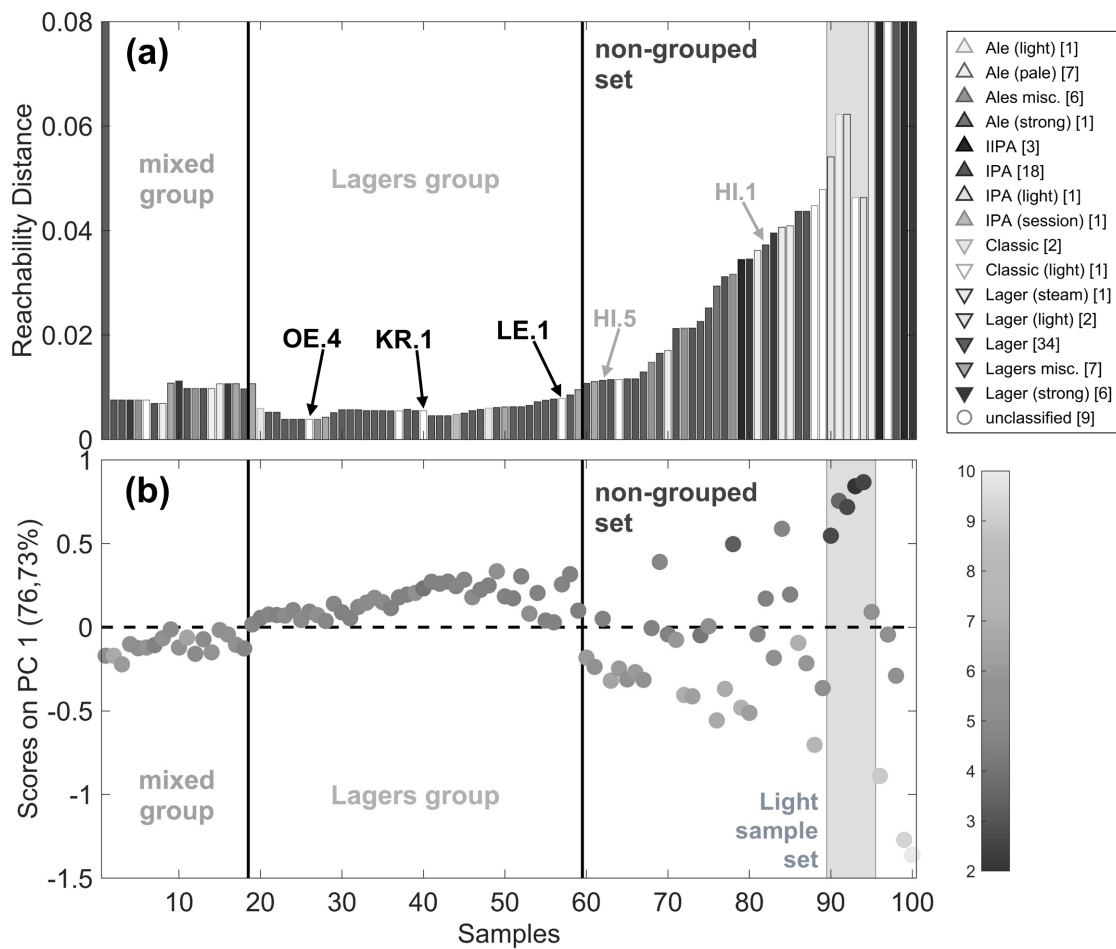


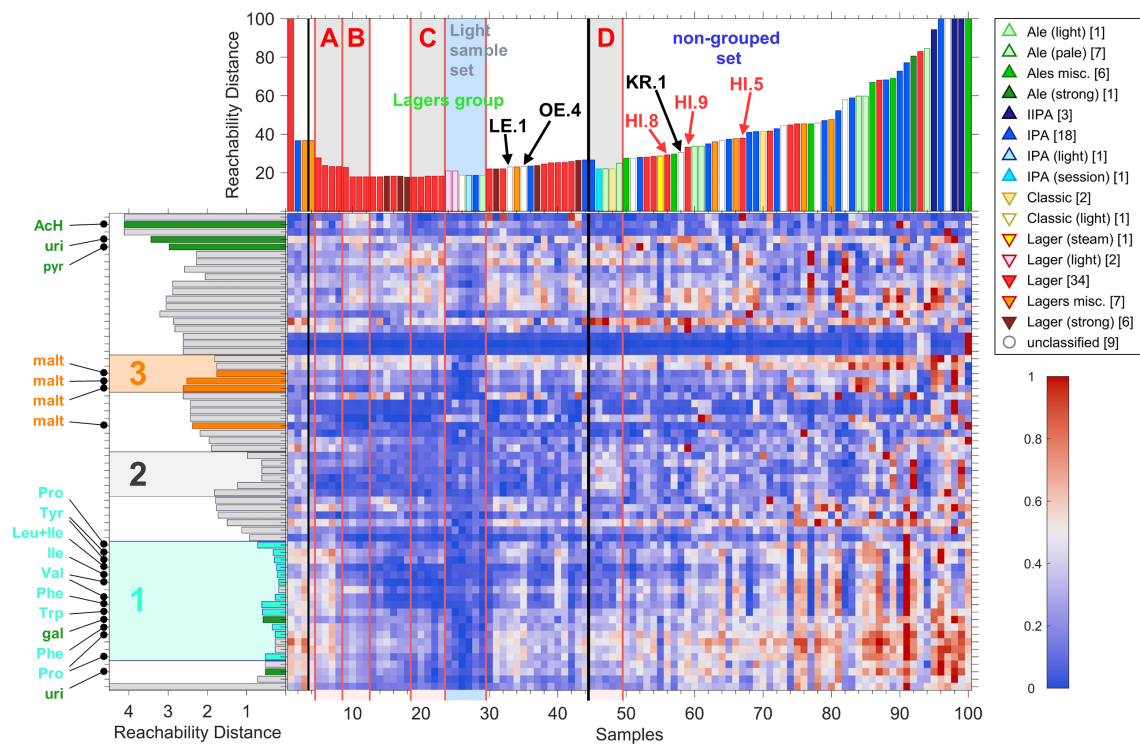
A



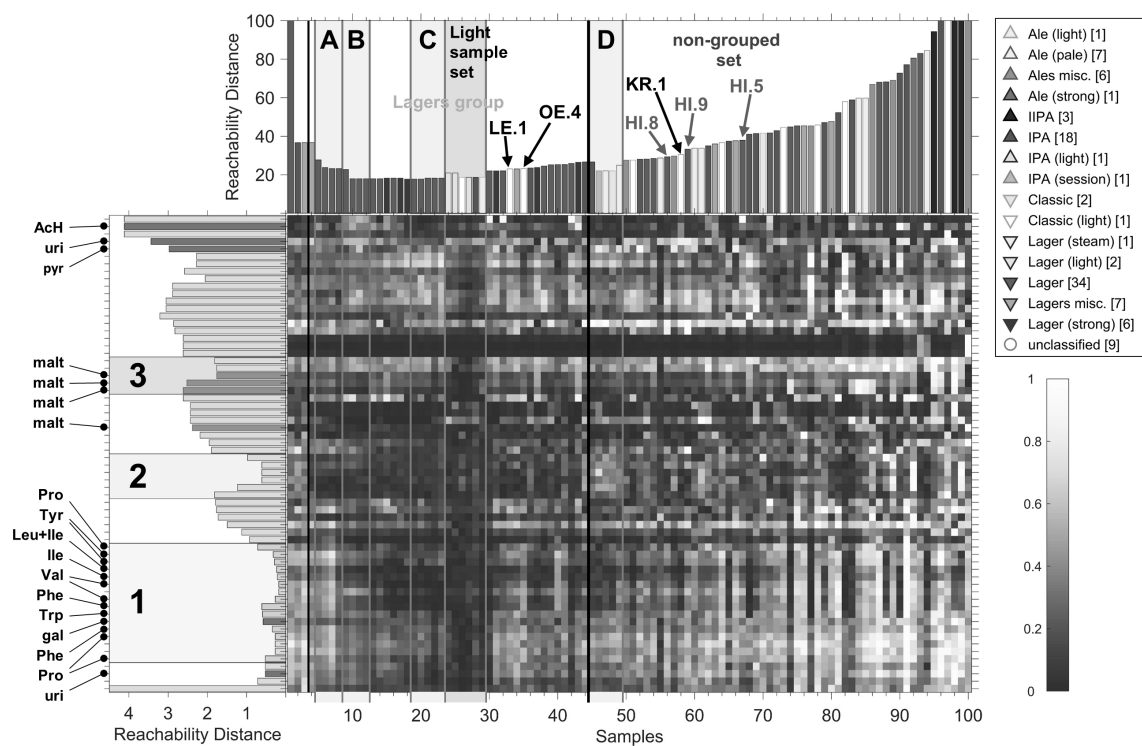


ACCEPTED

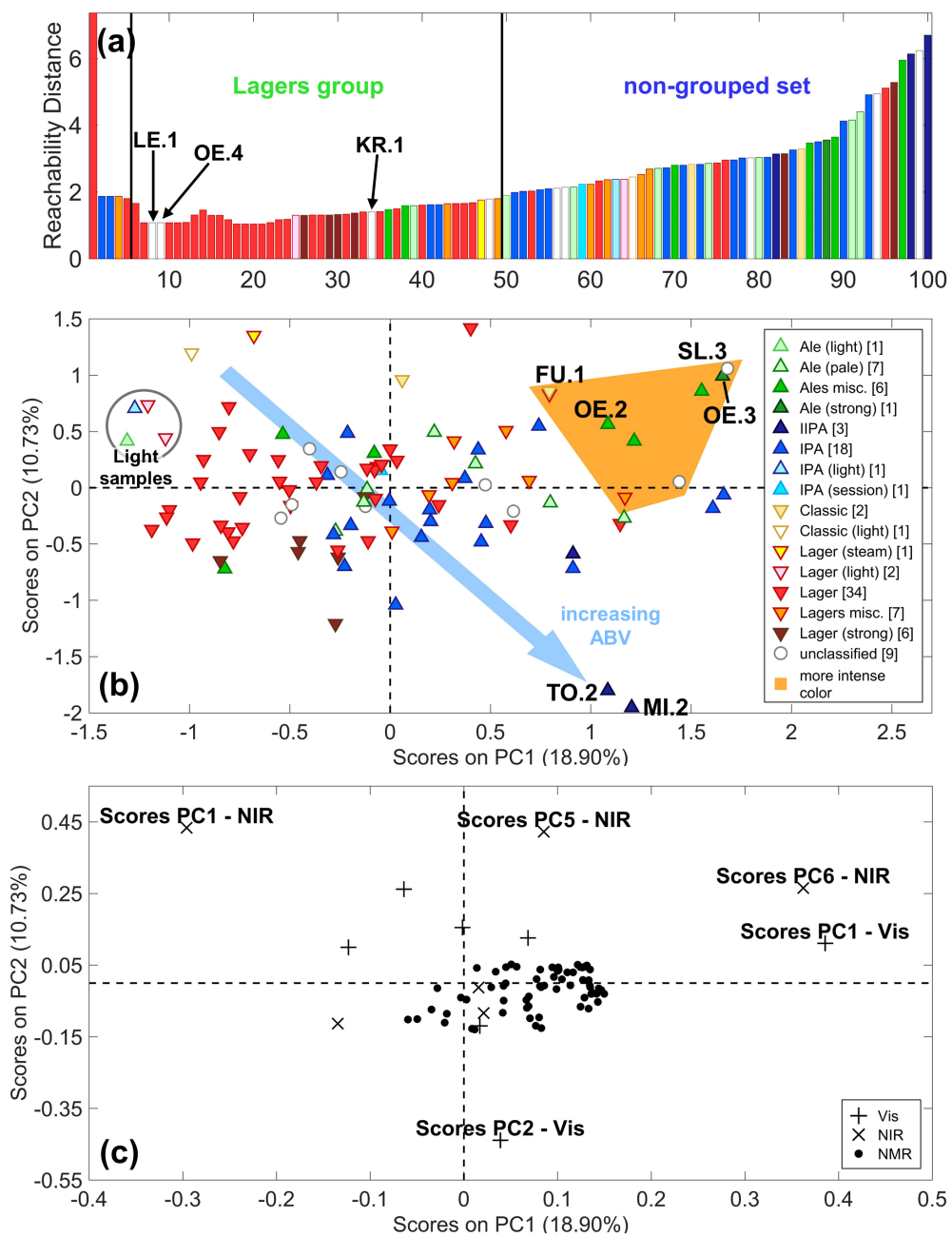


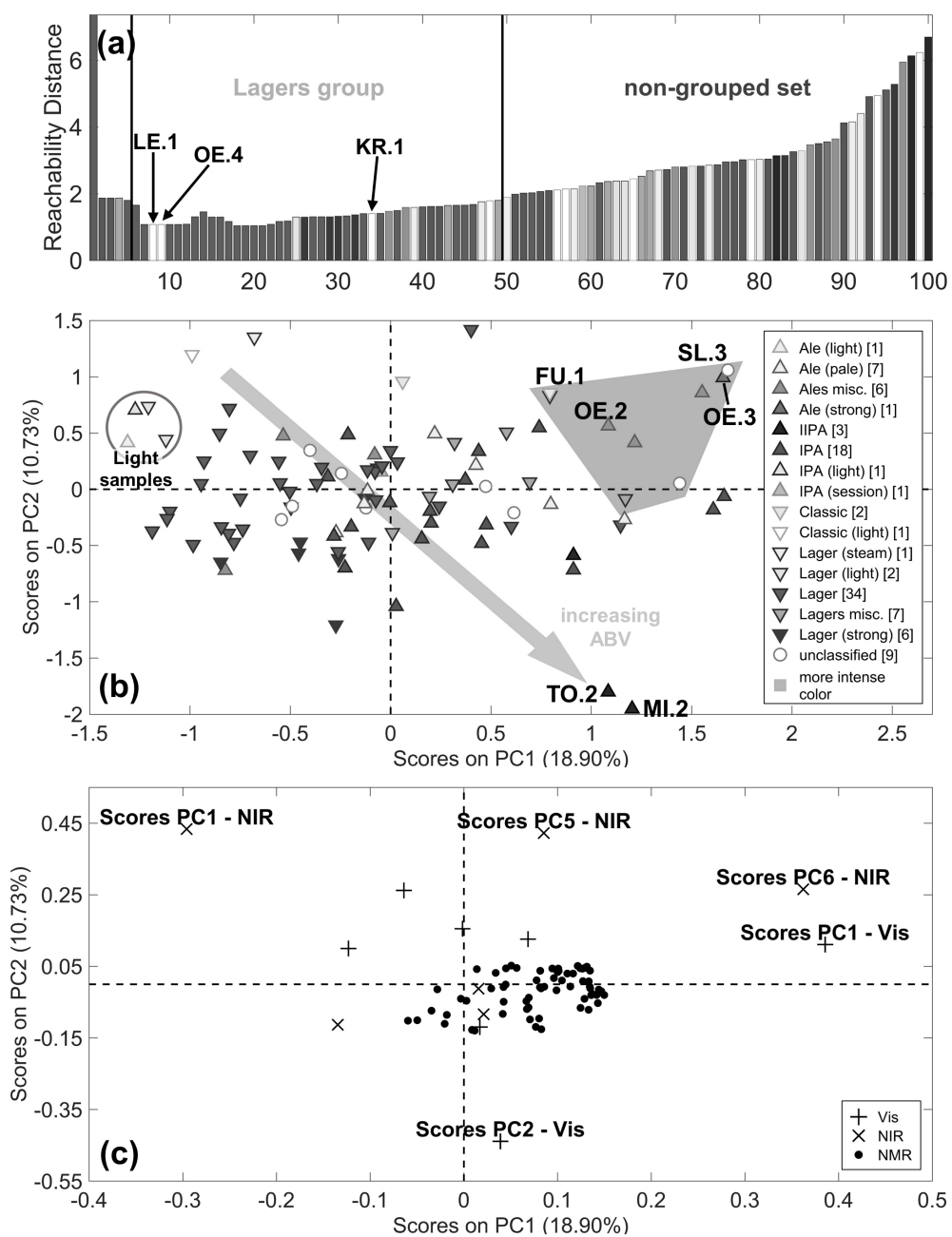


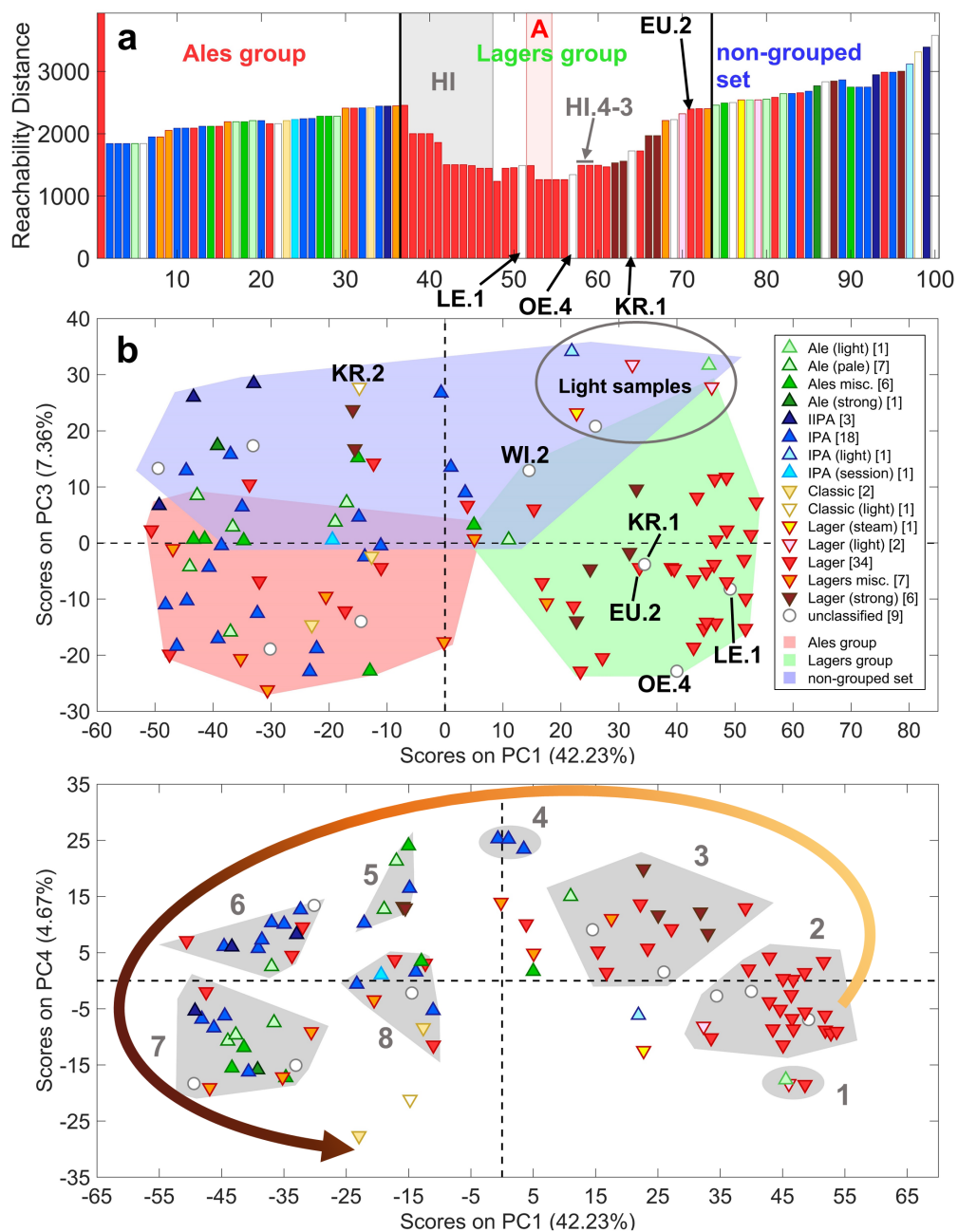
ACCEPTED



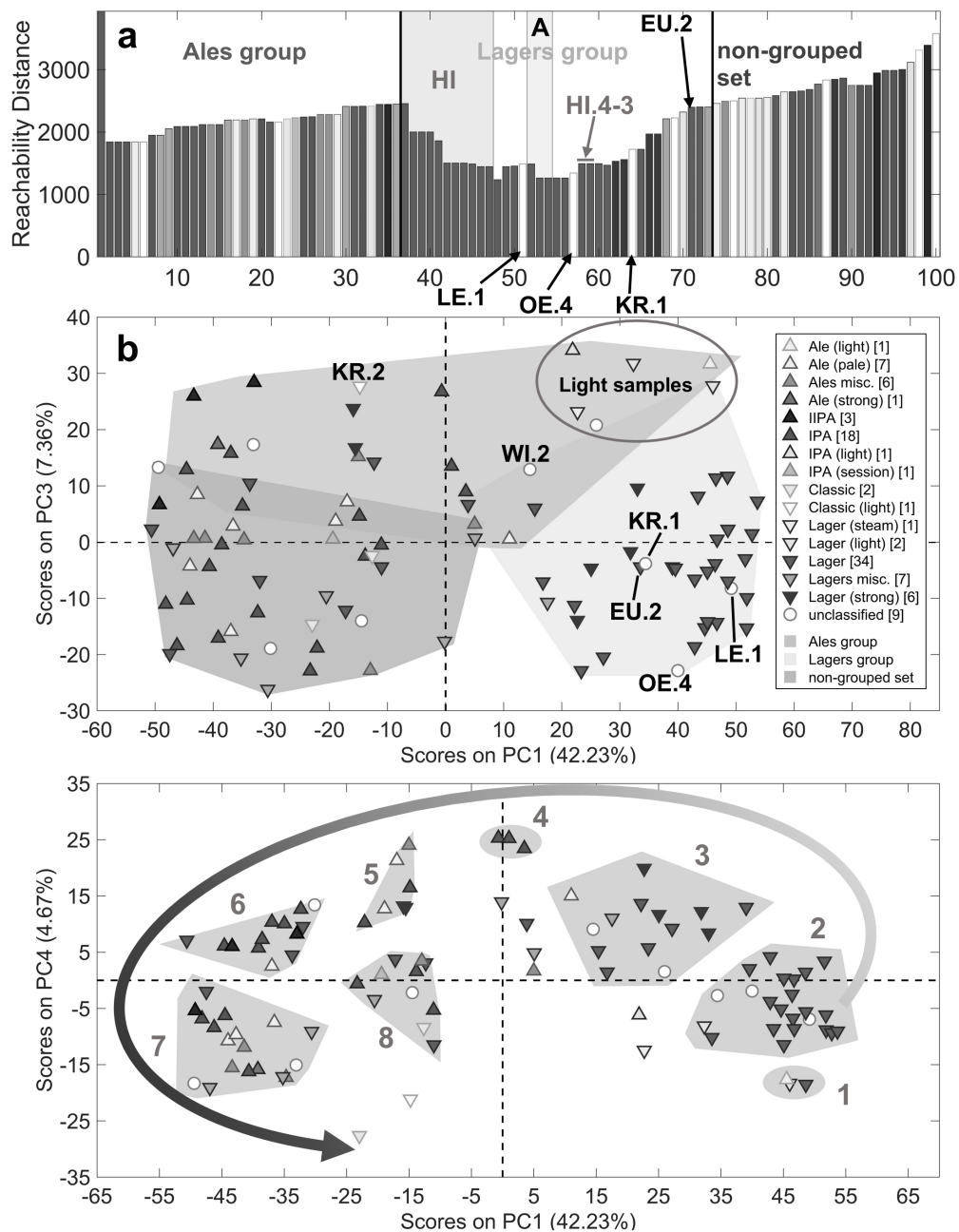
ACCEPTED



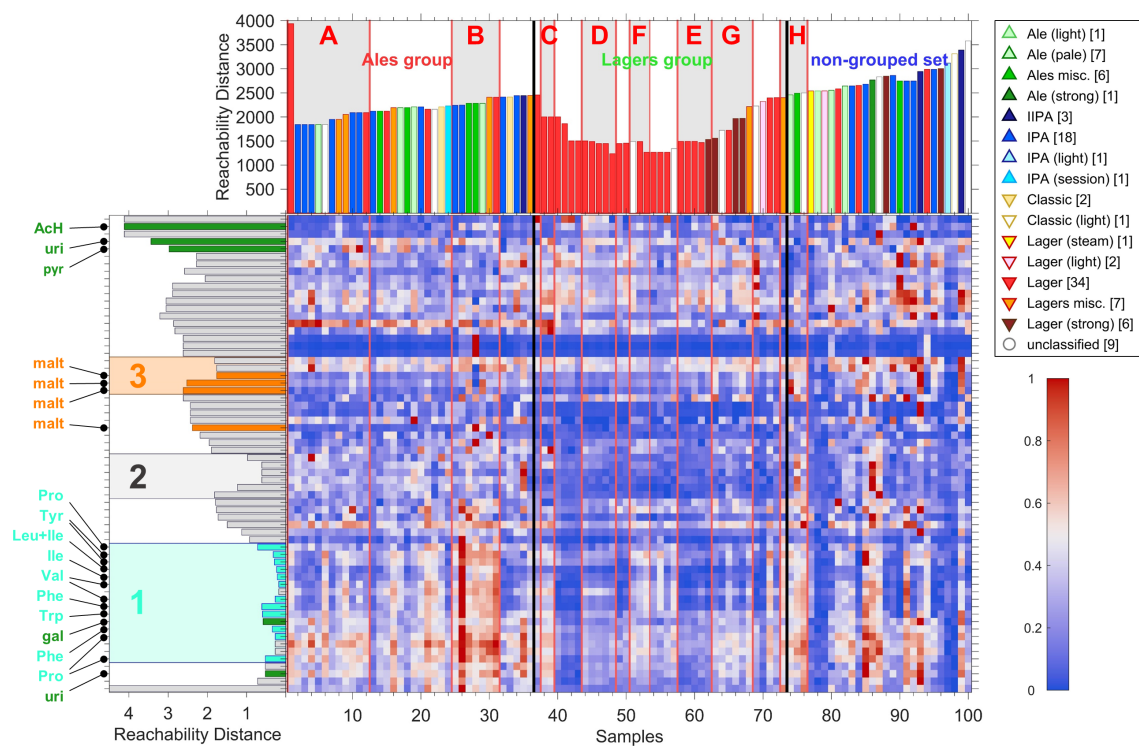


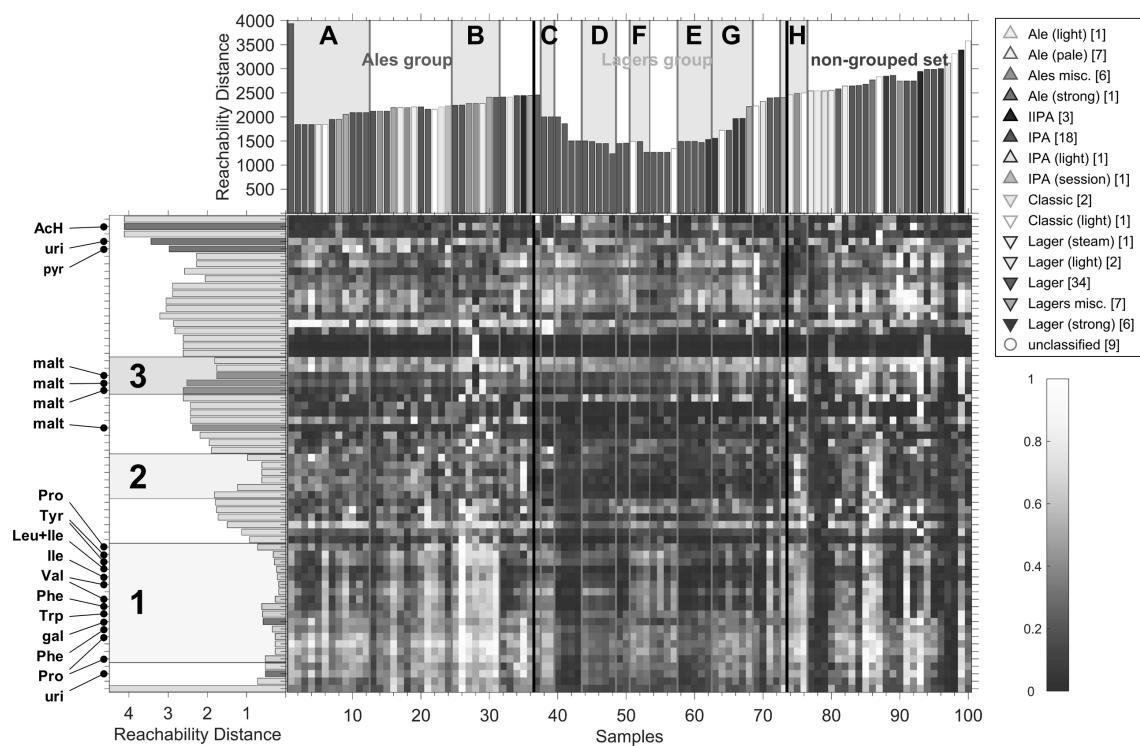


A



A





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: