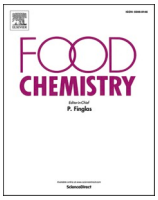
Food Chemistry 447 (2024) 138938

Contents lists available at ScienceDirect

Food Chemistry

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



Untargeted metabolomics and machine learning unveil quality and authenticity interactions in grated Parmigiano Reggiano PDO cheese

Pier Paolo Becchi [a], Gabriele Rocchetti [b,*], Pascual García-Pérez [a,c], Sara Michelini [d], Valentina Pizzamiglio [d], Luigi Lucini [a]

[a] Department for Sustainable Food Process, Università Cattolica del Sacro Cuore, Via Emilia Parmense 84, 29122 Piacenza, Italy
[b] Department of Animal Science, Food and Nutrition, Università Cattolica del Sacro Cuore, Via Emilia Parmense 84, 29122 Piacenza, Italy
[c] Nutrition and Bromatology Group, Analytical and Food Chemistry Department, Faculty of Food Science and Technology, Universidade de Vigo, Ourense Campus, 32004 Ourense, Spain
[d] Parmigiano Reggiano Cheese Consortium, Via J.F. Kennedy, 18, Reggio Emilia 42124, Italy

ARTICLE INFO

Keywords:
Food integrity
Multiblock orthogonal partial least squares
Foodomics
Multivariate statistics
Random Forest classification

ABSTRACT

The chemical composition of Parmigiano Reggiano (PR) hard cheese can be significantly affected by different factors across the dairy supply chain, including ripening, altimetric zone, and rind inclusion levels in grated hard cheeses. The present study proposes an untargeted metabolomics approach combined with machine learning chemometrics to evaluate the combined effect of these three critical parameters. Specifically, ripening was found to exert a pivotal role in defining the signature of PR cheeses, with amino acids and lipid derivatives that exhibited their role as key discriminant compounds. In parallel, a random forest classifier was used to predict the rind inclusion levels (> 18%) in grated cheeses and to authenticate the specific effect of altimetry dairy production, achieving a high prediction ability in both model performances (i.e., ~60% and > 90%, respectively). Overall, these results open a novel perspective to identifying quality and authenticity markers metabolites in cheese.

1. Introduction

Parmigiano Reggiano (PR) cheese is one of the most ancient long-ripened traditional Italian cheese characterized by a Protected Designation of Origin (PDO). The whole production chain, within a restricted geographic area in northern Italy, is strictly regulated by the application of production guidelines defined by the PR Consortium (PDO production specification of PR cheese: https://www.parmigianoreggiano.com/consortium/rules_regulation_2/default.aspx).

Although manufactured following a strict production specification, its chemical composition can be significantly affected by different factors across the dairy supply chain, including primary production practices and specific processing conditions (Santarcangelo et al., 2022; Zannoni, 2010). Among the potential factors influencing cheese quality, ripening has been highlighted as one of the main ones driving cheese composition, involving a series of biochemical reactions and microbial transformations influencing a given cheese product's overall flavour and texture characteristics (McSweeney, 2004). Specifically, the chemical signatures related to proteolytic and lipolytic events in hard cheese have been proposed as valuable tools for monitoring PDO cheese production's quality and authenticity (Rocchetti et al., 2018). Also, the rind inclusion level in grated cheese has been highlighted as a critical aspect considered in the context of the quality and authenticity of PR cheese production (Alinovi, Mucchetti, & Tidona, 2019). Due to the different chemical and textural properties, an excessive amount of rind in grated cheese has been outlined as an important concern, as it negatively impacts the organoleptic properties (Zannoni & Hunter, 2015). To date, no official analytical methods have been recognized for rind estimation in PR grated cheese, having 18% *w*/w of rind as a threshold value defined by the PR production specification. Approaches based on NIR hyperspectral imaging, S3 nanowire gas sensor devices combined with GC–MS, and high-resolution mass spectrometry have been proposed for this aim (Abbatangelo et al., 2018; Calvini, Michelini, Pizzamiglio, Foca, & Ulrici, 2020; Rocchetti, Michelini, Pizzamiglio, Masoero, & Lucini, 2021), although requiring a validation on larger datasets.

Regarding the sustainability of dairy production along the entire production chain, mountain dairy products have gained increasing attention for their distinctive added value in dealing with the

* Corresponding author.
E-mail address: gabriele.rocchetti@unicatt.it (G. Rocchetti).

https://doi.org/10.1016/j.foodchem.2024.138938
Received 23 October 2023; Received in revised form 28 February 2024; Accepted 2 March 2024
Available online 5 March 2024

sustainability claim (Mancini et al., 2019). The production of PR cheese is a tangible example of this trend, where the specificity of mountain productions and their connection to the local environment ensures the preservation of traditions, local producers, and ecological integrity of these areas (Arfini et al., 2019). From a chemical perspective, Becchi, Rocchetti, Vezzulli, Lambri, and Lucini (2023) recently showed that grassland biodiversity has a positive impact on the untargeted metabolite composition of hard cheese, with an increasing content of beneficial compounds (such as linolenic acid and stearidonic acid, both omega-3 polyunsaturated fatty acids, PUFAs). Similarly, Cavallini et al. (2023) recently used a ^{1}H NMR spectroscopy approach combined with multivariate data analysis to provide new insights into the definition of the chemical fingerprint of mountain certifications compared to conventional PR PDO cheese samples. However, studies dealing with the traceability of a mountain production are still limited (Bontempo, Lombardi, Paoletti, Ziller, & Camin, 2012; Moran, Aldezabal, Aldai, & Barron, 2019; Segato et al., 2017), and the global evaluation of its quality and authenticity is still challenging.

Starting from this complex background, in this work we used a combined untargeted metabolomics and machine learning-based chemometrics approach as a novel methodology to discriminate some of the most important interacting factors characterizing the metabolite profile of PR cheese, namely ripening, altitude of PR cheese production, and rind inclusion level. Specifically, this work aims to establish a novel and robust workflow allowing both a quality and authenticity assessment of PR cheese, going beyond the classical application of unsupervised/supervised multivariate statistics. For this purpose, a data mining perspective of the PR metabolomics profile was performed, based on analysis of variance (ANOVA)-multiblock orthogonal partial least squares (AMOPLS) modelling and random forest (RF) classification, to elucidate the impact of critical productive factors and shed light into their potential interactions.

2. Materials and methods

2.1. Collection of cheese samples and experimental design

PR PDO grated cheese samples (n = 60) were supplied by the Parmigiano Reggiano PDO cheese Consortium, classified according to three main variables, namely ripening time (12, 24, and 36 months), altitude of PR cheese production and percentage of rind included. Specifically, the reference used for classifying mountain dairies was based on the Regional Rural Development Program 2014–2020, in which mountain regions were classified as less-favoured areas due to natural, geographical, and economic constraints (Regulation (EU) N° 1305/2013, article 32.3: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32013R1305). In this work, 21 samples were collected from dairies located in mountain areas, whereas the remaining cheese samples were largely produced from lowland areas (i.e., 39 samples). Mixtures with different rind percentages were prepared manually by considering that the outer 6 mm of the wheel should be considered as rind (following the Parmigiano Reggiano PDO production specifications). Hence, the rind was separated from the pulp to obtain pure grated rind and pulp and mixtures with increasing percentages (0%, 18%, 23%, 30%, and 100% *w*/w) were prepared as previously reported by Rocchetti et al. (2021), accounting a total of 164 distinct cheese samples.

Moreover, the Parmigiano Reggiano cheese Consortium gratefully donated sixteen commercially available PR grated hard cheese samples without information regarding the product's ripening. The rind percentage of these samples was first estimated using the NIR instrument available at the Parmigiano Reggiano PDO cheese Consortium, which is routinely used for qualitative analysis. These samples have been considered to test our prediction models. The grated cheese samples used for this study were contained in airtight packages and stored at −18 °C until further metabolomics analysis. The comprehensive list of sample identifiers (IDs) and their classification for both factors under investigation can be found in the supplementary material (Table S1).

2.2. Extraction method of grated cheese samples

The grated cheese samples were thawed at room temperature and extracted as previously reported by Rocchetti et al. (2021), with minor adjustments. Given the extensive sampling conducted for this study, each sample underwent a single extraction process without including technical replicates. Briefly, 2 g of each individual sample was combined with 14 mL of acetonitrile (LC-MS grade, Sigma-Aldrich, Madison, CA) containing 3% (*v*/v) formic acid. The mixture was then vigorously vortexed for 3 min and subjected to a homogenizer-assisted extraction with an Ultra-turrax (Ika T10, Staufen, Germany) for 5 min at room temperature. Afterwards, samples were centrifuged at 7800 ×*g* for 15 min at 4 °C and placed at −18 °C overnight under dark conditions to enable protein precipitation. The resulting supernatants were filtered through 0.20 μm cellulose syringe filters and collected in amber vials until instrumental analysis. Furthermore, quality control (QC) samples were prepared by pooling 6 μL from each of the 180 PR cheese samples into the same UHPLC vial. During the metabolomics analyses of PR cheese samples, QC sample was randomly injected through the sequence, specifically at the beginning, at the end of the batch, and every ten analyzed samples. This QC preparation and injection procedure has been recognized in LC-MS-based untargeted metabolomics investigations to ensure the stability and reliability of the analytical method throughout the entire analysis process (Evans et al., 2020).

2.3. Untargeted metabolomics-based analysis

The fingerprinting of PR cheese samples was conducted using a Vanquish ultra-high-pressure liquid chromatography (UHPLC) equipped with a heated electrospray ionization (HESI)-II probe and coupled to a Q Exactive™ Focus Hybrid Quadrupole-Orbitrap Mass Spectrometer (Thermo Scientific, Waltham, MA, USA) (Rocchetti et al., 2021). Shortly, LC-MS grade water and acetonitrile (Sigma-Aldrich, Milan, Italy) were used as mobile phases, employing a gradient elution (6–94% acetonitrile in 35 min) and 0.1% formic acid as a phase modifier. The chromatographic separation of extracts was carried out on a Zorbax Eclipse Plus C18 column (50 × 2.1 mm, 1.8 μm) from Agilent Technologies (Santa Clara, CA, USA). The temperature on the column compartment was set at 30 °C and a sample volume of 6 μL was injected into the system with a constant flow rate of 200 μL/min. The mass spectrometric data acquisition involved integrating Full scan MS analysis and data-dependent MS/MS mode. Full scan MS analysis was carried out within the *m/z* range from 70 to 1200 using an automatic gain control (AGC) of 1×10^6, maximum injection time (IT) of 200 ms and a mass resolution of 70,000 (FWHM) at m/z 200. Otherwise, data-dependent MS/MS mode was applied only on QC spectra acquired during metabolomic analysis (i.e., reaching a number of 18 QC injections) with precursor fragmentation in positive polarity on the most abundant ions (Top N ions =3) with a reduced full scan mass resolution of 17,500 at *m/z* 200. An isolation window of 1.0 m/z and a maximum IT of 100 ms were applied, considering an AGC target value of 1×10^5. Also, stepped Normalized Collisional Energy (10, 20, 40 eV) was used to fragment the Top N ions. Finally, before analysis, the mass spectrometer was calibrated using Pierce™ positive ion calibration solution mix (Thermo Scientific, Waltham, MA, USA).

Afterwards, the acquired data (.RAW files) were then processed using the MS-DIAL software (version 4.90) (Tsugawa et al., 2015) for automated peak finding, LOWESS normalization, and annotation via spectral matching, using the comprehensive database FooDB and MoNA (Mass Bank of North America) for MS-MS annotation. Features in the mass range of 100–1200 *m/z* with a minimum peak height of 10,000 cps were explored, applying an accurate mass tolerance of 0.05 Da for MS peak centroiding and 0.1 Da for MS/MS. Retention time information was

not considered in the total score calculation. Identification of compounds was based on mass accuracy, isotopic pattern, and spectral matching, resulting in a total identification score with a minimum cut-off of 50%, considering the most common HESI+ adducts. Finally, the peak finder algorithm was used to fill the missing peaks with a 5-ppm tolerance for m/z values. According to Blaženović et al. (2019), a level 2 of identification (i.e., putatively annotated compounds) was reached in our untargeted-based experiments.

2.4. Data processing and statistical analysis

The statistical multivariate analysis was done using three different tools, namely Mass Profiler Professional B.12.06 (from Agilent Technologies) for cluster analysis, SIMCA version 16.0 (Umetrics, Malmo, Sweden) for OPLS-DA, and the "rAMOPLS" package (version 0.2) on R studio (version 4.2.3) for multifactorial ANOVA. Firstly, data were filtered by abundance (only the metabolites having an area > 10,000 counts were included), log2 converted, normalized at the 75th percentile, and baselined against the median calculated among all the samples. Afterward, unsupervised hierarchical cluster analysis (HCA) was performed to group samples according to their similarities by considering the three factors affecting sample variability in our dataset. Then, the metabolomics dataset was subjected to advanced supervised approaches, namely orthogonal partial least squares discriminant analysis (OPLS-DA) and ANOVA multi-block orthogonal partial least squares analysis (AMOPLS). For both methods, models were built including the three experimental factors (i.e., ripening, altimetric zone and rind inclusion level).

Regarding the supervised AMOPLS-based approach, models were built by carrying out 100 permutations and 10-step subsampling combined with 3-step parallelization to manage the variations in the sample population across groups caused by the unbalanced experimental design. Afterward, to ensure the consistency of the ANOVA partitioning, the smallest number of orthogonal components showing statistical significance was first selected, according to the "parsimony principle" shown by Boccard and Rudaz (2016). The generated AMOPLS model was evaluated in terms of statistical significance and predictive assessment through several statistical parameters to assess the statistical significance of each effect and their interaction, such as goodness of fit (R^2Y), residual structure ratio (RSR), residual sum of squares (RSS) and their associated *p*-values. Then, as described by González-Ruiz et al. (2017), the specific calculation of the squared variable importance in the projection (VIP^2) contribution was used to rank the influence of metabolites in prediction for each significant effect and to select the top-50 VIP^2 markers having the highest discrimination potential. In conclusion, by employing these comprehensive statistical approaches, we aimed to gain insight into the metabolomic differences and identify key compounds contributing to the discrimination of the studied conditions.

2.4.1. Machine learning-assisted classification

Classification models were performed using a Random Forest (RF) algorithm, available in the R studio statistical environment (https://CRAN.R-project.org/package=randomForest, accessed 24 May 2023). In particular, the RF algorithm provides a statistically powerful approach for classification and regression tasks in metabolomics research, employing an ensemble of decision trees with randomized feature selection (Kokla, Virtanen, Kolehmainen, Paananen, & Hanhineva, 2019). RF was used to classify the grated cheese samples according to the rind inclusion level and to discriminate mountain PR cheese productions compared to those from lowland areas. For both factors, the entire dataset of samples was divided into train and test groups in order to evaluate overall performance and validate their ability to make accurate predictions. A total of 132 cheese samples were selected to train the model (80% of total samples), whereas the remaining 20% was selected to test the built model. For each model, selecting the number of trees and the number of variables for each split (number of trees = 500; number of variables tried at each split: 52) is crucial for achieving optimal performance and excluding overfitting. Furthermore, once the model was built using the training set, the mean decrease Gini (MDG) index was evaluated to measure the consistency of each feature's importance. This value represents the cumulative reduction in Gini impurity across a given variable, normalized by the total number of trees (Han, Guo, & Yu, 2016). As described by Calle and Urrea (2011), this parameter represents a powerful tool to rank variables according to their importance in the calculated model, with strong and robust predictions. So, the top 30 markers having the highest MDG values were evaluated for each RF model calculated.

3. Results and discussion

3.1. One factor-chemometric analysis on the untargeted metabolomics profile of Parmigiano Reggiano PDO grated samples

In the first part of this work, untargeted metabolomics was employed to investigate the chemical fingerprint of grated PR cheese samples according to the three factors highlighted in the previous sections, namely ripening time (*ripening* factor), altitude of PR cheese production (*altimetric zone* factor), and rind percentage included in the pulp (*rind inclusion* factor). Our findings revealed a total of 2739 mass features that were putatively annotated across samples and, among them, 598 compounds were confirmed based on the MS/MS information (i.e., fragment ions) provided by QC samples. The comprehensive list of cheese metabolites annotated is provided in Table S2, together with their composite spectra, including retention time, raw abundances, accurate mass, MS isotopic spectrum, MS/MS spectrum, and adducts used for annotation.

Unsupervised cluster analysis (HCA) was carried out by considering the average fold-change variations in all the samples based on the three main factors under investigation (i.e., *ripening*, *altimetric zone,* and *rind inclusion*, respectively), to naïvely explore dataset patterns across treatments. Specifically, the HCA plot was split into two sections to better understand each factor's contribution to the overall metabolomic profile of PR cheeses. Fig. 1 showed the combined interaction of *ripening* and *altimetric zone*, while Fig. S1 reported the effect of rind inclusion on the overall PR cheese samples. As can be observed from Fig. 1, *ripening* was the hierarchically prevalent factor driving the clustering of cheese samples, reflecting a potential discriminant profile associated with 12-month ripened samples (except for one sample type that clustered rather distant from the others; 12 months; Lowland; 100% rind). In contrast, the longest ripening stages (i.e., 24 and 36 months) exhibited overlapping profiles. This indicates that the main changes in cheese metabolome due to ripening are prevalent over other factors. Also, a defined unique profile was reported for commercial samples, exhibiting a metabolic closeness to less ripened cheese samples. These findings suggest that they might have reached a minimum ageing threshold allowed by the specifications of Parmigiano Reggiano PDO. In parallel, based on *altimetric zone* information, inside the two main clusters enhanced by the ripening effect, mountain PR samples could be discriminated from cheeses produced mainly from lowland areas, enhancing the potential impact of feeding practices, climatic conditions and other factors in modulating the metabolomic profile of PR cheese. Moreover, looking at the HCA plot given by the *rind inclusion* effect (Fig. S1), no significant trends were observed regarding the different percentages of rind included in the grated samples, indicating that this factor has a lower influence on discrimination among the cheese samples. Lastly, the overall outcome provided by cluster analysis may indicate a complex and intricate interaction between all the factors involved in this study, meriting further investigation by applying machine learning algorithms as described later.

Subsequently, a supervised OPLS-DA analysis was conducted to examine and assess the individual impact of different groups of cheese metabolites for discrimination purposes. In particular, OPLS-DA models

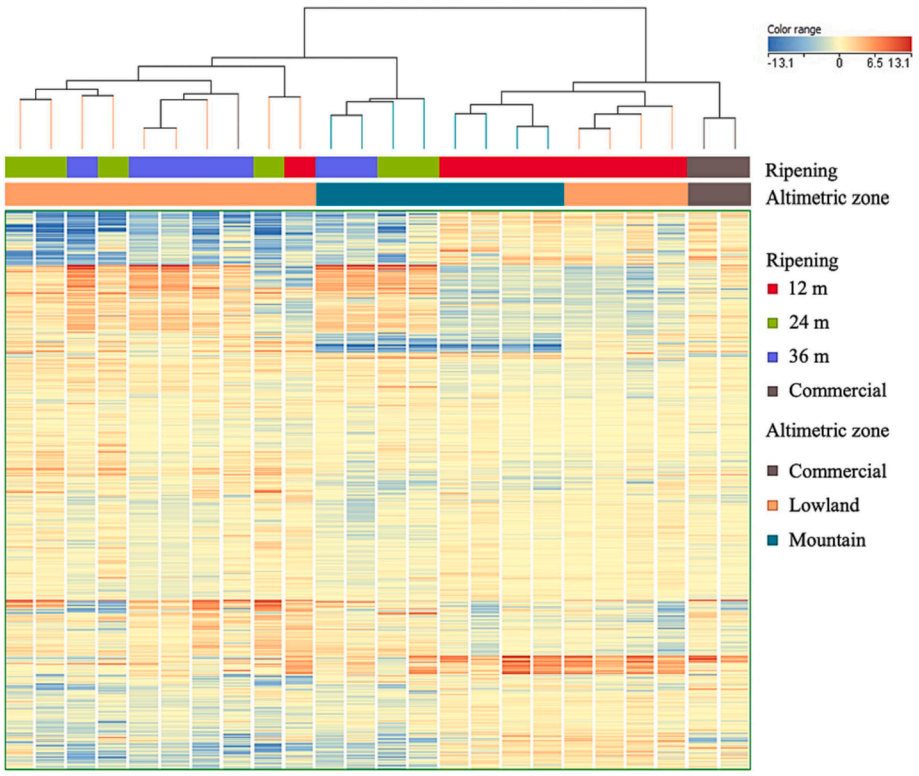


Fig. 1. Heat map resulting from the unsupervised hierarchical cluster analysis (HCA) on the metabolomic profile of the different Parmigiano Reggiano cheese samples under investigation for ripening and altimetric zone effects.

were carried out for each experimental factor, namely *ripening, altimetric zone* and *rind inclusion* factors (Fig. S2). The quality parameters in terms of goodness-of-fit, R^2Y, and goodness-of-prediction, Q^2Y, for each OPLS model indicate that both *ripening* and *altimetric zone* were found as discriminant factors ($R^2Y \approx 1.0$; $Q^2Y > 0.5$), whereas rind inclusion did not meet quality requirements for discrimination (Table S3). Moreover, OPLS-DA models were statistically validated by cross-validation ANOVA, highlighting significant *p*-values for *ripening* and *altimetric zone* factors ($p < 0.001$; Table S3), and overfitting was excluded by the ROC plots (Fig. S3), as well as a variation <30% between R^2Y and Q^2Y values (Veerasamy et al., 2011).

Providing the discriminant power attributed to *ripening* and *altimetric zone*, the corresponding OPLS models were combined with the variable importance in projection analysis (VIP) to identify the metabolites showing the highest contribution to discrimination for each factor, called as VIP markers (Tables S4 and S5, respectively), which were classified according to their chemical ontology provided by FooDB database in Table 1. Each class of VIP markers was combined with the average VIP score and log FC values to better interpret the net effect of each factor on the metabolome of PR samples.

A substantial separation among the different clusters was observed regarding the *ripening* effect, with the 24-month ripened samples positioned outside the 95% confidence limit, indicating significant distance from the other two classes (Fig. S2). This discrepancy could be attributed to variations in the metabolomic profiles among the different samples. Additionally, it is important to highlight that the model may be influenced by the disparity in the number of analyzed samples, with a lower representation of the 24-month category. In this regard, to enhance the accumulation trends of the different VIP metabolites in PR cheese samples, the VIP selection method was coupled with fold-change analysis (FC value >1.2) for the "36 m vs 12 m" comparison. Overall, as can be observed from Table 1, among the 50 compounds having the highest discriminant potential, we found that were mainly representative of the 36-month-PR cheeses, as highly positive log FC values were reported with respect to 12-month-cheese. Regarding the most enriched chemical classes highlighted among the VIP markers, we found mainly amino acids and derivatives (12 compounds), benzene derivatives (5 compounds), fatty acyls (4 compounds), and organoheterocyclic compounds (6 compounds). As can be observed, amino acids were identified as the most representative markers among the various VIP compounds, consistent with our previous findings regarding the overall impact of *ripening* on the PR cheese metabolomics profile (Becchi et al., 2023). Notably, proteolysis emerges as the principal biochemical process during cheese ageing, involving the breakdown of initial caseins into small peptides and amino acids by the enzymatic activity of lactic acid bacteria (LAB) and non-starter lactic acid bacteria (NSLAB) proteases (Sforza et al., 2012). In this regard, by looking at our findings, the VIP amino acids exhibited an average VIP score of 1.731 together with a high LogFC value for the "36 m vs 12 m" comparison (average LogFC value = 1.74), with asparaginyl-phenylalanine compound showing the highest VIP value associated to this chemical class (VIP score = 1.8). Specifically, as ripening progressed, all amino acids contributed positively to enriching the 36-month cheeses, with an overall up-accumulation trend detected among all the different dipeptides highlighted by the VIP approach. In particular, the two dipeptide derivatives of phenylalanine, namely lysyl-phenylalanine (LogFC = 2.43) and asparaginyl-phenylalanine (LogFC = 1.74) were characterized by high LogFC values, confirming phenylalanine's involvement in the proteolysis processes of long-ripened cheese (Ochi et al., 2013). The contribution of phenylalanine during cheese ripening could be related to the catabolism of *Lactobacillus* spp. and it has been identified in Cheddar cheese as a precursor of cheese flavour formed by the catabolism of lactic acid bacteria (Gummalla & Broadbent, 2001; Singh, Drake, & Cadwallader, 2003). Moreover, by inspecting the other dipeptides found as discriminants in the *ripening* OPLS-DA model, we found several amino acids that have been already highlighted as the main precursors of key compounds, such as tryptophan, leucine and valine (Atanasova et al., 2021). Specifically, their presence in cheese is potentially derived from

Table 1
Comprehensive summary of the discriminant compound classes calculated for the top-50 VIP markers of OPLS-DA models built considering the cheese samples classified according to ripening and altimetric zone factors.

OPLS-DA model	Enriched chemical class	Number of VIP discriminant metabolites	Most discriminant compounds (OPLS-DA)	Average VIP scores	Average LogFC scores
Ripening					*(36 m* vs *12 m)*
	Amino acids, peptides, and analogues	12	Asparaginyl-Phenylalanine (VIP score = 1.8)	1.731	1.74
	Benzene and substituted derivatives	5	2-(1-Naphthyl)acetamide (VIP score = 1.79)	1.713	2.78
	Carbohydrates and carbohydrate conjugates	3	2,3-Butanediol apiosylglucoside (VIP score = 1.766)	1.733	2.74
	Carbonyl compounds	1	3,5-Didecanoylpyridine	1.723	5.33
	Cinnamic acids and derivatives	1	alpha-Terpinyl cinnamate	1.683	2.75
	Fatty Acyls	4	Isobutyryl-L-carnitine (VIP score = 1.757)	1.726	2.2
	Glycerolipids	1	MG(14:1(9Z)/0:0/0:0)	1.705	1.41
	Linoleic acids and derivatives	2	Stearidonic acid (VIP score = 1.78)	1.76	1.98
	Nucleosides, nucleotides, and analogues	2	S-adenosyl-L-methionine (VIP score = 1.715)	1.704	1.36
	Organic acids and derivatives	3	N1-Acetylspermidine (VIP score = 1.704)	1.692	1.99
	Organic nitrogen compounds	1	Dehydrospermidine	1.798	2.96
	Organoheterocyclic compounds	6	2-Acetyl-1,5,6,7-tetrahydro-6-hydroxy-7-(hydroxymethyl)-4H-azepine-4-one (VIP score = 1.778)	1.74	2.23
	Organosulfur compounds	1	Urocanic acid	2.033	0.09
	Purine nucleosides	1	N-6-Isopent-2-enyl-adenosine	1.795	3.53
	Sesquiterpenoids	1	alpha-Santalyl acetate	1.786	3.21
	Other	6	4-Amino-2-methyl-1-naphthol (VIP score = 1.723)	1.692	1.63
Altimetric zone					*(Mountain* vs *Lowland)*
	Amines	3	Palmitoleoyl ethanolamide (VIP score = 2.772)	2.676	−1.3
	Amino acids, peptides, and analogues	6	Pentacosanoylglycine (VIP score = 3.233)	2.652	0.85
	Benzene and substituted derivatives	1	5-Hydroxynorvaline-betaxanthin	3.149	1.76
	Carbonyl compounds	1	4,4'-Thiobis-2-butanone	2.795	0.74
	Diterpenoids	1	Crocin 2	2.731	2.68
	Fatty Acyls	11	Octyl phenylacetate (VIP score = 2.934)	2.529	−3.51
	Glycerolipids	1	TG(18:2(9Z,12Z)/16:0/18:3(9Z,12Z,15Z))[iso6]	2.24	−6.07
	Glycerophospholipids	4	PA(16:0/20:1(11Z)) (VIP score = 3.191)	2.958	−6.4
	Indoles and derivatives	1	Nb-Tricosanoyltryptamine	2.443	−8.59
	Lipids and lipid-like molecules	7	Cycloartanyl ferulate (VIP score = 3.171)	2.62	−4.57
	Organic acids and derivatives	3	Protocatechuic acid 4-O-sulphate (VIP score = 3.038	2.653	−1.84
	Organoheterocyclic compounds	4	4,5-Dimethyl-2-octylthiazole (VIP score = 3.118)	2.61	0.01
	Organosulfur compounds	2	2-Methyl-3-(methylthio)pyrazine (VIP score = 2.36)	2.36	0.71
	Phenols	3	Norcapsaicin (VIP score = 2.572)	2.371	−1.4
	Triterpenoids	2	Feruloyldihydro-beta-sitosterol (VIP score = 2.911)	2.808	−3.66

casein hydrolysis by starter or non-starter/secondary starter bacteria, accounting for 56–70% of the recorded free amino acids after cheese proteolysis. Finally, also suberylglycine was detected as a potential *ripening* biomarker and its impact on the ageing time of Parmigiano Reggiano cheese has already been depicted in our previous work (Becchi et al., 2023). Additionally, the lipid fraction consisted mainly of fatty acyls (4 compounds), glycerolipids (1 compound), and linoleic acid derivatives (2 compounds). All these chemical classes exhibited overall accumulation trends associated with the more aged cheeses (i.e., 36 months-PR cheese samples), with higher accumulation trends detected for fatty acyls (average LogFC value = 2.2). Specifically, among fatty acyls, two different fatty acyl esters (i.e., 1-ethylhexyl tiglate, isobutyryl-L-carnitine), one hydroxy derivative (i.e., (9S,10E,12Z)-9-hydroperoxy-10,12-octadecadienoate), and one N-acyl derivative (i.e., N-palmitoyl alanine) were identified. In particular, among lipid compounds, the presence of fatty acyl esters played a major role in the development of traditional sensory characteristics of Grana cheese and acted as precursors in a wide range of different reactions for the synthesis of alcohols, aldehydes, and lactones (Thierry et al., 2017).

Regarding the OPLS-DA model built for *altimetric zone* factor, as reported by Table S3, the comparison *"mountain vs. lowland"* revealed a clear differentiation between the two product categories with high fitting and prediction performances (i.e., R^2cum = 0.934; Q^2cum = 0.806). As can be observed from Table 1, several chemical classes of the 50 most discriminant VIP compounds (having a VIP score > 2.200) were found to mainly characterize *lowland* PR cheese samples. Among them, the chemical classes of indoles (average LogFC value = −8.59), glycerophospholipids (average LogFC value = −6.4), glycerolipids (average LogFC value = −6.07), lipid derivatives (average LogFC value = −4.57), triterpenoids (average LogFC value = −3.66) and fatty acyls (average LogFC value = −3.51) molecules chemical classes showed the lowest contribution in terms of average LogFC values for the comparison under investigation. Interestingly, fatty acyls were found to be the most representative class, with 11 compounds that were highlighted by the VIP approach. As can be observed, fatty acyl VIP markers consisted mainly of long-chain fatty acids and fatty alcohol derivatives (i.e., 5 and 4 compounds, respectively), with the highest VIP contribution showed by octyl phenylacetate (VIP score = 2.934). Moreover, glycerophospholipids also exhibited a significantly high average VIP score (average VIP score = 2.958) and a high average LogFC value of −6.4, indicating their potential enrichment in lowland PR cheese samples. In particular, 4 different glycerophospholipids were found by our VIP approach, mainly 3 different phosphatidic acids (i.e., PA(P-16:0/18:2(9Z,12Z)), PA(16:0/18:0), PA(16:0/20:1(11Z))) and one lysophosphatidylcholine (LysoPC(18:0/0:0)). Moreover, among the lipid-derived compounds, we found some specific feed-derived compounds, such as cycloartanyl ferulate (VIP score = 3.171), 2,4-Methylene cholesterol (VIP score = 2.412) and SM(d19:0/24:1(15Z)) (VIP score = 2.262). Particularly, cycloartanyl ferulate has been reported to be one of the main steryl ferulates found in millets and cereals (Tsuzuki et al., 2018).

Specifically, this compound has been found to account for >95% of the total steryl ferulate content in barley, commonly used as concentrates to complete the cow's daily ration in PR cheese production (Buonaiuto et al., 2021). Moreover, by inspecting the other chemical classes, no clear trends were revealed by the accumulation trends for amino acids, showing in the VIP markers an almost equal distribution among the two different types of PR cheeses under investigation, with Pentacosanoylglycine (VIP score = 3.233) showing the higher VIP score value. Finally, several derivative organic compounds have been detected as discriminant metabolites, particularly those derived from organic acids (3 compounds), organosulfur (2 compounds), and organoheterocyclic compounds (4 compounds). To date, no literature reports have identified these compounds as potential markers attributed to differences in altitude of cheese production.

3.2. AMOPLS models based on ripening and altimetric zone factors

Considering the discriminating effect and the possible interaction between *ripening* and *altimetric zone* on the metabolomic fingerprint of PR cheese, an AMOPLS approach was further applied to decipher the influence of each specific factor involved in Parmigiano Reggiano PDO production. This approach was applied due to the limits of the previous supervised OPLS-DA models in the accurate discrimination of cheese samples with different ageing and to better extrapolate the relative importance of each factor under investigation. AMOPLS approach incorporates a modelling strategy by simultaneously considering multiple response variables (Guisset, Martin, & Govaerts, 2019). This is particularly advantageous when dealing with multi-dimensional datasets, as it considers the eventual interactions among the response variables, leading to a more comprehensive and deeper analysis. Overall, starting from the selection of orthogonal components, three statistically significant models were obtained (p-value = 0.01) and the "parsimony principle" of taking one distinctive orthogonal component (to) was chosen, with an R^2Y of 0.95. The plot containing the first three orthogonal components with the R^2Y parameters, and associated p-values, is reported in Fig. S4. The resulting AMOPLS score plots are reported in Fig. 2, and well-defined clusters were provided for *ripening*, *altimetric zone*, and the combined effect derived from the two interacting factors (i.e., *ripening* × *altimetric zone* interaction). The statistical parameters associated with each effect are summarised in Table 2, with relative sum of squares (RSS), residual structure ratio (RSR), and their corresponding p-values, as well as the contributions to each predictive and orthogonal component (i.e., tp and to, respectively). The results indicate that *ripening* was found to be a statistically significant discriminant factor (RSR p-value = 0.01), as it showed the highest contribution to the model (RSS = 36.6%), as well as the highest RSR value, 2.273 (Table 2). Moreover, this effect was also associated with the highest contribution to the two principal predictive components, tp1 and tp2 (99.6% and 98.3%, respectively), together with the lowest contribution for the orthogonal component, 14.0%. In contrast, the rest of the effects, i.e., the *altimetric zone* and the interaction *ripening* × *altimetric zone* did not show a statistical significance, with RSS values below <10%, being poorly representative of the variable observed. However, all factors presented significant R^2Y values, indicating their discriminant power, which agreed with the previous results of OPLS modelling. Additionally, the residuals accounted for the remaining 46.4% of the total observed variability, indicating the presence of other sources of variation within the dataset.

Considering these results, *ripening* was confirmed as the most critical factor regarding PR production, and its associated AMOPLS model was combined to determine the variables playing the strongest influence in the projection. In this case, the robustness and efficacy ascribed to AMOPLS modelling enabled the cumulative contribution of each metabolite based on their relative importance in the discrimination among each of the tested factors, selecting the *ripening* effect as the ruling information to obtain the VIP^2 markers. Table 3 shows the first 50 top metabolites highlighted through the VIP^2 approach, grouped according to their chemical class, in which it was reported the cumulative and effect-specific VIP^2 values for *ripening* factor related to the AMOPLS model, as well as LogFC values and regulation for the comparison "36 m vs 12 m". As can be observed, the markers showing the highest VIP^2 score value for *ripening* factor were alpha-santalyl acetate (VIP^2 score *ripening* = 2.606), dehydrospermidine (VIP^2 score *ripening* = 2.605), N-palmitoyl alanine (VIP^2 score *ripening* = 2.551) or asparaginyl-tryptophan (VIP^2 score *ripening* = 2.547). As depicted by the previous OPLS-DA model, several amino acidic and lipid derivatives were

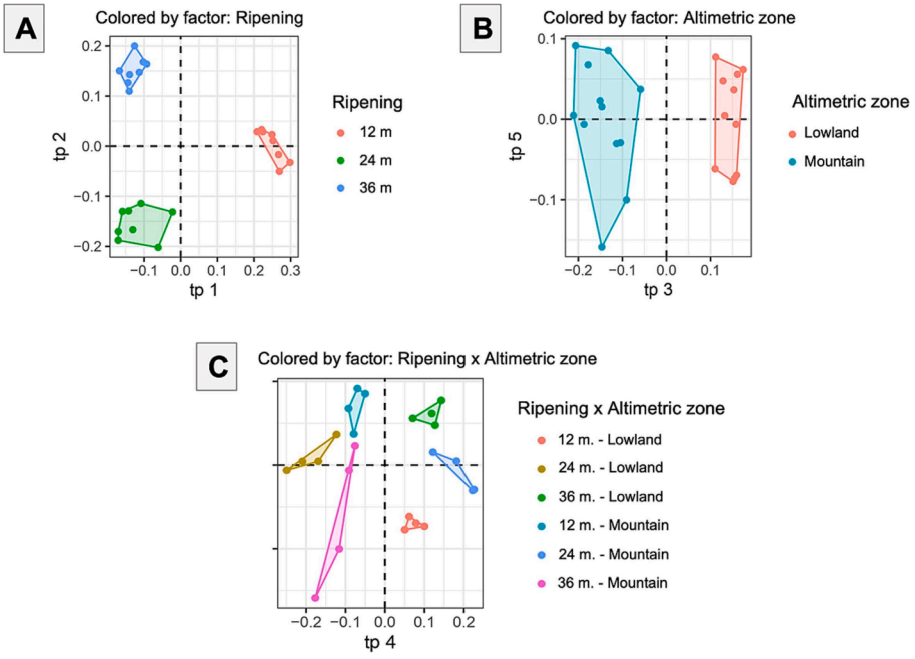


Fig. 2. Supervised AMOPLS analysis score plots built considering the metabolomic profile of the different grated cheese samples using (a) *ripening*, (b) *altimetric zone* and (c) *ripening x altimetric zone* factors as discriminant parameters.

Table 2
Relative variability and block contributions of the AMOPLS analysis of Parmigiano Reggiano cheese data acquired considering the three different factors under investigation (i.e., *ripening*, *altimetric zone*, and *ripening x altimetric zone*, respectively).

Effect	RSS (%)	RSR	RSS p-value	RSR p-value	R^2Y p-value	Block contributions (%)					
						tp1	tp2	tp3	tp4	tp5	to
Ripening	36.6%	2.273	0.01	0.01	0.01	99.6%	98.3%	0.3%	0.3%	7.7%	14.0%
Altimetric zone	8.7%	1.21	1.00	1.00	0.01	0.1%	0.5%	98.5%	0.7%	14.5%	26.3%
Ripening x Altimetric zone	8.3%	1.145	1.00	1.00	0.01	0.1%	0.5%	0.6%	98.2%	60.2%	27.8%
Residuals	46.4%	1.00	N/A	N/A	N/A	0.2%	0.6%	0.7%	0.8%	17.5%	31.9%

Table 3
Cumulative and effect-specific VIP^2 values for the 50 compounds showing the largest contribution for *ripening* factor related to the AMOPLS model.

Classification	VIP^2 compound	VIP^2 score Ripening	VIP^2 score Altimetric zone	VIP^2 score Rip x Altim. zone	VIP^2 cum scores	LogFC scores *(36 m* vs *12 m)*
Alkaloids and derivatives	7-Acetylintermedine	2.347	0.251	0.003	2.601	5.10
Amino acids, peptides, and analogues	Asparaginyl-Tryptophan	2.547	0.130	0.322	2.999	3.82
	Lysyl-Phenylalanine	2.519	0.026	0.061	2.606	2.43
	Hydroxyprolyl-Arginine	2.476	0.056	0.271	2.803	4.27
	N-Benzoylaspartic acid	2.450	0.070	0.015	2.535	2.34
	Asparaginyl-Phenylalanine	2.446	0.076	0.392	2.914	1.74
	Suberylglycine	2.425	0.046	0.059	2.529	1.94
	Serylmethionine	2.351	0.262	0.208	2.822	0.53
	N-Methyl-D-aspartic acid	2.337	0.002	0.744	3.083	0.83
	Arginylisoleucine	2.325	0.421	0.528	3.274	−6.93
	Isoleucyl-Arginine	2.312	0.014	0.356	2.682	3.00
Benzene and substituted derivatives	Fabatin	2.533	0.096	0.510	3.139	3.99
	2-(1-Naphthyl)acetamide	2.445	0.136	0.395	2.975	2.82
	Ethyl 3,4,5-trimethoxybenzoate	2.302	0.071	0.068	2.441	−1.06
Carbohydrates and carbohydrate conjugates	Isopropyl beta-D-glucoside	2.333	0.018	0.035	2.386	1.20
	trans-1,2,10-Trihydroxydihydrolinalyl oxide 7-glucoside	2.316	0.240	0.568	3.125	4.44
	2,3-Butanediol apiosylglucoside	2.303	0.002	0.047	2.352	1.44
	beta-D-Glucosamine	2.494	0.057	0.086	2.637	2.59
Fatty Acyls	N-palmitoyl alanine	2.551	0.044	0.217	2.812	2.33
	1-Ethylhexyl tiglate	2.423	0.312	0.087	2.822	2.07
	Dodecanamide	2.310	0.610	0.013	2.933	3.08
Flavonoids	7-Hydroxy-3′,4′,5,6-tetramethoxyflavone	2.356	0.082	0.016	2.455	0.89
	Myricetin 3-[galloyl-(− > 2)-4-acetyl-a-L-rhamnoside]	2.329	0.005	0.026	2.360	−2.95
Glycerolipids	3-beta-D-galactosyl-sn-glycerol	2.497	0.044	0.270	2.811	0.98
Indoles and derivatives /D	5-Methoxydimethyltryptamine	2.367	0.378	0.151	2.895	2.12
Lignans, neolignans and related compounds	Isolariciresinol 9'-O-alpha-L-arabinofuranoside	2.397	0.009	0.470	2.876	3.59
Linoleic acids and derivatives	(9Z,11E,14Z)-(13*S*)-hydroperoxyoctadeca-(9,11,14)-trienoate	2.474	0.011	0.040	2.525	1.32
	Stearidonic acid	2.458	0.110	0.351	2.918	2.63
	(9S,10E,12Z)-9-hydroperoxy-10,12-octadecadienoate	2.334	0.018	0.306	2.658	1.40
Lipids and lipid-like molecules	(1(10)E,4E,6a,8b)-8-Angeloyloxy-14-oxo-1(10),4,11(13)-germacratrien-12,6-olide	2.391	0.060	0.026	2.477	1.97
Nucleosides, nucleotides, and analogues	S-adenosyl-L-methionine	2.449	0.000	0.387	2.836	1.18
	Orotidylic acid	2.313	0.027	0.459	2.800	−0.97
Organic acids and derivatives	N1-Acetylspermidine	2.334	0.623	0.061	3.018	2.70
Organic acids and derivatives	3-Hydroxydodecanedioic acid	2.333	0.317	0.035	2.685	1.86
Organic nitrogen compounds	Dehydrospermidine	2.605	0.010	0.151	2.766	2.96
Organoheterocyclic compounds	2-Acetyl-1,5,6,7-tetrahydro-6-hydroxy-7-(hydroxymethyl)-4H-azepine-4-one	2.505	0.288	0.018	2.811	3.57
	Junosine	2.403	0.016	0.557	2.977	1.10
	2-Heptyl-4,5-dimethylthiazole	2.371	0.566	0.046	2.982	2.98
Purine nucleosides	N-6-Isopent-2-enyl-adenosine	2.515	0.046	0.125	2.686	3.53
Sesquiterpenoids	alpha-Santalyl acetate	2.606	0.108	0.244	2.957	3.21
Purine nucleosides	N-6-Isopent-2-enyl-adenosine	2.515	0.046	0.125	2.686	3.53
Other compounds	9-(3,4-Dimethoxyphenyl)-2-methoxy-1H-phenalen-1-one	2.490	0.126	0.113	2.729	2.78
	1-Aminopyrene	2.454	0.054	0.017	2.525	2.05
	Salsoline-1-carboxylate	2.410	0.113	0.015	2.538	2.20
	2-Hydroxyestrone-1-S-glutathione	2.390	0.014	0.474	2.878	3.52
	Arctic acid B	2.320	0.306	0.069	2.694	3.29
	Prostaglandin E-2	2.319	0.164	0.837	3.321	4.26
	3-Hydroxy-carbofuran	2.319	0.100	0.002	2.420	5.46
	N-Acetylputrescine	2.317	0.601	0.436	3.354	0.84
	1,3,5-Trihydroxy-10-methylacridone	2.302	0.231	0.894	3.426	3.92

identified as metabolite markers, enhancing the potential major contribution of proteolysis and lipolysis processes during the ageing process of Parmigiano Reggiano PDO cheese. Delving deeper into the amino acid class, the AMOPLS analysis revealed the major impact of amino acids in describing the 36-month cheeses, with the overall up-accumulation trends also observed for amino acid VIP^2 markers. Consistently to has been highlighted previously by the OPLS-DA model, we found dipeptide derivatives of phenylalanine and tryptophan, such as Asparaginyl-Tryptophan (VIP^2 score ripening = 2.547; Log FC value = 3.82), Lysyl-Phenylalanine (VIP^2 score ripening = 2.519; Log FC value = 2.43) and Asparaginyl-Phenylalanine (VIP^2 score ripening = 2.446; Log FC value = 1.74), enhancing their role in enriching cheese with longer ageing time. However, new dipeptides of arginine and aspartic acids were found with AMOPLS approach, namely hydroxyprolyl-arginine (VIP^2 score *ripening* = 2.476; LogFC value = 4.27), N-benzoylaspartic acid (VIP^2 score *ripening* = 2.45; LogFC value = 2.34), *N*-methyl-D-aspartic acid (VIP^2 score *ripening* = 2.337; LogFC value = 0.83) and arginylisoleucine (VIP^2 score *ripening* = 2.325; LogFC value = −6.93). Moreover, going into more detail about the VIP^2-effect specifically, a few amino acids were also related to the interaction of *ripening* and *altimetric zone* factors, such as *N*-methyl-D-aspartic acid (VIP^2 *Rip x altim.zone* score = 0.744), arginylisoleucine (VIP^2 *Rip x altim. zone* score = 0.528) and asparaginyl-phenylalanine (VIP^2 *Rip x altim. zone* score = 0.392). This suggests that the amino acid composition might be mainly affected by the ageing process of cheese, but a secondary influence could be attributed to the interaction with *altimetric zone* effect. Furthermore, by inspecting the lipid derivatives highlighted by the model, our approach highlighted the contributions of included fatty acid amides (i.e., dodecanamide, N-palmitoyl alanine), one fatty acid ester (i.e., 1-Ethylhexyl tiglate), one glycerolipid (i.e., 3-beta-D-galactosyl-sn-glycerol) and linolenic acid derivatives (i.e., (9Z,11E,14Z)-(13*S*)-hydroperoxyoctadeca-(9,11,14)-trienoate, stearidonic acid, (9S,10E,12Z)-9-hydroperoxy-10,12-octadecadienoate) (Table 3). In particular, by observing the LogFC values associated, they were found to mainly characterize cheese samples with the higher ripening stage, with positive values detected among all the lipid VIP^2 markers. Regarding the role of lipid metabolites in the model, most showed the largest contribution to *ripening*, with lower VIP^2 contributions related to the other effects considered. This indicates that lipid changes during the cheese maturation process could substantially contribute to its distinctive characteristics, impacting its flavour and overall quality profile. The only compounds that exhibited a particular influence with the combined effect of *ripening* and *altimetric zone* were stearidonic acid (VIP^2 score *rip* x *altimetric zone* = 0.351) and (9S,10E,12Z)-9-hydroperoxy-10,12-octadecadienoate (VIP^2 score *rip* x *altimetric zone* = 0.306). Regarding the other chemicals that were found particularly predominant in describing the most ripened cheeses, we found 4 different carbohydrates characterized by high LogFC score values, such as three different *O*-glycosyl compounds, such as isopropyl beta-D-glucoside (LogFC value = 1.2), trans-1,2,10-trihydroxydihydrolinalyl oxide 7-glucoside (LogFC value = 4.44), 2,3-Butanediol apiosylglucoside (LogFC value = 1.44) and one hexose, namely beta-D-Glucosamine (LogFC value = 2.59). Overall, these observations justify the deep knowledge provided by the AMOPLS approach, filling the gap obtained by one-factor OPLS-DA modelling and highlighting the critical role of ripening on the metabolomic profile of PR PDO cheese. In addition, the efficacy of AMOPLS modelling in simultaneously identifying the VIP^2 markers associated with discrete effects opens a wide perspective on identifying food markers. This way, although *altimetric zone* and the interaction *ripening* × *altimetric zone* did not show statistically significant discrimination, they showed associated VIP^2 markers (Tables S6 and S7, respectively). Interestingly, the VIP^2 markers reported for *ripening* × *altimetric zone* suggest that *altimetric zone* played a higher contribution within this interaction, as indicated by the higher VIP^2 scores observed for the single effect of *altimetric zone* than those observed for *ripening* (Table S7), as mainly found for 1-pyrrolidinecarboxaldehyde (VIP^2 cumulative score = 9.184, VIP^2 *altimetric zone* score = 4.021), 2,4-pentanedione (VIP^2 cumulative score = 9.202, VIP^2 *altimetric zone* score = 3.754), and N-ethylpropionamide (VIP^2 cumulative score = 9.106, VIP^2 *altimetric zone* score = 3.718). Overall, these trends revealed the potential effect of *altimetric zone* in modulating the chemical profile of cheese, specifically the enhancing impact of the mountain PR cheese production, outlining the importance of production area, milk processing, agricultural practices, and all the comprehensive additional quality parameters on the chemical fingerprint of this unique and distinctive cheese product.

3.3. Random forest algorithm discerns rind inclusion level in grated hard cheese and mountain production of PR cheese

3.3.1. Random forest classifier on the rind inclusion of PR cheese

In the final part of this work, a Random Forest (RF) classifier algorithm was further employed to assess the predictive accuracy of rind inclusion levels in PR cheese samples based on their metabolomic profiles. As previously reported, OPLS-DA modelling did not satisfactorily discriminate the effect of *rind inclusion*, with very low fitting and prediction ability, thus driving the application of RF machine learning approaches to gain insight into the definition of this effect. Once built the model starting from our data, we applied the algorithm to assess its efficacy in classifying rind inclusion levels in commercial PR cheese samples. For these samples, rind inclusion levels were previously estimated by using NIR analysis, which is commonly used by the PR Consortium to determine general quality parameters, including the rind content. Due to the limited predictive capacity of NIR estimation, samples were classified only if the estimated rind content fell below or exceeded 18%, as defined by the production regulations of Parmigiano Reggiano PDO. Starting from our data, the first RF algorithm model was implemented by setting two groups: ≤18% for low-rind samples and > 18% for high-rind samples. The results of the obtained RF classifier models are shown in Table S8. Starting from the calibration samples, the predictive capacity was tested to estimate the rind content on our grated samples. As can be observed, the model showed a prediction ability of ≈ 60% (i.e., 59.38%), a significantly better result compared to the OPLS-DA model performance built for *rind inclusion*, with a higher error rate in estimating high-rind samples (>18%; Table S8). The obtained error rate, around 30% (i.e., 31.06%), could respond to the heterogeneity of the established groups in terms of rind proportion, as the low-rind group, ≤18%, included samples with 0% and 18% rind, whereas the high-rind group, >18%, contained samples presenting 23%, 30%, and 100%. As a further investigation on the *rind inclusion* level on cheese profile, we restricted the dataset, focusing on samples with rind percentages of only 18% and 30%. Considering the 18%-limitation imposed by PDO regulation for rind content in grated cheese, their discrimination could be a major concern for the PR dairy industry because most frauds were potentially concentrated in this range. As shown in Table S8, an error rate of 54.17% was obtained, although a high 68.00% predictive power was reported on test samples. This higher error rate can be attributed to using a smaller and less representative dataset, which aimed to explore a narrow range of rind content levels.

As the next step, the obtained RF models were further investigated to determine the importance of each metabolite on the classification of cheese samples by the MDG index. Based on the previous comparisons, Figs. S5a and S5b show the top 30 compounds reflecting the strongest influence on RF classification models. As can be observed, the significant variability of chemical features obtained by employing these RF classification models highlighted the overall chemical complexity of the rind metabolome. In particular, exposure to environmental conditions mainly leads to these differences during ripening, resulting in decreased moisture content, proteolytic activity, and a higher degree of oxidation (Rocchetti et al., 2021). Looking at our findings, tetraphyllin B, 2-indolecarboxylic acid, 1-metoxyindole-3-carbaladehyde and valine were found as critical metabolites in both models, establishing them as

potential markers for anomalous rind inclusion level in PR grated cheese. Interestingly, among the markers of the restricted model (i.e., 18% vs 30%), four different lysophospholipids were highlighted, namely LysoPE(0:0/24:0), LysoPE(0:0/14:0), LysoPC(16:0) and LysoPC(18:4 (6Z,9Z,12Z,15Z)/0:0). These compounds are typically by-products deriving from phospholipid oxidation and their presence in cheese could be related to the higher rind exposure to air and humidity, compared to the whole inner part of wheel (Subbanagounder, Watson, & Berliner, 2000). Also, as reported by Rocchetti et al. (2021), the phospholipid oxidation within cheese might generate a considerable variety of products, including potential by-products deriving from the cleavage of the oxidized fatty acyl chain. Together with lysophospholipids, 5-hydroxyeicosatetraenoic acid has been linked to phospholipid oxidation, as reported by Reis and Spickett (2012), enhancing the major role of this process in discriminating rind from cheese pulp. Galactitol was also reported as a classifier metabolite for *rind inclusion* (i.e., ≤18% vs >18%). Typically, galactitol in Grana-type cheese could be attributed to certain bacteria using galactose for their metabolism during ageing. In particular, in the early stages of ripening in PR cheese, galactose is completely degraded by thermophilic starter lactic acid bacteria (SLAB) (Lindner et al., 2008). Monfredini, Settanni, Poznanski, Cavazza, and Franciosi (2012) demonstrated that microbial composition and spatial distribution vary throughout the cheese ageing process. Particularly, it has been found that the number of SLAB bacteria in the cheese core was lower compared to those found within the rind portion after a long ageing period. These findings support the galactitol contribution in defining the rind inclusion level in combination with the ripening stage and make this a potential compound related to an anomalous rind inclusion level in grated hard cheese.

Afterwards, applying both RF models on commercial samples exhibited 93.75% and 81.25% success when predicting rind percentage according to NIR estimates on commercial grated cheeses (Table S8). Overall, these findings potentially unravelled a good correlation between the NIR determination and the chemical fingerprint of PR PDO cheese, supporting future exploitation with a more robust dataset with a narrower range of rind inspected (for instance, 18%–30% *w*/w of rind in cheese pulp).

3.3.2. *Random forest classifier on the mountain production of PR cheese*

An RF classifier modelling was also applied to discriminate PR cheeses produced by dairies located in mountain areas. To this aim, PR cheese samples were classified for *altimetric zone* factor parameter, as reported before (i.e., *mountain* vs *lowland*, respectively). Overall, the statistical performances of the RF algorithm are displayed in Table S8. In this scenario, the RF classification proved efficiency in detecting *mountain* cheese according to its UHPLC-Orbitrap metabolomic profile, with only a 4.55% error rate, further supported by a 90.63% predictability when testing the model. Then, the top 30 markers with the highest importance in the classification are depicted in Fig. S5c, according to the MDG index. To better investigate the distinctiveness of potential markers related to the *altimetric zone* parameter on the traceability of mountain PR cheese, a comparative study including the associated VIP markers from the OPLS model (Table S5), VIP^2 markers from the AMOPLS model (Table S6), and the critical compounds found from RF classification (Fig. S5c), was further performed. The obtained Venn diagram (Fig. 3) indicates 17 common markers from these three different approaches and a detailed list of these compounds can be found in Fig. 3. Overall, the majority of these markers were mainly assigned to amino acid and lipid derivatives. The pivotal role of these chemical classes as discriminants for mountain cheeses has already been reported in our previous study, in which the PR cheese profile was found to be modulated by different cow feeding regimes when comparing a diet based on mountain permanent meadows and a more intensive diet based on lowland-alfalfa (Becchi et al., 2023). Looking at our findings, two distinct N-acyl-alpha amino acids, namely pentacosanoylglycine and N2-succinylarginine, were observed. Additionally, we detected a gamma-amino acid, aminobutyric acid-betaxanthin, and an aspartic acid derivative, L-N-(1H-Indol-3-ylacetyl)aspartic acid. Regarding lipid composition, three glycerophospholipids were identified such as PA (16:0/20:1(11Z)), PA(P-16:0/18:2(9Z,12Z)), and SM(d19:0/24:1 (15Z)), alongside with a fatty acyl and a steroid derivative, namely 22-oxo-docosanoate and Cycloartanyl ferulate, respectively. Noteworthy is the presence of the sphingolipid SM(d19:0/24:1(15Z)), previously recognized as a significant marker of milk used for grana-type cheese production associated with a diet based on fresh forage/hay. (Rocchetti, Gallo, Nocetti, Lucini, & Masoero, 2020). Finally, our analysis also revealed two distinct triterpenoids, namely 26-Methyl nigranoate and Feruloyldihydro-beta-sitosterol, suggesting a potential transfer of plant-derived metabolites within the cheese samples.

In this scenario, the potential impact of the mountain feeding system and the increased grazing accessibility of cows in mountain areas could play a crucial role in defining the chemical fingerprint of mountain PR

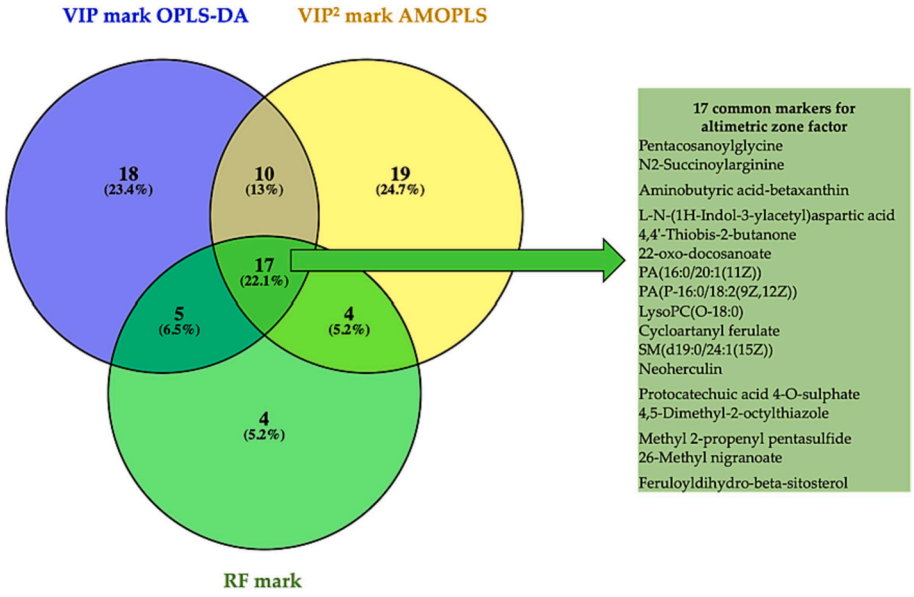


Fig. 3. Venn diagram obtained by combining VIP markers from OPLS model (Table S5) and VIP^2 markers from AMOPLS model (Table S6) and critical compounds found from RF classification (Fig. S5c) considering the *altimetric zone* as discriminant factor.

cheese production. Overall, the results obtained from this RF model are highly promising, enabling a preliminary classification of PR cheeses produced in mountain dairies. This study could also hold significant relevance for the Consortium activities dedicated to safeguarding Grana-type cheese, as it reinforces the excellence of mountain production by promoting the sustainability of these disadvantaged regions compared to the lowland ones and keeping them alive.

4. Conclusions

In this study, untargeted metabolomics was used to discriminate Parmigiano Reggiano PDO samples based on three distinct factors influencing the quality and integrity of this PDO product, namely *ripening*, *altimetric zone*, and *rind inclusion* level in grated hard cheese samples. In this scenario, applying the supervised AMOPLS approach allowed us to consider these experimental factors in a unique model, providing insightful observations among the various sources of variation within our complex dataset. Specifically, this supervised method gave us better results when compared to OPLS-DA, which singularly evaluates the discriminative power for one singular experimental factor at a time. Overall, looking at our findings, the naïve patterns from unsupervised HCA analysis showed that ripening was the variable with the highest hierarchical impact on our data and resulted to be statistically significant when compared to the other studied factors in our AMOPLS model. Regarding the *altimetric zone* factor point of view, comparing PR cheese samples derived from dairies located at different altitude levels provided high-quality results in our supervised models, with a clear separation of PR cheese produced in mountain areas. Furthermore, due to the low-performance OPLS-DA model for *rind inclusion*, random forest (RF) classification models were then performed to classify cheese samples according to the rind inclusion limit defined by the production regulations of Parmigiano Reggiano (not >18%), achieving acceptable accuracy levels on model performances. Additionally, our predictions closely aligned with estimates from NIR measurements on commercial samples, confirming the potential of the NIR technique in predicting rind inclusion levels in grated hard cheese samples. Finally, a specific RF model was further implemented to assess the possibility of correctly classifying dairy farms' altimetry in PR cheese production. Our findings revealed a high-quality prediction performance with a strong capability to differentiate the mountain from lowland PR cheese samples. Taken together, our results highlight the potential of multivariate statistics, multiblock orthogonal partial least squares, and machine learning-based approaches to discriminate three main factors characterizing the quality of PR grated cheese. These findings introduce new potential opportunities for the Consortium's promotional activities related to the protection and sustainability of mountain-based productions. In fact, under real conditions, a set of different interacting factors contribute to quality and authenticity. Despite the large datasets provided by untargeted metabolomics, adequate interpretations are essential to elucidate the role of such interconnected factors. Still, further investigations focusing on a narrower range of rind inspected in grated cheese (i.e., for instance, 18%–30% *w*/w of rind in cheese pulp) appear of great interest for ensuring the authenticity of PR products and mitigating fraudulent practices during cheese grating operations.

CRediT authorship contribution statement

Pier Paolo Becchi: Data curation, Formal analysis, Investigation, Software, Validation, Visualization, Writing – original draft. **Gabriele Rocchetti:** Data curation, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Pascual García-Pérez:** Data curation, Investigation, Validation, Visualization, Writing – review & editing. **Sara Michelini:** Conceptualization, Supervision. **Valentina Pizzamiglio:** Conceptualization, Supervision, Writing – review & editing. **Luigi Lucini:** Methodology, Resources, Supervision, Writing – original draft, 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.

Data availability

Data will be made available on request.

Acknowledgments

P.P.B. was recipient of a Ph.D. Fellowship from the Doctoral School of the Agro-Food System (Università Cattolica del Sacro Cuore, Piacenza, Italy). The authors also thank the "Romeo ed Enrica Invernizzi" Foundation (Milan, Italy) for supporting the metabolomic facility at Università Cattolica del Sacro Cuore and the "Consorzio del Formaggio Parmigiano Reggiano" (Italy) for providing technical advice.

Appendix A. Supplementary data

The supplementary material file is divided into a word file with the supplementary figures and an excel file containing sheets with the different supplementary tables. Regarding the supplementary tables, the excel file containing the following sheets: **S1)** sample legend; **S2)** metabolomics dataset resulting from UHPLC-Orbitrap analysis; **S3)** Model's cross-validation parameters of the three different OPLS-DA models; **S4)** VIP markers related to OPLS-DA *ripening* model; **S5)** VIP markers related to OPLS-DA *altimetric zone* model; **S6)** 50 most discriminant VIP^2 markers related to *altimetric zone* factor of AMOPLS model; **S7)** 50 most discriminant VIP^2 markers related to *ripening x altimetric zone* factor of AMOPLS model; **S8)** Overview of the performance statistical parameters of three RF models under investigation.

References

Abbatangelo, M., Núñez-Carmona, E., Sberveglieri, V., Zappa, D., Comini, E., & Sberveglieri, G. (2018). Application of a novel S3 nanowire gas sensor device in parallel with GC-MS for the identification of rind percentage of grated Parmigiano Reggiano. *Sensors, 18*(5), 1617. https://doi.org/10.3390/S18051617

Alinovi, M., Mucchetti, G., & Tidona, F. (2019). Application of NIR spectroscopy and image analysis for the characterisation of grated Parmigiano-Reggiano cheese. *International Dairy Journal, 92*, 50–58. https://doi.org/10.1016/J.IDAIRYJ.2019.01.010

Arfini, F., Antonioli, F., Cozzi, E., Donati, M., Guareschi, M., Mancini, M. C., & Veneziani, M. (2019). Sustainability, innovation and rural development: The case of Parmigiano-Reggiano PDO. *Sustainability, 11*(18), 4978. https://doi.org/10.3390/SU11184978

Atanasova, J., Dalgalarrondo, M., Iliev, I., Moncheva, P., Todorov, S. D., & Ivanova, I. V. (2021). Formation of free amino acids and bioactive peptides during the ripening of Bulgarian white brined cheeses. *Probiotics and Antimicrobial Proteins, 13*(1), 261–272. https://doi.org/10.1007/S12602-020-09669-0

Becchi, P., Rocchetti, G., Vezzulli, F., Lambri, M., & Lucini, L. (2023). The integrated metabolomics and sensory analyses unravel the peculiarities of mountain grassland-based cheese production: The case of Parmigiano Reggiano PDO. *Food Chemistry, 428*, Article 136803. https://doi.org/10.1016/J.FOODCHEM.2023.136803

Blaženović, I., Kind, T., Sa, M. R., Ji, J., Vaniya, A., Wancewicz, B., … Fiehn, O. (2019). Structure annotation of all mass spectra in untargeted metabolomics. *Analytical Chemistry, 91*(3), 2155–2162. https://doi.org/10.1021/ACS.ANALCHEM.8B04698

Boccard, J., & Rudaz, S. (2016). Exploring omics data from designed experiments using analysis of variance multiblock orthogonal partial least squares. *Analytica Chimica Acta, 920*, 18–28. https://doi.org/10.1016/J.ACA.2016.03.042

Bontempo, L., Lombardi, G., Paoletti, R., Ziller, L., & Camin, F. (2012). H, C, N and O stable isotope characteristics of alpine forage, milk and cheese. *International Dairy Journal, 23*(2), 99–104. https://doi.org/10.1016/J.IDAIRYJ.2011.10.005

Buonaiuto, G., Palmonari, A., Ghiaccio, F., Visentin, G., Cavallini, D., Campidonico, L., … Mammi, L. M. E. (2021). Effects of complete replacement of corn flour with sorghum flour in dairy cows fed Parmigiano Reggiano dry hay-based ration. *Italian Journal of Animal Science, 20*(1), 826–833. https://doi.org/10.1080/1828051X.2021.1916408

Calle, M. L., & Urrea, V. (2011). Letter to the editor: Stability of random Forest importance measures. *Briefings in Bioinformatics, 12*(1), 86–89. https://doi.org/10.1093/BIB/BBQ011

Calvini, R., Michelini, S., Pizzamiglio, V., Foca, G., & Ulrici, A. (2020). Exploring the potential of NIR hyperspectral imaging for automated quantification of rind amount

in grated Parmigiano Reggiano cheese. *Food Control, 112*, Article 107111. https://doi.org/10.1016/J.FOODCONT.2020.107111

Cavallini, N., Strani, L., Becchi, P. P., Pizzamiglio, V., Michelini, S., Savorani, F., … Durante, C. (2023). Tracing the identity of Parmigiano Reggiano "Prodotto di Montagna - Progetto Territorio" cheese using NMR spectroscopy and multivariate data analysis. *Analytica Chimica Acta, 341761*. https://doi.org/10.1016/J.ACA.2023.341761

Evans, A. M., O'Donovan, C., Playdon, M., Beecher, C., Beger, R. D., Bowden, J. A., … Vuckovic, D. (2020). Dissemination and analysis of the quality assurance (QA) and quality control (QC) practices of LC–MS based untargeted metabolomics practitioners. *Metabolomics, 16*(10), 1–16. https://doi.org/10.1007/S11306-020-01728-5

González-Ruiz, V., Pezzatti, J., Roux, A., Stoppini, L., Boccard, J., & Rudaz, S. (2017). Unravelling the effects of multiple experimental factors in metabolomics, analysis of human neural cells with hydrophilic interaction liquid chromatography hyphenated to high resolution mass spectrometry. *Journal of Chromatography A, 1527*, 53–60. https://doi.org/10.1016/J.CHROMA.2017.10.055

Guisset, S., Martin, M., & Govaerts, B. (2019). Comparison of PARAFASCA, AComDim, and AMOPLS approaches in the multivariate GLM modelling of multi-factorial designs. *Chemometrics and Intelligent Laboratory Systems, 184*, 44–63.

Gummalla, S., & Broadbent, J. R. (2001). Tyrosine and phenylalanine catabolism by Lactobacillus cheese flavor adjuncts 1. *Journal of Dairy Science, 84*, 1011–1019. https://doi.org/10.3168/jds.S0022-0302(01)74560-2

Han, H., Guo, X., & Yu, H. (2016). Variable selection using mean decrease accuracy and mean decrease Gini based on random forest. In *, Vol. 0. Proceedings of the IEEE international conference on software engineering and service sciences, ICSESS* (pp. 219–224). https://doi.org/10.1109/ICSESS.2016.7883053

Kokla, M., Virtanen, J., Kolehmainen, M., Paananen, J., & Hanhineva, K. (2019). Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: A comparative study. *BMC Bioinformatics, 20*(1), 1–11. https://doi.org/10.1186/S12859-019-3110-0/FIGURES/5

Lindner, J. D., Bernini, V., De Lorentiis, A., Pecorari, A., Neviani, E., & Gatti, M. (2008). Parmigiano Reggiano cheese: Evolution of cultivable and total lactic microflora and peptidase activities during manufacture and ripening. *Dairy Science & Technology, 88*, 511–523. https://doi.org/10.1051/dst:2008019

Mancini, M. C., Menozzi, D., Donati, M., Biasini, B., Veneziani, M., & Arfini, F. (2019). Producers' and consumers' perception of the sustainability of short food supply chains: The case of Parmigiano Reggiano PDO. *Sustainability, 11*(3), 721. https://doi.org/10.3390/SU11030721

McSweeney, P. L. (2004). Biochemistry of cheese ripening. *International Journal of Dairy Technology, 57*(2–3), 127–144. https://doi.org/10.1111/J.1471-0307.2004.00147.X

Monfredini, L., Settanni, L., Poznanski, E., Cavazza, A., & Franciosi, E. (2012). The spatial distribution of bacteria in grana-cheese during ripening. *Systematic and Applied Microbiology, 35*(1), 54–63. https://doi.org/10.1016/J.SYAPM.2011.07.002

Moran, L., Aldezabal, A., Aldai, N., & Barron, L. J. R. (2019). Terpenoid traceability of commercial sheep cheeses produced in mountain and valley farms: From pasture to mature cheeses. *Food Research International, 126*, Article 108669. https://doi.org/10.1016/J.FOODRES.2019.108669

Ochi, H., Sakai, Y., Koishihara, H., Abe, F., Bamba, T., & Fukusaki, E. (2013). Monitoring the ripening process of Cheddar cheese based on hydrophilic component profiling using gas chromatography-mass spectrometry. *Journal of Dairy Science, 96*(12), 7427–7441. https://doi.org/10.3168/JDS.2013-6897

Reis, A., & Spickett, C. M. (2012). Chemistry of phospholipid oxidation. *Biochimica et Biophysica Acta (BBA) - Biomembranes, 1818*(10), 2374–2387. https://doi.org/10.1016/J.BBAMEM.2012.02.002

Rocchetti, G., Gallo, A., Nocetti, M., Lucini, L., & Masoero, F. (2020). Milk metabolomics based on ultra-high-performance liquid chromatography coupled with quadrupole time-of-flight mass spectrometry to discriminate different cows feeding regimens. *Food Research International, 134*, Article 109279. https://doi.org/10.1016/J.FOODRES.2020.109279

Rocchetti, G., Lucini, L., Gallo, A., Masoero, F., Trevisan, M., & Giuberti, G. (2018). Untargeted metabolomics reveals differences in chemical fingerprints between PDO and non-PDO Grana Padano cheeses. *Food Research International, 113*, 407–413. https://doi.org/10.1016/J.FOODRES.2018.07.029

Rocchetti, G., Michelini, S., Pizzamiglio, V., Masoero, F., & Lucini, L. (2021). A combined metabolomics and peptidomics approach to discriminate anomalous rind inclusion levels in Parmigiano Reggiano PDO grated hard cheese from different ripening stages. *Food Research International, 149*, Article 110654. https://doi.org/10.1016/J.FOODRES.2021.110654

Santarcangelo, C., Baldi, A., Ciampaglia, R., Dacrema, M., Di Minno, A., Pizzamiglio, V., … Daglia, M. (2022). Long-aged Parmigiano Reggiano PDO: Trace element determination targeted to health. *Foods, 11*(2), 172. https://doi.org/10.3390/FOODS11020172/S1

Segato, S., Galaverna, G., Contiero, B., Berzaghi, P., Caligiani, A., Marseglia, A., & Cozzi, G. (2017). Identification of lipid biomarkers to discriminate between the different production systems for Asiago PDO cheese. *Journal of Agricultural and Food Chemistry, 65*(45), 9887–9892. https://doi.org/10.1021/ACS.JAFC.7B03629

Sforza, S., Cavatorta, V., Lambertini, F., Galaverna, G., Dossena, A., & Marchelli, R. (2012). Cheese peptidomics: A detailed study on the evolution of the oligopeptide fraction in Parmigiano-Reggiano cheese from curd to 24 months of aging. *Journal of Dairy Science, 95*(7), 3514–3526. https://doi.org/10.3168/JDS.2011-5046

Singh, T. K., Drake, M. A., & Cadwallader, K. R. (2003). Flavor of Cheddar cheese: A chemical and sensory perspective. *Comprehensive Reviews in Food Science and Food Safety, 2*(4), 166–189. https://doi.org/10.1111/J.1541-4337.2003.TB00021.X

Subbanagounder, G., Watson, A. D., & Berliner, J. A. (2000). Bioactive products of phospholipid oxidation: Isolation, identification, measurement and activities. *Free Radical Biology and Medicine, 28*(12), 1751–1761. https://doi.org/10.1016/S0891-5849(00)00233-1

Thierry, A., Collins, Y. F., Abeijón Mukdsi, M. C., McSweeney, P. L. H., Wilkinson, M. G., & Spinnler, H. E. (2017). Lipolysis and metabolism of fatty acids in cheese. *Cheese: Chemistry, Physics and Microbiology, 1*, 423–444. https://doi.org/10.1016/B978-0-12-417012-4.00017-X. Fourth Edition.

Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., … Arita, M. (2015). MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature Methods, 12*(6), 523–526. https://doi.org/10.1038/nmeth.3393

Tsuzuki, W., Komba, S., Kotake-Nara, E., Aoyagi, M., Mogushi, H., Kawahara, S., & Horigane, A. (2018). The unique compositions of steryl ferulates in foxtail millet, barnyard millet and naked barley. *Journal of Cereal Science, 81*, 153–160. https://doi.org/10.1016/J.JCS.2018.04.006

Veerasamy, R., Rajak, H., Jain, A., Sivadasan, S., Varghese, C. P., & Agrawal, R. K. (2011). Validation of QSAR models—Strategies and importance. *International Journal of Drug Design and Discovery, 2*, 511–519.

Zannoni, M. (2010). Evolution of the sensory characteristics of Parmigiano–Reggiano cheese to the present day. *Food Quality and Preference, 21*(8), 901–905. https://doi.org/10.1016/J.FOODQUAL.2010.01.004

Zannoni, M., & Hunter, E. A. (2015). Relationship between sensory results and compliance scores in grated Parmigiano-Reggiano cheese. *Italian Journal of Food Science, 27*(4), 487–494. https://doi.org/10.14674/1120-1770/IJFS.V41