Dairy Product Quality Using Screening of Aroma Compounds by Selected Ion Flow Tube–Mass Spectrometry: A Chemometric Approach

Jackie E. Wood[a]*, Brendon D. Gill[a], Wendy M. Longstaff[a], Robert A. Crawford[a], Harvey E. Indyk[a], Roger C. Kissling[a], Yi-Hsuan Lin[b], Carlos A. Bergonia[a], Lisa M. Davis[a], Anna Matuszek[b]

[a] Fonterra Waitoa, PO Box 7, Waitoa, New Zealand

[a] Wintec, Private Bag 3036, Waikato Mail Centre, Hamilton 3240, New Zealand

* Corresponding author

Abstract

Selected ion flow tube-mass spectrometry (SIFT-MS) can be used to analyse the concentration of volatile compounds in the headspace over food samples. Utilising chemometric classifiers, the concentration of aroma compounds detected by SIFT-MS was used to differentiate products. Nineteen compounds most useful in differentiating a range of dairy products were identified from the results of classification and selected for the development of preliminary threshold models to distinguish acceptable products from those containing off-aromas. Product differentiation was used to select the compounds for the threshold models, because sensory panel analysis rarely detects off-aromas in the products being examined. Threshold models for these compounds in the different products were developed using the 95% percentiles for the concentrations of these compounds that sensory panels found to be acceptable. These models have been used successfully during routine analysis to distinguish good products from marginal or off-aroma products, thereby lowering the demand on sensory panels.

1. Introduction

Chemometric analysis utilises mathematical and statistical models to associate analytical chemical data with the physicochemical properties of a product, which then allows the quality of unknown product batches to be predicted (Forina, Casale, & Oliveri, 2009), and is currently being used to determine

Wood, J.E.; Gill, B.D.; Longstaff, W.M.; Crawford, R.A.; Indyk, H.E.; Kissling, R.C.; Lin, Y.-H.; Bergonia, C.A.; Davis, L.M.; Matuszek. A. (2021) Dairy product quality using screening of aroma compounds by selected ion flow tube–mass spectrometry: A chemometric approach. *International Dairy Journal* 121, 105107. doi.org/10.1016/j.idairyj.2021.105107

authenticity, geographical origin, varietal differences and stage of maturation of different food products (Arvanitoyannis & Tzouros, 2005; Forina et al., 2009).

The majority of chemometric methods used in published dairy studies have included principal component analysis (PCA), discriminant functional analysis, and linear regression techniques including partial least squares discriminant analysis (PLS-DA) or linear discriminant analysis (LDA) (Karoui & De Baerdemaeker, 2007). Recently, data obtained using gas chromatography-mass spectrometry (GC-MS) was applied to the identification of lipid oxidation compounds and predict fishy off-flavours in dairy powders by random forest classifiers (Chen, Husny, & Rabe, 2018). Milk origin and quality were evaluated with an E-nose utilising data pre-processing using PCA followed by logistic regression and random forest classifiers (Chen, Husney, & Rabe, 2018; Mu, Gu, Zhang, & Zhang, 2020). Naïve Bayes classifiers have been used to distinguish fresh beef and pork based on E-nose headspace data (Wijaya, Sarno, & Daiva, 2017).

Threshold models using aroma compound screening for determining products with acceptable aromas has been a relatively novel approach in the dairy industry. There are a few prior investigations of off-aroma analysis for dairy product quality, performed predominantly using headspace GC-MS (Karoui & De Baerdemaeker, 2007). Early studies measure hexanal levels in fresh milk powder to evaluate lipid auto-oxidation (Ulberth & Roubicek, 1995) or microbial off-aroma compound thresholds as a guide for quality of cheddar cheese (Dunn & Lindsay, 1985). More recently, proton transfer reaction-mass spectrometry (PTR-MS) has been used alongside various multivariate analysis methods to examine the quality of butter and butter oil (van Ruth et al., 2008). A few studies with multivariate analysis have examined the quality of milk powders and whey powder (Makhoul et al., 2016), anhydrous milk fat (Makhoul et al., 2016; Pedrotti, 2018, 2020) and raw milk (Asaduzzaman, Kerschbaumer, & Scampicchio, 2018).

Dairy companies and other food manufacturers use human sensory panel analysis for routine quality control before sale of their products and for new product development (Baietto & Wilson, 2015; Carbonell-Barrachina, 2007; Drake, 2007; ISO/IDF 2009). However, there are some advantages to using instruments rather than panels, i.e., they do not suffer sensory fatigue or adaptation and the results are not affected by other physiological or psychological factors; therefore, the results should be highly reproducible (Smyth & Cozzolino, 2013). However, instrumental techniques may not be as sensitive to some aromas as trained panellists and may not detect all aromas accurately, because of the complex nature of aroma-chemical interactions within the sample (Chambers & Koppel, 2013; Smyth & Cozzolino, 2013).

Dairy proteins are large non-volatile molecules that can bind volatile aroma compounds in solution, causing them not to be released into the headspace (Kühn, Considine, & Singh, 2006). Commercial dairy protein products are typically not pure protein isolates and contain moisture, minerals, lipids, lactose, and feed taints, which can be aroma compounds themselves, or may lead to the formation of aroma compounds

(Lee, Laye, Kim, & Morr, 1996; Mehta, Bassette,& Ward,1974; Smith, Campbell, Jo, & Drake, 2016). A variety of different aroma compounds have previously been reported in different types of manufactured dairy proteins using GC-MS; these include aldehydes, ketones, furans, esters, lactones, sulphides, alcohols, amines, pyrroles, pyrazines and carboxylic acids (Carter & Drake, 2018; Evans, Zulewska, Newbold, Drake, & Barbano, 2009; Javidipour & Qian, 2008; Mortenson, Vickers, & Reineccius, 2008).

Pathways leading to the formation of aroma compounds in dairy products have been described, with two predominant mechanisms being lipid oxidation and thermal processing (Shipe et al., 1978). Aldehydes are the primary products of lipid oxidation and these can form in milk or whey stored at 4 C (Frankel, 1984). Secondary lipid oxidation products are alcohols, ketones, carboxylic acids, hydrocarbons, furans, lactones and esters (Frankel, 1984). In whey, aldehydes can be formed by the cheese starter (Whitson, Miracle, Bastian, & Drake, 2011). Many artefacts are formed as a result of thermal processing; pyrroles, pyrazines, lactones, and furans are formed in Malliard reactions (Calvo & de la Hoz, 1992; Ferretti & Flanagan, 1971), sulphur dioxide from methionine sulphur oxidation (Wüst & Pischetsrieder, 2016), ammonia and amines from the deamination of arginine and glutamine (Borad, Kumar, & Singh, 2017; Metwalli & van Boekel,1998; Sohn & Ho,1995), and thiols released from cysteine and cystine (Vazquez-Landaverde, Torres, & Qian, 2006; Volkin & Klibanov, 1987). Lecithin, an additive added to enhance solubility, can also release trimethylamine on heating (Lunden Gustafsson, Imhof, Gauch, & Bosset, 2002).

Other aroma compounds occur because strong feed taints can be transferred into the milk from either the digestive tract or lungs via the bloodstream of the cow, which then can diffuse into the bovine mammary gland where they are incorporated into the milk and manufactured dairy products (Babcock, 1938; Carter & Drake, 2018; Kilcawley, Falkner, Clarke, O'Sullivan & Kerry, 2018).

Selected ion flow tube-mass spectrometry (SIFT-MS) is a direct variant of mass spectrometry that can be used to provide highly sensitive, real-time analysis of volatile compounds present in the headspace of powder and liquid samples (Langford et al., 2012; Patana-anake & Barringer, 2016). The use of SIFT-MS to provide rapid screening of products is a novel approach for this type of study and the aim of this preliminary study was to utilise this technique to create aroma compound threshold models from a small number of compounds identified by chemometric classifier methods as important aroma discriminators for the milk protein, lactose and milk mineral products examined.

2. Materials and methods

2.1. Reagents

The SIFT-MS was calibrated using a Scott Mini-Max ultra-pure gas calibration cylinder (Aire Liquide, TX, USA). The cylinder contains 2.04 ppm benzene, 2.06 ppm ethylene, 2.08 ppm isobutane, 2.07 ppm octafluorotoluene, 2.05 ppm hexafluorobenzene, 2.14 ppm 1,2,4,5,-tetrafluorobenzene, 2.05 ppm toluene, and 2.05 ppm *p*-xylene, with the balance of the cylinder being nitrogen. A SIFT-MS standard solution was prepared using analytical grade water (18.2 MΩ cm^{-1}, Barnstead Genpure, ThermoFisher, Waltham, MA USA) comprising 10 ppb acetone (AR-grade, Fisher Chemical, Hampton, NH, USA), 1 ppb 3-methylbutanal (97%), 4 ppb acetaldehyde (≥ 99.5%), 0.4 ppb butyric acid (≥ 99%), 1 ppb 2,3-butanedione (analytical standard), 2 ppb hexanal (98%), 0.4 ppb dimethyl sulphide (≥ 99%) from Sigma-Aldrich (St. Louis, MO, USA), 0.4 ppb ethyl acetate (ChromAR grade), and 0.4 ppb glacial acetic acid (100% Empure) from Mallinckrodt (St. Louis, MO, USA) was prepared. This solution was diluted 1:100 (v/v) with analysis grade water, and was used to check the performance of the SIFT-MS instrument at beginning of each batch analysed.

2.2. Sample analysis

The aroma detection instrument used in this study was the Voice 200 SIFT-MS (Syft Technologies, Christchurch, NZ) fitted with a Gerstel Multipurpose auto-sampler comprising a six vial incubation chamber configured with a 2.5 mL gas-tight Gerstel headspace syringe (Mülheim an der Rurh, Germany).

A literature review of aroma compounds in dairy products was used to identify 81 compounds that could be detected by SIFT-MS and had previously been associated with sensory defects in dairy products. These compounds were entered into the SIFT-MS selected ion monitoring software, allowing the headspace concentration of ions associated with these compounds to be measured in the different samples. As a consequence of the way that products are manufactured, it is assumed that each product would have a unique aroma profile and therefore be able to be differentiated.

Instrument settings were optimised by initial trials on lactose, whey protein concentrates, and milk protein concentrates at three incubation times (5, 10 and 15 min), at three different temperatures (40, 55, and 70 °C) and the optimal settings were established to be incubation at 70 °C for 15 min. The concentrations of these 81 compounds were analysed in 300 samples, comprising of 23 mineralised and lactic acid casein, 17 calcium caseinate, 14 lactalbumin (highly heat-treated whey protein product), 34 milk protein concentrate, 56 sodium caseinate, 41 whey hydrolysate, 27 whey protein concentrate, 33 whey protein isolate,18 lactose, and 37 milk mineral powders. The samples came from manufacturing sites in the North

Island of New Zealand. All samples of each product had previously been assessed by the sensory panel to have a typical aroma profile. These samples were collected and analysed by SIFT-MS from June to December 2016.

Samples (3.0 ± 0.05 g) were weighed into 20 mL (Supleco, Bellefonte, PA, USA) screwcap headspace vials. Two blank samples were run at the beginning of each analysis to provide an atmospheric baseline, and two blank samples were run between different protein types to ensure that there was no carry over between the different types. Although the SIFT-MS can detect compounds down to 1 ppt, repeatability was variable at this concentration. For products where the concentration was above 1 ppb, the repeatability was ≤ 15.5% (n = 5 duplicates). Ammonia was initially selected as one of the 19 compounds, but due to poor repeatability it was omitted from the threshold models. The vials were loaded into the auto-sampler, which then transferred them sequentially to the incubation chamber, in which they were heated to 70 °C for 15 min without shaking. Several samples could be being incubated at one time. The headspace gas (2.5 mL) from these samples was injected directly into the instrument, which generated data for the 81 compounds.

Subsequently, the data from the 81 compounds were analysed using chemometric classifier methods to identify a smaller set of compounds, which were used for differentiating between types of protein products, lactose, and milk minerals. Initially, one model was developed for each product type using the threshold for each of the selected compounds. However, when more samples were analysed on a routine basis, it was found that samples containing lecithin gave rise to higher amounts of trimethylamine and one of the whey hydrolysates gave high levels of ammonia, so separate threshold models were developed for these products.

The final experimental settings were optimised to routinely analyse the selected compounds, allowing for at least 20 sample scan points per compound using 10,000 ion counts and a scan duration of 125 s. The syringe had an injection time of 15 s and a settle time of 110 s. The background was subtracted prior to the chemometric analysis but otherwise no pre-processing was carried out.

2.3. Statistical analysis

Three different classifier methods from the Weka 3 Machine Learning software (open-source data-mining software developed at Waikato University, Hamilton, New Zealand) were used to differentiate the protein products, lactose and milk minerals (Maimon & Rokach, 2010; Witten & Eibe, 2005). The classifier methods used were random forest, simple logistic and naïve Bayes. Differences between the methods were explained by Ng and Jordan (2001) and Menze, Geremia, Ayache, and Szekely (2012). Once the compounds had been selected, linear discriminant analysis (LDA) with crossvalidation in Minitab (State College, PA,

USA) was used to assess the probability of correct classification (Balakrishnama, Ganapathiraju, & Picone, 1999).

Because of low numbers of samples for some types of protein product, the accuracy of the models was assessed using ten-fold cross-validation, rather than using distinct training and validation test sets. The cross-validation test sets were chosen randomly using the Weka algorithm and used to assess the classifier methods. The classifier methods were trained using the whole data set and then the data was split into test sets, which were then iteratively classified by the trained model to determine the model's accuracy.

The threshold models for each product type were developed using data for the selected compounds in at least 60 samples of each product type that had previously been deemed to be of acceptable quality by the sensory panel (Esbensen &Wagner, 2017). The threshold value for each compound was set at the 95th percentile for each compound.

2.4. Sensory analysis

Sensory analysis methodology was performed by evaluating differences from a control sample using the difference from control method. The control was a sample with good sensory, functional and chemistry properties saved from a recent batch and stored in heat sealed, nitrogen gas flushed foil bags. Trained sensory panellists were used to access the aroma. The results for the sample were recorded as typical if the aroma was similar to the control, or atypical if the aroma was different.

For lactose samples, aroma was measured on liquid samples made by dissolving 100 g lactose in 300 g deionised water in a 600 mL glass beaker. For all other products, the control bag aroma was sampled immediately before the sample bag.

2.5. Method application

The SIFT-MS has been routinely used to screen protein, milk minerals and lactose products since 2017 using the threshold models. In the first year after the development of the thresholding models, approximately 1900 samples were screened, and in the second year, 3400 samples were screened.

3. Results

3.1. Initial studies

Initial optimisation trials on a sodium caseinate, milk protein concentrate, whey protein concentrate, and lactose samples demonstrated that, due to the nature of the auto-sampler set-up, the 15 min incubation time proved to be optimal.

Unsurprisingly, the compounds were at their most concentrated in the vial incubated at 70 °C. An absence of change in the concentration of Maillard reaction products at the three different temperatures, confirmed that analysis conditions of 15 min at 70 °C did not cause the appearance of heat marker compounds. Longer incubation times did not appear to increase the concentration of the compounds in the headspace, suggesting that the headspace had reached equilibrium. Replicates of the SIFT-MS standard had a RSD of 9.9–12.3% (n = 29 days).

3.2. Classification modelling

Classifier methods were used to determine which aroma compounds were most important in differentiating the products. When all products were classified together, neither the naïve Bayes method nor the random forest method was entirely successful in classifying all samples. The naïve Bayes method correctly classified 264 out of the 300 samples, corresponding to 88% correct classifications, while the random forest model, with only 17 samples being incorrectly classified, yielded over 94% correct classifications. Confusion matrices for both these model types, showing how the protein products, milk minerals and lactose were classified, are given in Supplementary material Tables S1 and S2. LDA could not be used to carry out the initial classification using all 81 compounds because of the high co-linearity between some of the compounds; however, it was used to assess the 19 compounds comprised of ammonia, short chain tertiary amines, sulphides, mercaptans, ketones, straight and branched chain aldehydes, and a methoxylated phenol, once they had been selected.

Because the products were not classified entirely correctly initially, they were broken down into pairwise subsets and reclassified using both these same two methods and also the simple logistic method. In summary, 135 pairwise classifications were prepared; these examined 45 product pairs using the three methods, as shown in Supplementary material Tables S3–S5. Fig. 1 shows the numbers of product pairs, out of the 45 pairs, that were classified by the different models.

The 19 compounds selected were those that: featured most frequently in the visual random forest method output; featured most frequently in the simple logistic equation method output; or had naïve Bayes method means that were more than one standard deviation apart. These compounds most frequently occurred for the 103 models that either correctly classified the protein products or misclassified only one are displayed in Fig. 2. Despite Fig. 2 showing that 13 compounds may be sufficient to explain most of the variance, the results from LDA found that, when 13 compounds were used, 77% of the products were correctly classified, whereas, when 19 compounds were used, 84% of the products were correctly classified. By examining the PCA and the loadings plot (Fig. 3 and Table 1) for all products examined using the selected compounds as predictor variables, 12 factors explained 96% of the variation which contained contributions

from all 19 compounds. The confusion matrices using LDA for all products and the pairwise product comparisons are shown in Supplementary material Tables S6 and S7, respectively.

In some cases, the compounds selected for use in the model were highly correlated because aroma compounds in this study were predominantly formed by lipid oxidation or as thermal artefacts. Because these compounds were formed via a limited number of pathways, they were often members in the same homologous series increasing by one methylene group. The number of Pearson correlations between the compounds for the 57 sodium caseinate samples is shown in Fig. 4.

3.3. Method application

In the first year (the New Zealand dairy season for June 2017–May 2018), all product samples were sent to both the SIFT-MS and the sensory panel. The SIFT-MS limits were set more tightly than the sensory panel limits, such that all products that failed the sensory panel also failed when analysed on the SIFT-MS. During the second year (June 2018–May 2019), one in every ten product samples, and those which were found to have off-aroma when using SIFT-MS, were subsequently assessed by the sensory panel. Supplementary Figs. 1–4 are PCA diagrams for the different sample types with approximately 100 samples of each product type analysed. The caseins in Supplementary material Fig. S1, including acid casein, calcium caseinate, and sodium caseinate, can be differentiated based on the geographical location of manufacture. When any product sample was an outlier and failed two or more compound thresholds of the models, it was sent to the sensory panel for further testing.

The PCA diagram generated for whey protein products (Supplementary material Fig. S2) is based on data for whey powders, whey protein concentrates, whey protein isolates and two distinct whey protein hydrolysates. One of the hydrolysates gave rise to significantly higher concentrations of headspace ammonia.

Both SIFT-MS and the sensory panel found six instances where the product had sensory defects. There was only one instance of SIFT-MS finding a product to be acceptable and the sensory panel finding it unacceptable; however, the compounds responsible for causing the off-aromas were semi-volatile indoles, which are not detectable by SIFT-MS.

4. Discussion

Although heated milk proteins and lactose can give rise to Maillard reaction products, samples need to be heated to above 70 °C before any Maillard products develop (Ajandouz, Desseaux, Tazi, & Puigserver, 2008; Cadwallader & Singh, 2009; Lee et al., 1996; Nie et al., 2013; Shimamura & Ukeda, 2012). The

caramelisation of any residual sugars in the protein products will usually occur at temperatures > 100 °C (Ajandouz et al., 2008).

As freshly manufactured product was predominantly used in the initial study, the data set was not well balanced (Cadwallader & Singh, 2009; Weiss, 2010) and as the sample size was small, cross-validation was used instead of validating with a test set. To be effective, all samples in the training set must be used in the cross-validation (Hawkins, Basak, & Mills, 2003). The main disadvantage of cross-validation is that it tends to overrate the effectiveness of the classifier methods (Baron, 2016; Hawkins et al., 2003).

The naïve Bayes model was selected due to its reputation for handling small data sets (Gertz, Gertz, Matthäus, & Willenberg, 2019), with random forest chosen as it can mitigate the known problems of a small data set, such as being unbalanced and co-linearity (Brown & Mues, 2012; Tomaschek, Hendrix, & Baayen, 2018). Co-linearity was expected to be present since the selected aroma compounds are often a homologous series of small organic compounds produced by the same mechanism, e.g., low molecular mass aldehydes and ketones.

Simple logistic, which is a logistics regression model used to classify pairs, can deal with multi co-linearity, and although not as good as random forest, such logistic regression classifiers perform better than other classifiers with unbalanced data sets (Brown & Mues, 2012; Lieberman & Morris, 2014).

When the results from the three classifier methods were combined, the compounds identified were largely successful in identifying product batches containing off-aromas. There were, however, two complications with the data set used to select the compounds used in the classifier methods.

The first complication was misclassification due to the rare sample effect; that is, products with smaller numbers tended to be misclassified, especially when using the naïve Bayes and logistic regression classifiers (McKnight, Wilcox, & Hripcsak, 2002; King & Zeng, 2001; Weiss, 2010). Random forest classifiers are more successful in mitigating the rare sample effect with small sample sizes (Mi, Huettmann, Guo, Han, & Wen, 2017). Rare sample effects can be overcome by prior correction or by weighting the log likelihood ratio depending on the samples and the proportion in the population. Prior correction is preferable in cases of small sample size (King & Zeng, 2001). The LDA classifier coped well with samples with few representatives, despite the fact that under-sampling has a negative effect on performance, as shown in Supplementary material Tables S7 and S8 (Xie & Qui, 2007).

The second complication occurred because the compounds used in the models were highly correlated with each other, causing multiple co-linearity. One-way of coping with highly correlated data is to pre-process the data using PCA and use the resultant PCA eigenvector outputs as the predictor variables; however, in doing this, it would be difficult to ascertain which compounds to measure (Linting, van Os, & Meulman, 2011; Zhang & Sun, 2009). Although one of the rules of the naïve Bayes classifier is that the predictor

variables need to be independent, this classifier often produces more accurate output than would be expected with data for which the features are not independent (Rish, 2001; Rish, Hellerstein, & Thathachar, 2001; Scott et al., 2013). The simple logistic method is also adversely affected by highly correlated features, and this can produce large mean square errors (Barker & Brown, 2001). The random forest classifier method is considered to be better at coping with highly correlated predictor variables; however, it has been shown that highly correlated important variables may be selected more often for the ensemble of decision trees than those variables that are important but not highly correlated (Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008). The LDA method could not cope with the co-linearity of the data. Due to this study generally being carried out to establish which compounds to use in the threshold models, co-linearity was accepted and was not corrected for as the compounds cannot be obtained from PCA factors.

The optimal headspace measurement with the SIFT-MS meant that a maximum of 19 compounds was optimal within the dwell time for the SIFT-MS. To avoid the problem of model overfitting through using too many predictor variables, the histogram of the prevalence of the 19 compounds was examined, and this showed that there was a decline after the 13th compound, suggesting that these may be enough variables to use. However, LDA analysis using 13 compounds gave only 77% correct classifications, which improved to 84% when all 19 compounds were used (Babyak, 2004).

Future research that could be considered is to use ensemble models rather than combining the results from the three separate models. Ensemble models may have resulted in better predictions due to less noise, making the models more robust; however, this increases the complexity of the model, making some results more difficult to interpret (Ramzai, 2019).

Because of the high similarity of whey protein isolate (~90% whey protein) and whey protein concentrate (~80% whey protein), these protein products are difficult to differentiate from each other. Protein samples from dairy processing plants have been analysed as a screening tool in routine product release by the SIFT-MS technique for approximately 2 years since the selected compound threshold models were developed. If instrumental screening does not identify any atypical compounds in a sample taken from a production batch, then no further sensory evaluation of that batch needs to be performed; however, if the screening tool identifies an atypical result, then further evaluation of that sample is undertaken by a trained sensory panel.

As with the caseins, milk protein concentrates and lactalbumin can be differentiated on the basis of the location of their manufacture (Supplementary material Fig. S5). Different milk protein concentrates can be identified by the amount of fat or lactose contained within the protein product. It was important to consider the spread of results when creating the models, as it was found that different factories and geographic locations gave different compound concentrations for a typical product.

5. Conclusions

This method proved to be an effective means of selecting the marker compounds that were used to form the threshold models for SIFT-MS. It was also shown that different types of classifier method are better at coping with naturally problematic data sets and that indicative results from several classifier or ensemble methods may be preferable to using a single method. The SIFT-MS has been successfully used to routinely screen protein products, lactose, and milk minerals for off-aromas since 2017 and offers an instrumental technique complementary to traditional sensory panel analysis.

CRediT author statement

Jackie E. Wood: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Supervision; Validation; Writing original draft; Writing, review & editing. Wendy Longstaff: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Resources; Validation. Robert A. Crawford: Conceptualization; Methodology; Writing, review & editing. Brendon D. Gill: Visualization; Writing, review & editing. Yi-Hsuan Lin: Formal analysis. Carlos Bergonia: Conceptualization; Project administration; Resources; Supervision. Roger C. Kissling: Conceptualization, Methodology; Writing, review & editing. Lisa M. Davis: Project Administration, Investigation. Anna Matuszek: Supervision, Project Administration. Harvey E. Indyk: Writing, review & editing.

Declaration of competing interest

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.

Acknowledgements

The authors would like to thank Cherilyn Bunning and Ces Tooke for project administration support as well as Claire Woodhall and Steve Holroyd for editing this manuscript. This project could not have been carried out without the assistance of the technical staff and sensory panel staff at the Fonterra Waitoa site. This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.idairyj.2021.105107.

References

Ajandouz, E. H., Desseaux, V., Tazi, S., & Puigserver, A. (2008). Effects of temperature and pH on the kinetics of caramelisation, protein cross-linking and Maillard reactions in aqueous model systems. *Food Chemistry*, 107, 1244–1252.

Arvanitoyannis, I. S., & Tzouros, N. E. (2005). Implementation of quality control methods in conjunction with chemometrics toward authentication of dairy products. *Critical Reviews in Food Science and Nutrition*, 45, 231–249.

Asaduzzaman, M., Kerschbaumer, M., & Scampicchio, M. (2018). Rapid and non-invasive multivariate approach for the quality control of raw milk from mountain areas based on proton transfer reaction mass spectrometry data. *Rapid Communications in Mass Spectrometry*, 32, 1379–1386.

Babcock, C. J. (1938). Feed flavours in milk and milk products. *Journal of Dairy Science*, 21, 661–668.

Babyak, M. A. (2004). What you see may not be what you get: A brief, non-technical introduction to overfitting in regression-type models. *Psychosometric Medicine*, 66, 411–421.

Baietto, M., & Wilson, A. D. (2015). Electronic-nose applications for fruit identification, ripeness and quality grading. *Sensors*, 15, 899–931.

Balakrishnama, S., Ganapathiraju, A., & Picone, J. (1999). Linear discriminant analysis for signal processing problems. *Proceedings of the IEEE Southeast Con*, 1999, 36–39.

Barker, L., & Brown, C. (2001). Logistic regression when binary predictor variables are highly correlated. *Statistics in Medicine*, 20, 1431–1442.

Baron, G. (2016). Comparison of cross-validation and test sets approaches to evaluation of classifiers in authorship attribution domain. In T. Czachórski, E. Gelenbe, K. Grochla, & R. Lent (Eds.), *Computer and information sciences. ISCIS 2016. Communications in computer and information science* (Vol. 659, pp. 81–89). Cham, Switzerland: Springer.

Borad, S. G., Kumar, A., & Singh, A. K. (2017). Effect of processing on nutritive values of milk protein. *Critical Reviews in Food Science and Nutrition*, 57, 3690–3702.

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39, 3446–3453.

Cadwallader, K. R., & Singh, T. K. (2009). Flavours and off-flavours in milk and dairy products. In P. L. H. McSweeney, & P. F. Fox (Eds.), *Lactose, water, salts and minor constituents* (Vol. 3, pp. 631–690). New York, NY, USA: Springer. Advanced dairy chemistry.

Calvo, M. M., & de la Hoz, L. (1992). Flavour of heated milks. A review. *International Dairy Journal*, 2, 69–81.

Carbonell-Barrachina, A. A. (2007). Application of sensory evaluation of food to quality control in the Spanish food industry. *Polish Journal of Food and Nutrition Sciences*, 57, 71–76.

Carter, & Drake, M. A. (2018). The effects of processing parameters on the flavour of whey protein ingredients. *Journal of Dairy Science*, 101, 6691–6702.

Chambers, E., & Koppel, K. (2013). Associations of volatile compounds with sensory aroma and flavour: The complex nature of flavour. *Molecules*, 18, 4887–4905.

Chen, C., Husney, J., & Rabe, S. (2018). Predicting fishiness off-flavour and identifying compounds of lipid oxidation in dairy powders by SPME-GC/MS and machine learning. *International Dairy Journal*, 77, 19–28.

Drake, M. A. (2007). Sensory analysis of dairy foods. *Journal of Dairy Science*, 90, 4925–4937.

Dunn, H. C., & Lindsay, R. C. (1985). Evaluation of role of microbial Strecker- derived aroma compounds in unclean-type flavors of cheddar cheese. *Journal of Dairy Science*, 68, 2859–2874.

Esbensen, K. H., & Wagner, C. (2017). Introduction to process sampling. *Spectroscopy Europe/Asia*, 29, 26–30.

Evans, J., Zulewska, J., Newbold, M., Drake, M. A., & Barbano, D. M. (2009). Comparison of composition, sensory, and volatile components of thirty-four percent whey protein and milk serum protein concentrates. *Journal of Dairy Science*, 92, 4773–4791.

Ferretti, A., & Flanagan, V. P. (1971). Lactose–casein (Maillard) browning system: Volatile components. *Journal of Agricultural and Food Chemistry*, 19, 245–249.

Forina, M., Casale, M., & Oliveri, P. (2009). Application of chemometrics to food chemistry. In S. D. Brown, R. Tauler, & B. Walczak (Eds.), *Comprehensive chemometrics* (1st ed., pp. 75–128). Amsterdam, Netherlands: Elsevier.

Frankel, E. N. (1984). Lipid oxidation: Mechanisms, products and biological significance. *Journal of the American Oil Chemists' Society*, 61, 1908–1910.

Gertz, C., Gertz, A., Matthäus, B., & Willenberg, I. (2019). A systematic chemometric approach to identify the geographical origin of olive oils. *European Journal of Lipid Science and Technology*, 121, Article 1900281.

Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing model fit by cross-validation. *Journal of Chemical Information and Computer Sciences*, 43, 579–586.

IDF/ISO. (2009). *Milk and milk products d sensory analysis d Part 3: Guidance on a method for evaluation of compliance with product specifications for sensory properties by scoring*. IDF 099-3. Brussels, Belgium: International Dairy Federation.

Javidipour, I., & Qian, M. C. (2008). Volatile component change in whey protein concentrate during storage investigated by headspace solid-phase microextraction gas chromatography. *Dairy Science & Technology*, 88, 95–104.

Karoui, R., & De Baerdemaeker, J. (2007). A review of the analytical methods coupled with chemometric tools for the determination of the quality and identity of dairy products. *Food Chemistry*, 102, 621–640.

Kilcawley, K. N., Faulkner, H., Clarke, H. J., O'Sullivan, M. G., & Kerry, J. P. (2018). Factors influencing the flavour of bovine milk and cheese from grass based versus non-grass based milk production systems. *Foods*, 7, Article 37.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9, 137–163.

Kühn, J., Considine, T., & Singh, H. (2006). Interactions of milk proteins and volatile flavor compounds: Implications in the development of protein foods. *Journal of Food Science*, 71, R72–R82.

Langford, V. S., Reed, C. J., Milligan, D. B., McEwan, M. J., Barringer, S. A., & Harper, W. J. (2012). Headspace analysis of Italian and New Zealand Parmesan cheeses. *Journal of Food Science*, 77, C719–C726.

Lee, Y. B., Laye, I., Kim, Y. D., & Morr, C. V. (1996). Formation of volatile compounds in whey protein concentrate during elevated temperature storage as a function of water activity. *International Dairy Journal*, 6, 485–496.

Lieberman, M. G., & Morris, J. D. (2014). The precise effect of multicollinearity on classification prediction. *Multiple Linear Regression Viewpoints*, 40, 5–10.

Linting, M., van Os, B. J., & Meulman, J. J. (2011). Statistical significance of the contribution of variables to the PCA solution: An alternative permutation strategy. *Psychometrika*, 76, 440–460.

Lunden, A., Gustafsson, V., Imhof, M., Gauch, R., & Bosset, J. (2002). High trimethylamine concentration in milk from cows on standard diets is expressed as fishy off-flavour. *Journal of Dairy Research*, 69, 383–390.

Maimon, O. Z., & Rokach, L. (2010). Introduction to knowledge discovery and data mining. In O. Z. Maimon, & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (2nd ed., pp. 1–18). New York, NY, USA: Springer.

Makhoul, S., Yener, S., Khomenko, I., Capozzi, V., Cappellin, L., Aprea, E., et al. (2016). Rapid non-invasive quality control of semi-finished products for the food industry by direct injection mass spectrometry headspace analysis: The case of milk powder, whey powder and anhydrous milk fat. *Journal of Mass Spectrometry,* 51, 782–791.

McKnight, L. K., Wilcox, A., & Hripcsak, G. (2002). The effect of sample size and disease prevalence on supervised machine learning of narrative data. *Proceedings of the AMIA Symposium*, 1, 519–522.

Mehta, R. S., Bassette, R., & Ward, G. (1974). Trimethylamine responsible for fishy flavor in milk from cows on wheat pasture. Journal of Dairy Science, 57, 285–289.

Menze, B. H., Geremia, E., Ayache, N., & Szekely, G. (2012). Segmenting glioma in multi-modal images using a generative-discriminative model for brain lesion segmentation. *Proceedings of MICCAI-BRATS*, 56–63.

Metwalli, A. A. M., & van Boekel, M. A. J. S. (1998). On the kinetics of heat-induced deamidation and breakdown of caseinate. *Food Chemistry*, 61, 53–61.

Mi, C., Huettmann, F., Guo, Y., Han, X., & Wen, L. (2017). Why choose random forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. *PeerJ*, 5, Article e2849.

Mortenson, M. A., Vickers, Z. M., & Reineccius, G. A. (2008). Flavor of whey protein concentrates and isolates. *International Dairy Journal*, 18, 649–657.

Mu, F., Gu, Y., Zhang, J., & Zhang, L. (2020). Milk source identification and milk quality estimation using an electronic nose and machine learning techniques. *Sensors*, 20, 4238.

Ng, A. Y., & Jordan, M. I. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naïve Bayes. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *In Proceedings of the 14th international conference on neural information processing systems: Natural and synthetic* (pp. 841–848). Cambridge, MA, USA: MIT Press.

Nie, S., Huang, J., Hu, J., Zhang, Y., Wang, S., Li, C., et al. (2013). Effect of pH, temperature and heating time on the formation of furan in sugar–glycine model systems. *Food Science and Human Wellness*, 2, 87–92.

Patana-anake, P., & Barringer, S. A. (2016). The effect of food additives in fruit drinks on the nosespace using selected ion flow tube mass spectrometry (SIFT-MS). *Access Journal of Food and Agriculture*, 1, 1–10.

Pedrotti, M., Khomenko, I., Cappellin, L., Fontana, M., Somenzi, M., Falchero, L., et al. (2018). Rapid and noninvasive quality control of anhydrous milk fat by PTR-MS: The effect of storage time and packaging. *Journal of Mass Spectrometry*, 53, 753–762.

Pedrotti, M., Khomenko, I., Fontana, M., Somenzi, M., Falchero, L., Arveda, M., et al. (2020). The good, the bad and the aged: Predicting sensory quality of anhydrous milk fat by PTR/SRI-ToF-MS analysis and data mining. *International Dairy Journal*, 109, Article 104729.

Ramzai, J. (2019). *Simple guide to ensemble learning methods*. https://towardsdatascience.com/simple-guide-for-ensemble-learning-methodsd87cc68705a2. (Accessed 26 February 2021).

Rish, I. (2001). An empirical study of the naive Bayes classifier. IJCAI Workshop on Empirical Methods in Artificial Intelligence, 3, 41–46.

Rish, I., Hellerstein, J. L., & Thathachar, J. (2001). *An analysis of data characteristics that affect naive Bayes performance. Technical Report RC21993*. Yorktown Heights, NY, USA: IBM T.J. Watson Research Center.

Scott, I. M., Lin, W., Liakata, M., Wood, J. E., Vermeer, C. P., Allaway, D., et al. (2013). Merits of random forests emerge in evaluation of chemometric classifiers by external validation. *Analytica Chimica Acta*, 801, 22–33.

Shimamura, T., & Ukeda, H. (2012). *Maillard reaction in milk – effect of heat treatment*. In W. L. Hurley (Ed.), Milk protein (pp. 147–158). London, UK: InTechOpen.

Shipe, W. F., Bassette, R., Deane, D. D., Dunkley, W. L., Hammond, E. G., Harper,W. J., et al. (1978). Off flavors of milk: Nomenclature, standards, and bibliography. *Journal of Dairy Science*, 61, 855–869.

Smith, T. J., Campbell, R. E., Jo, Y., & Drake, M. A. (2016). Flavor and stability of milk proteins. *Journal of Dairy Science*, 99, 4325–4346.

Smyth, H., & Cozzolino, D. (2013). Instrumental methods (spectroscopy, electronic nose, and tongue) as tools to predict taste and aroma in beverages: Advantages and limitations. *Chemical Reviews*, 113, 1429–1440.

Sohn, M., & Ho, C.-T. (1995). Ammonia generation during thermal degradation of amino acids. *Journal of Agricultural and Food Chemistry*, 43, 3001–3003.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9, 307–318.

Tomaschek, F., Hendrix, P., & Baayen, R. H. (2018). Strategies for addressing collinearity in multivariate linguistic data. Journal of Phonetics, 71, 249–267.

Ulberth, F., & Roubicek, D. (1995). Monitoring of oxidative deterioration of milk powder by headspace gas chromatography. *International Dairy Journal*, 5, 523–531.

van Ruth, S. M., Koot, A. H., Akkermans,W., Araghipour, N., Rozijn, M., Baltussen, M., et al. (2008). Butter and butter oil classification by PTR-MS. *European Food Research and Technology*, 227, 307–317.

Vazquez-Landaverde, P. A., Torres, J. A., & Qian, M. C. (2006). Quantification of trace volatile sulfur compounds in milk by solid-phase microextraction and gas chromatography–pulsed flame photometric detection. *Journal of Dairy Science*, 89, 2919–2927.

Volkin, D. B., & Klibanov, A. M. (1987). Thermal destruction processes in proteins involving cystine residues. *Journal of Biological Chemistry*, 262, 2945–2950.

Weiss, G. M. (2010). Mining with rare cases. In O. Z. Maimon, & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (2nd ed., pp. 747–758). New York, NY, USA: Springer.

Whitson, M. E., Miracle, R. E., Bastian, E., & Drake, M. A. (2011). Effect of liquid retentate storage on flavour of spray-dried whey protein concentrate and isolate. *Journal of Dairy Science*, 94, 3747–3760.

Wijaya, D. R., Sarno, R., & Daiva, A. F. (2017). Electronic nose for classifying beef and pork using Naïve Bayes. In *International Seminar on Sensors, Instrumentation, Measurement and Metrology (ISSIMM*), Surabaya (Vol. 2017, pp. 104–108).

Witten, I. H., & Eibe, F. (2005). *Data mining. Practical machine learning tools and techniques* (2nd ed.). Amsterdam, Netherlands: Elsevier.

Wüst, J., & Pischetsrieder, M. (2016). Methionine sulfoxide profiling of milk proteins to assess the influence of lipids on protein oxidation in milk. *Food & Function*, 15, 2526–2536.

Xie, J., & Qui, Z. (2007). The effect of imbalanced data sets on LDA: A theoretical and empirical analysis. *Pattern Recognition*, 40, 557–562.

Zhang, Q., & Sun, S. (2009). Weighted data normalization based on eigenvalues for artificial neural network classifications. In C.-S. Leung, M. Lee, & J. H. Chan (Eds.), *16th international conference on neural information processing (ICONIP 2009)* (pp. 349–356). Heidelberg, Germany: Springer. Part 1.

Table 1. HPLC gradient conditions used following automated cartridge cleanup

Factor	% Cumulative variance explained	Major contributing compounds
1	37.4	1, 2, 3, 4, 5, 8, 9 (+)
2	51.6	11, 12, 14, 19 (+),3 (−)
3	61.8	6, 7, 18 (+), 8, 15, 16, 17 (−)
4	70.0	4, 5, 6 (+), 7, 10, 13, 16 (−)
5	75.5	10, 14, 17 (−)
6	80.3	10, 19 (+), 13, 14 (−)
7	84.7	14, 16, 17, 18 (+)
8	88.5	17, 18 (+), 5, 15, 16 (−)
9	91.3	14, 16, 19 (+), 11, 12, 17 (−)
10	93.7	15, 18 (+), 19 (−)
11	95.4	5, 9, 15 (+), 4, 16 (−)
12	96.5	4, 7, 8 (+), 3, 5 (−)

[a] Major contributing compounds are those with coefficient < −0.25 or > +0.2) and direction + or −. Factors 13–19 cumulatively explained 97.5, 98.3, 98.9, 99.3, 99.7, 99.9 and 100% of the variance. Compound order same as in occurrence histogram, e.g., Compound 1 most prevalent.

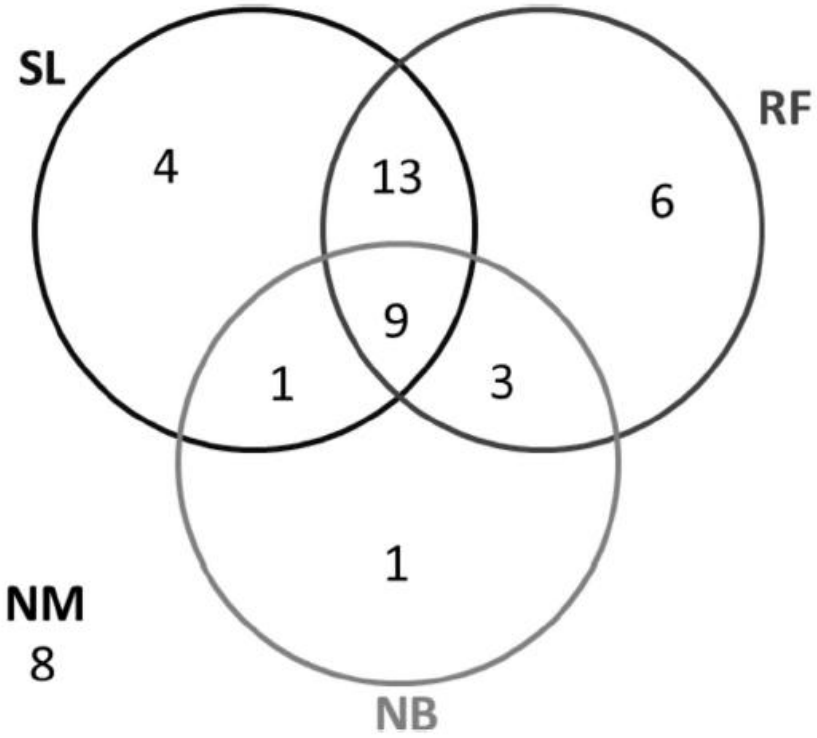

Fig 1. Venn diagram showing which of the models classified the 45 pairs: SL, single logistics; RF, random forest; NB naïve Bayes; NM, not classified by any of the classifier methods

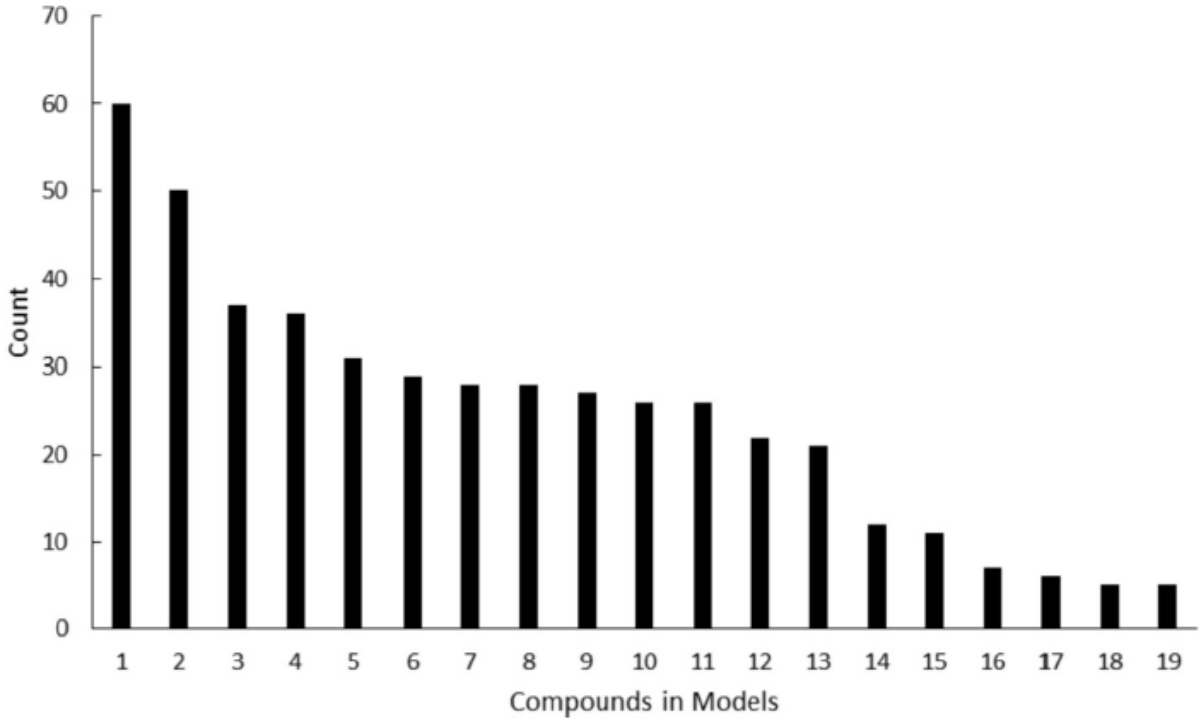

Fig 2. Occurrence of 19 selected predictor variable compounds in the successful models.

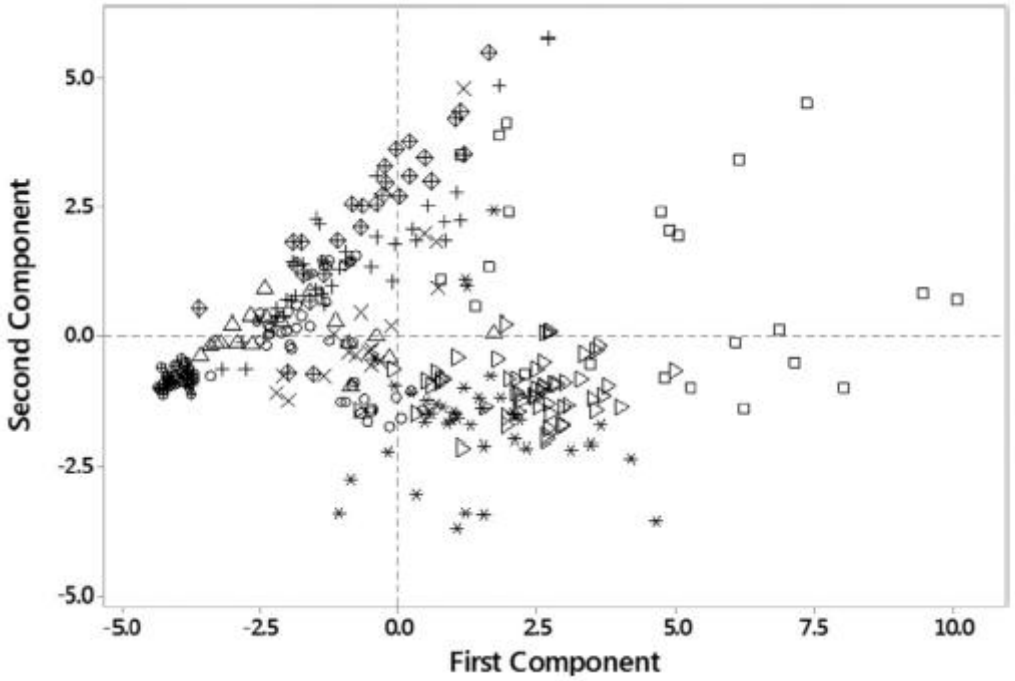

Fig 3. PCA plot of proteins, lactose and milk minerals using the 19 compounds: □, acid casein; ⊠calcium caseinate; ○ , lactalbumin; △ , lactose; ⊕ , milk minerals; ✚ , milk protein concentrate; ▷ , sodium caseinate; ☆ , whey hydrolysate; ◇ , whey protein concentrate; ○ , whey protein isolate