THESIS

ABILITY, REPEATABILITY, AND REPRODUCIBILITY OF RAPID EVAPORATIVE IONIZATION MASS SPECTROMETRY TO PREDICT BEEF QUALITY ATTRIBUTES

Submitted by

Michael J. Hernandez Sintharakao

Department of Animal Sciences

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Fall 2022

Master's Committee:

Advisor: Mahesh N. Nair
Co-Advisor: Jessica E. Prenni

James B. Morgan
Julia L. Sharp

ABSTRACT

ABILITY, REPEATABILITY, AND REPRODUCIBILITY OF RAPID EVAPORATIVE IONIZATION MASS SPECTROMETRY TO PREDICT BEEF QUALITY ATTRIBUTES

Tenderness, juiciness, and flavor are beef quality attributes that influence consumer satisfaction eating beef. Rapid evaporative ionization mass spectrometry (REIMS) is a novel technique that provides chemical information of biological tissues with the potential to predict beef quality attributes. Two studies were conducted to evaluate the ability of REIMS to predict quality attributes of beef (study I) and to evaluate the repeatability and reproducibility of REIMS in a beef matrix (study II).

In study I, USDA Select or upper two-thirds Choice (n = 42, N=84) striploins and tenderloins were collected approximately 36h post-mortem from a commercial beef abattoir. Slivers of the *longissimus dorsi* muscle between the 12-13th rib were collected during grading (GR, 36h post-mortem) and analyzed using REIMS. Striploins (LM) and tenderloins (PM) were cut into portions and assigned to 6 aging periods (3, 14, 28, 42, 56, and 70 days). However, only samples aged 3, 14, and 28 days were used to represent industry practices in study I. After aging, portions were cut into 2.54-cm steaks to analyze juiciness, tenderness, and 10 flavor attributes with a trained sensory panel. In addition, tenderness measures were performed using slice shear force (SSF) and Warner-Bratzler shear force (WBF). Samples were categorized by SSF, WBF, and sensory panel tenderness (PT) into “tough” and “tender”; by juiciness into “dry” and “juicy”; and by flavor into “acceptable” and “unacceptable” classes using a composite score of all flavor descriptors. Combinations of three dimensionality reduction methods (principal component analysis [PCA],

feature selection, [FS], and a combination of both [PCA-FS]) with 13 machine learning algorithms were used to create classification models based on REIMS data for tenderness, juiciness, and flavor classes at the three aging periods. The predictive ability of the models was assessed with the overall accuracy resulting from 10-fold cross-validation. Among all machine learning algorithms evaluated, the maximum classification accuracies for days 3, 14, and 28 were 94, 87, and 83% for PT; 86, 85, 92% for SSF; 87, 82, and 95 for WBF; 85, 84, and 86% for juiciness; and 87, 89, and 81% for flavor classes, respectively. FS performed the best as a dimensionality reduction method for all PT, juiciness, flavor, and SSF on day 3 and WBF on days 3 and 14. PCA-FS was the best dimensionality reduction method for SSF on days 14 and 28, and WBF on day 28. Extreme gradient boosting machine learning algorithm was the highest performing algorithm for all juiciness models, flavor model on day 28, PT on days 3 and 14, SSF on days 14 and 28, and WBF on days 3. Partial least squared discriminant analysis (PLSDA) performed better for PT day 28 and flavor day 14, while elastic-net regularized generalized linear model, random forest, and support vector machine were the highest performing algorithms for SSF day 3, and WBF days 14 and 28, respectively. Results demonstrated that the chemical fingerprints obtained with REIMS could potentially be used as in situ and real-time technique to sort carcasses by flavor, juiciness, and tenderness. However, overlaps between classes affected REIMS results, and unbalanced data negatively affected model accuracies. Therefore, exploring the full potential of REIMS will require increasing the sample size and developing a sampling method that allows increased separation between sensory evaluations.

Study II was performed with REIMS data from all LM and PM samples from the six aging periods (n=1008), two sets of GR samples (n=168, N=84), and quality control (QC) samples (n=29) made from homogenized ground beef. Except for the second set of GR samples, REIMS

analysis of all samples was performed at Colorado State University (CSU) using a meat probe as the sampling device. Analysis of all samples was performed over 5 days, including two batches per day. GR samples were evaluated on the first day, and LM and PM data were randomly analyzed on the remaining days. QC samples were analyzed at the beginning, middle, and end of each batch. The second set of GR samples was analyzed at Texas Tech University (TTU) using different mass spectrometry (MS) instruments, technicians, and an iKnife as the sampling device. The stability of REIMS data between burns, batches, and days was evaluated with QC data. Day effect and robustness of REIMS data were evaluated with data from LM and PM samples, and interlab reproducibility was evaluated with data from GR samples. Multiple classification models of muscle type and aging were built with LM and PM data to evaluate the robustness of REIMS and day-to-day variability. Models to predict sensory attributes of beef were used to assess the robustness of REIMS with respect to interlab variability. Coefficients of variation (CV) between burns of the mass bins representing 90% of the total ion current were between 0.7 to 0.98, while the most relevant mass bins showed CV less than 0.3. Variances between batches and collection days were not significant ($P < 0.05$). PCA of LM and PM showed that data variability by collection day was stronger than muscle type and aging time variability. However, data could classify samples into muscle types and two distant aging times with accuracies higher than 95.6% and 91.0%, respectively. PCA of GR samples showed that data collected in both labs differed, and the predictive models developed with the CSU data did not appropriately predict the quality classes with the TTU data. REIMS collected with the meat probe provides a chemometric profile of beef samples with good repeatability and interday reproducibility but low interlab reproducibility. Consequently, optimization and standardization of sampling methods will be required to improve the interlab reproducibility of REIMS.

ACKNOWLEDGEMENTS

I am deeply grateful to my advisor and co-advisor, Dr. Nair Mahesh and Dr. Jessica Prenni, for allowing me to be a part of this program, for their guidance, and for contributing to my academic and professional development. I would also express my gratitude to Dr. Morgan Brad and Dr. Sharp Julia for their contribution as committee members and Dr. Woerner Dale, Chandler Sarchet, and Chaoyu Zhai for their assistance and contribution in developing this project. I am also thankful to Dr. Georgnaras Gina and Dr. Delmore Bob for their patience and help and for allowing me to learn from their experience and knowledge in the microbiology and meat lab.

Lastly, I thank my partner, family, and friends. I could not have undertaken this journey without my partner Sara and her unconditional support and patience. I am grateful to my family for their support and always pushing me to work harder and pursue my passion. I want to thank my friends Chaoyu, Colton, Caleb, Tyler, Emily and Aeriel, who made this journey more pleasant and funny.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: REVIEW OF LITERATURE

The United States is the largest beef consumer and producer worldwide. In 2020, the US produced 27.6 million pounds of beef and exported 10.9% of its production, representing 83.5 million dollars of the local market and 6.8 million dollars in exports (USDA, 2021). It is estimated that the annual consumption of beef is 58 pounds per capita, making it the second most consumed animal protein in the US (USDA, 2021). One of the reasons for the success of the beef industry has been the development and implementation of the USDA grading system.

USDA beef grading system

The USDA beef grading is a voluntary marketing program developed and maintained by USDA's Agricultural Marketing Service (AMS). This program includes the USDA yield grade, which estimates the percentage of saleable beef in the carcasses, and the USDA quality grade, which seeks to predict beef palatability of carcasses (USDA, 2019). Even when the USDA grading is not mandatory, around 96% of steers and heifers slaughtered in commercial facilities were graded in 2017 (AMS, 2018a). The USDA grading system has been a successful program in the beef industry because it facilitates commercial transactions between producers and farmers, and consumers recognize it as a symbol of safe and high quality (USDA, 2019). The grading system has made it possible to establish a grid price for carcasses that incentivize producers to increase beef quality and cattle practices.

Cutability refers to the amount of edible meat from a carcass. The USDA yield grading predicts beef cutability by classifying the carcasses into grades 1, 2, 3, 4, and 5. Yield grade 1 represents carcasses with the highest yield of saleable beef and grade 5 with the lowest yield. Cutability is estimated with a mathematical model that includes the hot carcass weight, an estimate

of the percentage of kidney, pelvic, and heart fat, external fat thickness over the ribeye, and the ribeye area. Fat thickness and ribeye are measured in a transversal cut of the ribeye muscle between the 12^{th} and 13^{th} rib (AMS, 2018b).

The USDA quality grade attempts to predict palatability (tenderness, flavor, and juiciness) of carcasses based on marbling score and animal maturity. From the highest expected palatability to the lowest, the USDA quality grades are Prime, Choice, Select, Standard, Commercial, Utility, Cutter, and Canner (AMS, 2018b). The final grade results from a combination of marbling score and maturity. The marbling score represents the amount of intramuscular fat found in the ribeye muscle. This score is assessed with Video Imagen Analysis Systems (VIA) or visual appraisal of the lean ribeye. Animal maturity includes classes A for young animals (under 30 months) and B, C, and D for animals over 30 months. Maturity A is established by dentition or age documentation, and for older carcasses, maturity is determined by skeletal and muscular traits (USDA, 2017). Only carcasses with maturity A or B can be classified as Prime, Choice, and Select, which are the grades most consumers recognize and consider for food-grade labels. Consequently, most carcasses with maturity C or older, less than 1.8% of carcasses harvested in commercial facilities, are sold ungraded (Boykin et al., 2017; Ward, 2021).

Quality grades segregate carcasses by the probability of having a positive eating experience, but maturity and marbling scores do not explain great variability in palatability traits (Emerson et al., 2013; Acheson et al., 2014; Gredell, 2018). O'Quinn et al. (2018) reported that nearly 91%, 87%, 83%, and 75% of consumers that eat Prime, top Choice, low Choice, and Select steaks have a positive eating experience, respectively. However, according to Acheson et al. (2014) and Semler et al. (2016), there are no differences in palatability scores of striploin from maturity-A carcasses compared to B-C maturity. Literature has reported that instrumental marbling score

explains between 24-45% variability in juiciness, 27-40% variability in tenderness, and 34-61% overall palatability of longissimus muscle of A-maturity carcasses (Smith et al., 1985; Emerson et al., 2013).

Beef tenderness

Tenderness is undoubtedly an important attribute that influences consumer satisfaction when eating beef. The odds of having a negative eating experience when tenderness is unacceptable is more than twice the odds when this attribute is acceptable (O'Quinn et al., 2018). Additionally, 80% of consumers would be willing to pay a premium for the guarantee of having a tender steak (Miller et al., 2001). Although the current USDA quality grade program seeks to predict palatability traits, the criteria for quality grading do not explain great variability in tenderness. There are no differences in tenderness in carcasses with maturity A-C, and studies have demonstrated that marbling score explains between 40-5% of the variability in tenderness (Wheeler et al., 1994; Platter et al., 2003; Emerson et al., 2013; Acheson et al., 2014; Semler et al., 2016).

Tenderness depends on muscle metabolic and structural characteristics, including sarcomere length, proteolysis of proteins, marbling, amount of connective tissue, and the cross-links within collagen molecules (Webb et al., 1964; Purchas, 2014). Sarcomeres are structural units of the muscle fibers responsible for contraction. During the conversion of muscle to meat, a series of biochemical causes an overlap of myofibrillar proteins and, therefore, sarcomere shortening. A greater degree of shortening of the sarcomere is associated with tough meat (Smulders et al., 1990). Pre- and post-harvest factors such as storage temperature or animal stress influence the degree of shortening. Tenderness increases with aging time due to proteolysis of proteins. Endogenous proteolytic enzymes break down muscle structures, increasing meat tenderness (Casas et al., 2006).

The extent and rate of proteolysis are influenced by breed, temperature, aging time, and electrical stimulation (Casas et al., 2006; Bhat et al., 2018). Meat with a higher concentration of connective tissue tends to be tougher, and cooking conditions greatly impact the final tenderness of the meat because it affects the solubility of the connective tissue (Purchas, 2014). Higher intramuscular fat has been related to higher tenderness. Although the explanations of how fat affects tenderness are not clear, some hypotheses establish that fat acts as a lubricant during chewing, influencing the perception of tenderness and juiciness; connective tissues are diluted with fat during cooking; or that meat with a higher degree of marbling has less probability of shortening before rigor (Ellis et al., 1998; Nishimura et al., 1999; Purchas, 2014).

Since tenderness is affected by multiple factors, including cooking conditions, common methods to measure tenderness are based on the biting force on cooked products (Purchas, 2014). Warner-Bratzler Shear Force (WBF) and Slice Shear Force (SSF) are mechanical methods for measuring the force required to cut cooked muscle fibers perpendicularly. The main difference between the two assays is that WBF measures force from cutting cylindrical 0.5-in-diameter cores and SSF from a 1x5cm slice (Wheeler et al., 2005). WBF allows measuring shear force from different muscles and animal species since only small cores are required, but it is time-consuming. On the other hand, SSF was originally developed to measure tenderness in recently cooked beef longissimus steaks, and its application in another type of muscle or smaller animal species is limited because it requires a sizeable steak (Shackelford et al., 1999).

SSF is a simpler, faster, more repeatable, and more accurate protocol to measure beef tenderness than WBF. The coring process in WBF is done after the steaks are cooked and cooled for a firmer texture and to make core extraction easier. While in SSF, slices are cut almost immediately after cooking for better results (Wheeler et al., 2005). The SSF has a stronger

correlation (r=-0.82) with the tenderness perceived by trained consumers than WBF (r=-0.77, Shackelford et al., 1999). Regarding repeatability, SSF explains 89-91% of the variance in longissimus tenderness between animals, while WBF explains 53-87% (Wheeler et al., 1996; Wheeler et al., 1997; Shackelford et al., 1999).

Non-destructive methods to measure tenderness from intact carcasses have been tested, but none have been implemented so far. The Tendertec Mark III Beef Grading Probe, an instrument developed to estimate steaks tenderness from beef carcasses, predicted WBF of steaks from old animals cooked at 60 to 65°C (Belk et al., 2001). However, this instrument could not predict the tenderness of steaks with a higher degree of doneness and from young animals. Other approaches have used video image analysis systems (VIA) to predict beef tenderness with muscle structure information (Vote et al., 2003; Wyle et al., 2003; Howard et al., 2010). Vote et al. (2003) evaluated the Computer Vision System equipped with a BeefCam, a VIA, to classify carcass into a "tender" category from lean color (L*, a* and b* values). Although they got 80% classification accuracy, this value was highly despaired by the marbling score. Howard et al. had better results using a new generation of VIA (Tenera Technology HighResolution Imaging system) that was able to classify tender steaks with 94.7% accuracy.

Spectroscopy and hyperspectral imaging as tenderness predictors has also been evaluated. Spectroscopy provides biochemical information that can be used to predict beef palatability traits. Research has found that outputs from Near-Infrared Spectrometry can explain between 53-68% variance in tenderness measured mechanically (Mitsumoto et al., 1991; Rosenvold et al., 2009; Yancey et al., 2010). Hyperspectral imaging combines biochemical information with a high spatial resolution (Konda Naganathan et al., 2016). Since tenderness is affected by the biological structure of the muscle and its biochemistry, it was expected that hyperspectral imaging has a better ability

to predict tenderness than the other technologies. Predictive models using hyperspectral imaging yield between 77.0-86.7% accuracy in classifying carcasses on tenderness categories based on mechanical shear force measures (Konda Naganathan et al., 2008; Konda Naganathan et al., 2016).

Although tenderness is an important factor for consumer satisfaction, the only marketing program for tenderness in the United States is not popular among producers. The USDA Certified Tender or Very Tender is a voluntary program of the USDA's Agricultural Marketing Service (AMS). Producers can include the claims “certified tender” and “certified very tender” to different products (e.g., beef loin, tenderloin, butt, defatted) if they validate tenderness WBF or SSF from the *longissimus dorsi* muscle of the carcass using a third-party auditor. The "certified tender” claim only applies to products from carcasses with WBF<4.4kg or SSF<20.0kg, and “certified very tender” applies when WBF<3.9kg or SSF<15.4kg (ASTM, 2011). Including these claims could bring economic benefits to the industry since consumers are willing to pay a premium for a tender guarantee (Miller et al., 2001). However, this program is not popular among producers due to the high cost since it requires the waste of 1-in-thick steaks (Wheeler et al., 2002).

Considering that less than 5% of the carcasses in the United States are considered tough (Martinez et al., 2017), predicting tenderness represents an opportunity more than a problem. Premium beef programs could benefit from excluding carcasses that do not meet tenderness requirements for an acceptable eating experience, increasing reputation and consumer trust in the brands. Carcasses with low marbling scores receive discounts because they are expected to have low palatability (Gredell, 2018). However, many carcasses with low marbling scores are tender and have positive sensory performance (Emerson et al., 2013). Therefore, methods that could predict tenderness at a low cost and in real-time represent an opportunity to improve beef marketability.

Flavor

Numerous studies have defined beef tenderness as the most important factor for consumer satisfaction (Savell et al., 1987; Miller et al., 1995; Miller et al., 2001). However, when tenderness is acceptable, flavor is considered the main factor of beef palatability (Goodson et al., 2002; Behrends et al., 2005; Legako et al., 2015). O'Quinn et al. (2018) reported that flavors account for 49.4% of overall palatability, while tenderness and juiciness accounted for 43.4 and 7.4%. Additionally, since more than 95% of the beef carcasses in the United States are considered tender, flavor may be the main driver of consumer satisfaction (Hunt, 2013; Martinez et al., 2017).

Flavor perception results from taste, smell, and touch sensations (Kerth and Miller, 2015). Gustatory sensory cells in our tongue perceive the basic tastes, including salty, sweet, sour, bitter, and umami. Stimulation of tactile receptors from texture, viscosity, and the astringency of the food is part of oral-somatosensation or what we know as "mouth feel" (Spence and Piqueras-Fiszman, 2016). The olfactory system perceives aromas resulting from the stimulation of olfactory bulbs with volatile components (Haggard and de Boer, 2014; Legako et al., 2016).

Beef flavor development occurs through multiple chemical reactions during cooking. Raw meat does not have the meaty flavor we normally find in cooked meat. Instead, raw meat tastes salty, metallic, and bloody, with an aroma similar to serum (Wasserman, 1972). Non-volatile compounds in raw meat, such as amino acids, peptides, vitamins, nucleotides, and reducing sugars, are taste-active compounds and precursors of cooked meat flavor (Legako et al., 2015). During cooking, heat induces lipid degradation, Maillard reaction, Strecker degradation, and reactions between the Maillard reaction and lipid degradation products (Kerth and Miller, 2015; Legako et al., 2016). Volatile compounds products of those reactions determine the aroma and contribute to the characteristic flavor of meat (Shahidi, 1994).

The sweet flavor of beef is associated with sugars (glucose, fructose, and ribose) and some L-amino acids (Macleod, 1994). Sourness is due to some organic acids (lactic, aspartic, and succinic acid), amino acids (glutamic acid, histidine, and asparagine), ortho-phosphoric, and pyrrolidone carboxylic acids. Inorganic salts and sodium salts of glutamate and aspartate are associated with saltiness. Bitterness is caused by hypoxanthine, certain peptides (e.g., anserine and carnosine), and various L-amino acids, while the umami taste of beef comes from 5' -guanosine monophosphate (GMP), 5'-inosine monophosphate (IMP), glutamic acid, monosodium glutamate, and some peptides (Macleod, 1994).

Around 1000 volatile compounds have been identified in beef, but only a portion has been identified as aroma-active (Macleod, 1994). Volatiles can be classified into chemical groups, such as n-aldehydes, Strecker aldehydes, ketones, sulfur-containing compounds, and alcohols, to mention some of them. N-aldehydes are the most prominent group resulting from thermal lipid degradation and usually are negatively associated with flavor liking (Legako et al., 2016). Strecker aldehydes are aroma-active compounds associated with beef flavor that result from the degradation of α-amino acids in the presence of α-dicarbonyl compounds. For instance, 3-methylbutanal and phenylacetaldehyde are derivates from leucine and phenylalanine with a malty and honey-like aroma, respectively (Hofmann and Schieberle, 2000). In the ketones group, 2,3-butanedione and 3-hydroxy-2-butanone are associated with a buttery flavor (Raes et al., 2003; Machiels et al., 2004). Foraker et al. (2020) stated that only 2,3-butanediol, 2-methylbutanal, 3-methylbutanal, acetic acid, ethanol, phenylacetaldehyde, and toluene are strongly correlated ($r \geq |0.50|$) with some beef flavor attributes (beef flavor identity, sour, oxidized, nutty, musty, or liver-like attributes).

Meat flavor is typically evaluated with sensory panels that include trained panelists or general consumers. In sensory panels, panelists are asked to evaluate sensory differences between

multiple samples or palatability quality traits (tenderness, juiciness, and flavor) of cooked samples (Wheeler et al., 2015). Consumer panels involve regular consumers of meat products who are usually selected by demographic information and consumer habits. However, this panel type is highly variable because it depends on consumer preferences (Di Rosa et al., 2017). Trained panels seek to reduce subjectivity by training panelists to detect and evaluate specific flavor traits (Kerth and Miller, 2015). Whether trained panelists or consumers are used, both methods rely on the human sense of taste and consequently are not objective (Di Rosa et al., 2017).

Some studies have evaluated electronic tongues and noses as objective methods for flavor measure. An electronic tongue contains an artificial lipid membrane that detects tastes via electrostatic and hydrophobic interactions with food. Zhang et al. (2015) demonstrated that an electronic tongue could differentiate tastes, somatosensations, and proximate composition of beef from Wagyu, Angus, and Simmental breeds cattle. Ismail et al. (2020) characterized non-volatile compounds of beef cooked with two different sous-vide methods using an electronic tongue. Electronic tongues can only detect tastes and somatosensations but cannot detect smells that are an essential part of flavor. Electronic noses follow similar principles to the electronic tongue to detect smells. Kodogiannis (2017) and Mohammed et al. (2021) demonstrated that an electronic nose could accurately differentiate between spoilage and fresh beef samples, but there are no current applications to evaluate beef smell attributes. Although the combination of electronic tongue and nose could enhance the ability to assess beef flavor, the author has not found studies that combine both technologies.

Juiciness

It is estimated that juiciness contributes 7.4% to the overall palatability of eating beef steaks (O'Quinn et al., 2018). A bad score in juiciness can be compensated with high scores in tenderness

and flavor (Liu et al., 2020). However, the probability of having a negative eating experience when juiciness is unacceptable is 66% (O'Quinn et al., 2018). Moreover, juiciness highly impacts meat texture and influences tenderness perception (Juárez et al., 2012). Initially, juiciness was defined as the moisture from food released in the mouth when chewing and moisture from saliva (Christensen, 1984). Later studies found that physiological and psychological factors also influenced juiciness. This theory brought to the table dividing juiciness into initial and sustained juiciness, but later both concepts were combined as one factor (Juárez et al., 2012).

Antemortem animal stress can cause two quality defects of meat that affect juiciness called dark, firm, and dry (DFD) and pale, soft, and exudative (PSE). Water holding capacity (WHC) represents the ability to retain water and is usually used as an indicator of juiciness. This property is impacted by meat pH. The farter the pH is from the isoelectric point of meat, the higher the WHC of meat (Van Oeckel et al., 1999). Muscle of animals undergoing long-term stress contains low glycogen levels, resulting in a high muscle pH post-mortem. High meat pH produces DFD characterized by dark color, firm texture, and high WHC. On the other side, acute stress before the slaughter combined with a low chilling rate produces low meat pH with pale colors, soft textures, and low WHC meat (PSE). In beef, PSE is only observed in deep muscles from the hip (Juárez et al., 2012). Authors have found that DFD meat is usually juicier because it has a higher WHC (Grayson et al., 2016). However, long-term stress can also lead to animal dehydration, impacting juiciness negatively (Juárez et al., 2012).

Process factors such as mechanical tenderization, seasoning, freezing, and cooking affect meat juiciness. Blade tenderization reduces juiciness because of a disruption of meat structure that leads to a reduced WHC. Similarly, crystal ice formed during freezing damages cell membranes, reducing WHC and juiciness (Wheeler et al., 1990). Studies have demonstrated that injection of

salt solutions increases juiciness because salt solutions enhance WHC (Juárez et al., 2012). During cooking, proteins are denatured, decreasing WHC and increasing water loss by evaporation. Therefore, the degree of doneness and cooking rate greatly influence juiciness (Savell et al., 1999).

Rapid Evaporative Ionization Mass Spectrometry (REIMS)

Rapid Evaporative Ionization Mass Spectrometry (REIMS) is a novel ambient mass spectrometry that can generate a mass spectra fingerprint of biological tissues in real-time, in situ, and without tedious sample preparation (Balog et al., 2010). This technique combines two key elements: a vaporization sampling device and a time-of-flight (TOF) mass spectrometer (Barlow et al., 2021). The purpose of the sampling device is to produce high temperatures on the surface of the sample (also called "burns") and generate an aerosol rich in chemical components (Ross et al., 2021). Then, the produced aerosol is sucked through a flexible tube by a venturi pump to the mass spectrometry, where the vapor is ionized and transformed into mass spectra (Barlow et al., 2021). The generated mass spectra can be used as a metabolic fingerprint that can be analyzed to assess key attributes of the samples. Current research has demonstrated the ability of REIMS to identify animal tissues from different anatomical origins, breeds, and species with more than 97% accuracy (Balog et al., 2016); differentiate between beef from two different processing systems (straight dry-aged and stepwise dry-age, Zhang et al., 2021); discriminate between tainted, and untainted boars with 100% accuracy (Verplanken et al., 2017); classify beef samples into two tenderness levels with 90% accuracy (Gredell et al., 2019).

First applications and most of the studies of REIMS have been carried out using an electrosurgical tool best known as iKnife (Schäfer et al., 2009; Balog et al., 2013). The iKnife, a term for "intelligent knife," was developed to detect cancerous tissues during surgical interventions (Balog et al., 2013). This device consists of a cutting electrode blade and a return electrode pad

connected to an electrosurgical generator (St John et al., 2017). When both electrodes are in contact with the sample, a circuit is formed, and a high-frequency alternating current heat the end of the blade in contact with the sample (Barlow et al., 2021). Other studies have evaluated CO2 laser, bipolar food probes, soldering iron, and bipolar forceps as sampling devices to improve ionization or find a more suitable sampling technique for different applications (Strittmatter et al., 2014; Ross et al., 2021). Since the iKnife requires that a return electrode is in contact with the sample, its use is not appropriate for the meat industry because it can lead to cross-contamination. Therefore, other more suitable sampling techniques should be evaluated for meat applications.

Comparison of REIMS with traditional mass-spectrometry techniques

Traditional mass spectrometry techniques (MS), such as liquid chromatography/mass spectrometry (LC-MS), liquid chromatography/tandem mass spectrometry (LC-MS/MS), gas chromatography/mass spectrometry (GC–MS), have been used for food analysis to identify and quantify chemical components with high sensitivity, accuracy, and reproducibility (Careri et al., 2002; Malik et al., 2010). Although conventional mass spectrometry methods offer great advantages in food analysis, their application in the food and meat industry is limited due to high costs, time, and the technical expertise required (Birse et al., 2021).

REIMS is an attractive alternative that can provide valuable chemical information in real-time, lower cost, and without sample preparation that conventional MS techniques require. Because REIMS does not require sample preparation, data acquisition takes just a few seconds. The rapid acquisition of REIMS data could follow the pace of production lines, allowing to use it for in-situ measurements in control points (Verplanken et al., 2017). Although recent REIMS studies use high-resolution mass spectrometers (Ross et al., 2021), the costs of implementing this technology could be reduced by acquiring lower resolution equipment. The REIMS analysis is

used as a chemical fingerprint, and the data is binned in intervals of 0.1 to 1.0 m/z making high-resolution equipment unnecessary. Furthermore, Strittmatter et al. (2014) found that varying the bin size from 0.1 to 5.0 Da does not impair REIMS ability to predict species, genus, and gram levels in microorganisms.

Some of the limitations of REIMS compared to conventional MS are that it cannot detect low abundance or high-molecular-weight molecules (Ross et al., 2021). REIMS does not include a chromatography separation step compared to conventional GC-MS and LC-MS. In REIMS, all the compounds reach the mass spectrometer almost simultaneously, which could cause ion suppression of low abundance molecules due to the effect of the sample matrix (Volmer and Jessome, 2006). Verplanken et al. (2017) observed this phenomenon when they could not detect traces of components responsible for boar taint (indole, skatole, and androstenone) from neck skin samples using REIMS. The optimal mass bins for REIMS reported by Birse et al. (2021) were between 600-900 m/z for chicken models, while other studies have used ranges between 100 to 1500 m/z. Since the compounds need to be in aerosol to be detected by REIMS, only molecules with molecular weights around 1000 and 1500 Da can be analyzed by REIMS (Ross et al., 2021).

Compared to conventional MS protocols, REIMS is not an appropriate method to quantify chemical components. Since the spray generated during sampling is highly variable, a standard infused in the sample is required to build a calibration curve for quantification. Nevertheless, there is no straightforward method to infuse standards in solid samples that can work for quantification (Barlow et al., 2021; Ross et al., 2021). Moreover, the confidence for compound identification of REIMS is low since it only includes MS data and not MS/MS (Ross et al., 2021). Instead, REIMS can segregate or classify samples based on chemometric profiling. The metabolic fingerprint generated by REIMS contains mass-to-charge ratios (m/z) of all components from the sample that

reaches the MS. The resulting data is a rich source of chemical information that can be used to build statistical models for biological classification (Balog et al., 2013).

REIMS data analysis

Analysis of REIMS data usually involves four main steps: data pre-processing, dimensionality reduction, building predictive models, and model verification. Data pre-processing is done with specialized software that allows peak selection (burns), data calibration, and transformations to make it easier to work with the data. Dimensionality reduction involves deleting redundant features that increase noise or make models inefficient. This step can also be performed using transformations that optimize data representation in a low number of features (van der Maaten et al., 2009). Then, the reduced data is used to train predictive models with machine learning algorithms. The last term refers to a combination of statistical tools and computational algorithms that allow the generation and improvement of models through experience with the data. Algorithms that include a response variable are called supervised learning. In contrast, unsupervised learning does not include labels and seeks to find patterns in the data (Gareth et al., 2021a). As the main purpose is using REIMS to build predictive models, supervised machine learning algorithms are used to build models; however, unsupervised algorithms are usually used to reduce dimensionality or analyze the data to find patterns (Gareth et al., 2021b). Finally, the performance of the models is evaluated using cross-validation or calculating statistical indicators or performance.

Data pre-processing includes lock mass correction using a standard, background subtraction, and peak normalization based on total ion current (Jones et al., 2019). Depending on the type of analysis, exact m/z are used, or data is binned in intervals (e.g., 0.1, 0.5, 1.0 m/z bins). Some authors have analyzed exact mass as preliminary identification of specific compounds in

animal tissues and microorganisms (Strittmatter et al., 2014; Balog et al., 2016; Genangeli et al., 2019; He et al., 2021). However, identification of compounds using REIMS is only tentative because of the lack of procedure for feature correspondence and quantification (Ross et al., 2021). REIMS data is more suitable for untargeted analysis that uses chemometric profiling for sample classification. Since the chemometric profile analysis does not require exact m/z, the mass spectra are usually binned.

After binning, the size of the data obtained is still considerably large to build properly predictive models (e.g., 600, 1000, 1700 mass bins). Therefore, dimension reduction is recommended to reduce computational costs and the chance of overfitting the model (Kettaneh et al., 2005). Dimension reduction consists of applying transformations or deleting unnecessary features to reduce redundancy and increase the significance of the data in fewer dimensions (van der Maaten et al., 2009). Common dimension reduction methods used for REIMS are Principal Component Analysis (PCA), Partial Least Squares (PLS), and feature selection (FS). PCA is a mathematical algorithm that transforms the data by projecting it into new directions (principal components, PC) that maximize the data variation in smaller dimensions (Ringnér, 2008). The benefits of using PCA are that it reduces noise and collinearity while explaining variability into a smaller number of features (Kemsley, 1996). In addition, this tool can be used as unsupervised machine learning for cluster identification. PLS is a supervised machine learning algorithm that projects the data into new hyperplanes that optimize linearity between the data and the response variable (Gareth et al., 2021b). Feature selection (FS) consists of identifying a subset of features that better explains the observations (Hauskrecht et al., 2007). This process can involve multiple steps and criteria of selection. Some criteria used in metabolomics analysis are selecting the features with lower variability within samples or excluding highly-correlated data (Ross et al.,

2021). Other methods could include using backward or forward selection to eliminate variables that do not contribute to the model.

Following dimensionality reduction, statistical learning algorithms are applied to build predictive models. The most commonly used algorithms in REIMS are linear discriminant analysis (LDA), partial least square discriminant analysis (PLS-DA), and orthogonal partial least square discriminant analysis (OPLS-DA). Although the ones mentioned above are usually in studies, other algorithms should also be explored. Gredell et al. (2019) demonstrated that REIMS data is complex, and using just one approach to analyze all situations is not appropriate. This study used three different dimensionality reduction methods combined with eight machine learning to train four predictive models of beef attributes. As a result, the best machine learning for one model was not the best for the others.

A common technique to evaluate the performance of the predictive models is cross-validation. This technique consists of testing the predictive ability of the model with a new data set that was not used to train the models. The benefit of using a new data set to test the models is that it reduces the risk of overfitting the model. Since REIMS data is high-dimensional, it is easy to fit noise or random fluctuations in the predictive models. Therefore, cross-validation techniques are highly recommended to verify REIMS models. Particularly for classification models, the performance is evaluated with the accuracy, specificity, or sensitivity that resulted from cross-validation. The accuracy represents the percentage of correct classifications of the model. Sensitivity represents the ability of the model to classify a positive outcome correctly, while specificity measures the ability to classify a negative outcome correctly.

Some of the cross-validation methods are hold-out, *k*-fold, and leave-one-out cross-validation. In the hold-out approach, observations are divided into two subsets. Then one subset is

used to build the models, and the hold-out subset is used to validate the model (Gareth et al., 2021c). This method is the most simple and requires the least computational power. In k-fold cross-validation, the observations are randomly divided into *k* groups of similar size. Then, *k*-1 groups are used to train the model, and the remaining *k* group is used to validate the model. This process is repeated *k* times until all groups are tested (Kohavi, 1995). The leave-one-out cross-validation (LOOCV) takes one observation aside as a test set and trains the model with the remaining observations. This process is repeated until all the observations are tested without repetition. The drawback of LOOCV is that it is computationally intense since modeling/validation is realized the same number of times as observations (Kohavi, 1995). Currently, there is no clear definition of what method works better for REIMS because it depends on the underlying structure of the data. However literature on machine learning states that intermedium approaches (e.g., 10 or 5-k-fold cross-validation) are generally better than others because they have a balanced bias and variance (Kohavi, 1995; Gareth et al., 2021c).

REIMS and Meat Quality

Current research of REIMS in meat science is focused on evaluating its potential for applications in provenance and food fraud, processing systems, and meat quality. REIMS provides a chemometric profile in seconds, remotely, and with minimally invasive procedures. Therefore, REIMS could be used as an in-situ and real-time method to verify meat attributes. REIMS also could be used as a rapid screening method for chemical analysis since it provides chemical information without tedious sample preparation that conventional MS techniques implied.

The first published REIMS study in meat science emerged as a solution to detect adulterations of ground beef with horse meat or offal meat. Balog et al. (2016) evaluated the ability of REIMS to discriminate samples from different species and anatomical origins. Predictive

models with PCA-LDA algorithms could differentiate meat from different species (bovine, equine, venison) and between horse meat and liver with perfect accuracy. Although this technology could differentiate pure samples with high precision, it was unable to recognize when the samples were adulterated with less than 5% of another type of tissue. Similarly, Black et al. (2019) used PCA-LDA models to detect adulteration with offals in cooked and raw ground beef. They found that the detection limit of REIMS was 5% for adulteration of raw and cooked beef with brain and liver tissues and 1% for adulterations with heart, kidney, and large intestine. Genangeli et al. (2019) reported up to 100% accuracy of binary classification models for 6 tissue types from 6 animal species. Nevertheless, the latest study does not specify the number of animals per tissue type but rather the number of burns per tissue type. Hence, it is assumed that models do not account for variation across animals.

Marketing programs that use breed or production systems as differentiators could be validated with rapid identification methods (e.g., Certified Angus Beef). Currently, breed identification is carried out by phenotype identification or parentage genotype association. REIMS could be used as an in-line identification or verification system for the breed. Balog et al. (2016) sorted samples from four cattle breeds with 97% using PCA-LDA, while Gredell et al. (2019) differentiated beef samples from carcasses with Angus influence from non-Angus with 82.5% using a PLS-DA model. These results are only proof of concept since breed classes for both studies were determined by phenotype traits and not genetic classification.

REIMS also can identify variations in production systems. For example, REIMS was able to classify lamb samples into grass-fed and grain-fed categories with 93% accuracy (Zhai et al., unpublished data). He et al. (2021) achieved 96% accuracy with OPLS-DA models to discriminate never-frozen from frozen beef samples. REIMS can also identify the use of ractopamine (a

common β-agonist) in feeding regimes from pork loin, shoulder, and thigh samples. Guitton et al. (2018) classify samples of pigs fed with and without ractopamine regimens with 95% accuracy based on the metabolic profile generated by REIMS. These applications could be implemented at an industrial level to avoid mislabelling and increase meat marketability to the export market. In research, REIMS could also be used as a screening method to identify chemical variation generated by production methods. Zhang et al. (2021) found different metabolic profiles between dry-aged and stepwise-aging beef, and they made a preliminary identification of dipeptides and amines between the two processing systems. He et al. (2021) made inferences about differences in fresh and frozen beef metabolites using tentative identification of fatty acids.

Regarding quality attributes, Gredell et al. (2019) evaluated the ability of REIMS to classify beef strip loin samples into two categories of tenderness. They got a maximum accuracy of 90.8% using a feature selection combined with a support vector machine algorithm. REIMS can identify off-odor (boar taint) in pork neck samples. Verplanken et al. (2017) built predictive models using OPLS-DA with the chemometric profiling of REIMS that discriminated tainted from untainted boar samples with 99% accuracy. However, they were unsuccessful in identifying traces of the chemical compounds responsible for boar taint.

Repeatability and reproducibility of REIMS

The reliability of measurement techniques is determined by repeatability and reproducibility. A measurement method is repeatable when it can provide consistent measurements of the same sample measured with the same instrument, operator and within the same day. Reproducibility represents the variations in measurements obtained with different instruments and operators and in different locations and times (Tabb et al., 2010). Another important but less studied property of measurement techniques is robustness. Robustness represents the resilience of

an analytical technique to variation in measurement conditions without affecting the output. While robust methods are desirable for laboratory applications, robust techniques are almost mandatory for industrial applications. In general, experiments carried out in laboratories have controlled conditions, and sources of variability are reduced. On the other hand, controlling the source of variation without affecting the efficiency of production processes is generally challenging, so robustness is an important attribute to consider.

Since the sampling process in REIMS is thermally dependent, variables that influence the generation and transfer of heat impact measurement variability. Variables such as the type of sampling device, cutting speed, solvent flow rate, and cutting power influence peak intensity and mass spectra (Genangeli et al., 2019; He et al., 2021). The cutting power and geometry of the sampling electrode affect the signal intensity and noise by influencing the heat transfer rate and contact surface. Bodai et al. (2018) tested sharp versus round blades in a diathermy device and found that the sharp tip requires less power than tips with a larger contact area to get similar signals. Bodai et al. (2018) also found that the power intensity affects the signal-to-noise ratio, and an optimal intermediate value depends on the application that provides the best results. Similarly, He et al. (2021) determined that variations in the cutting power and the cutting speed affect the signal intensity of specific mass bins. The flow rate of the solvent injected into the MS source also affects the mass spectra by changing the matrix effect (He et al., 2021).

Genangeli et al. (2019) compared the performance of a CO_2 laser and a diathermic knife as sampling devices to classify biological tissues from 6 different species. They found that the variability of the total ion chromatogram between replicates of CO_2 laser is lower than the variability using a diathermic knife. The coefficient of variation (CV) of the CO_2 laser was between 9.76-12.00%, and the CV for a diathermic knife was 13.92-22.72%. However, it is important to

notice that the CO_2 laser was applied using a mechanical arm that stabilizes the sampling technique, while the diathermic knife was used by hand. He et al. obtained CV between 4.95-9.57% of specific ions on beef samples using a monopolar electrosurgical handpiece after optimizing solvent rate, cutting speed, and power intensity. Verplanken et al. (2017) evaluated REIMS repeatability by analyzing the variability of the ion peaks of an endogenous standard (m/z 699.497) in quality control samples (bovine meat). They found that 94% of the intensity peaks of the standard were within 2 standard deviations of the mean. Most studies used signal intensity or intensity of specific m/z to evaluate REIMS repeatability. However, because the spray generated in REIMS is highly variable, analysis of the relative abundance of specific components is more appropriate to determine repeatability.

High reproducibility and robustness of REIMS analysis have been demonstrated in the medical field to differentiate cancerous from noncancerous cells. Database from different surgical interventions, hospitals, and tissue types have been employed to develop predictive models that resulted in 97.7% sensitivity and 96.5% specificity for in-vivo histological diagnosis (Balog et al., 2013). In another study, a classification model to detect ovarian cancer was built with frozen samples and validated in-real time during multiple interventions resulting in an overall accuracy of 99.1% (Phelps et al., 2018). Strittmatter et al. (2014) analyzed the reproducibility of REIMS using a recursive maximum margin criterion (RMMC) plot, a supervised algorithm, including microorganism species and measures with three different mass spectrometers. Their RMMC plot showed that interspecies separation was more dominant than instrument clustering.

In food applications, Black et al. (2017) classified five different fish species with 99.0% accuracy in real-time using a model built with previously frozen samples. Verplanken et al. (2017) studied the reproducibility and robustness of REIMS to classify tainted boars, untainted boars, and

sows and to classify tainted and untainted boars. The predictive accuracy of their model with three classes dropped from 99% to 89% when building models using data from the same day and the same cone voltage to data collected on three consecutive days and varying cone voltage. However, the accuracy of the 2-classes model with a 100% accuracy was not affected when modeling with data collected on different days. He et al. (2021) obtained interday CV between 4.06-8.24% of specific ions in a beef sample.

REIMS is a non-destructive technique that has the potential to be implemented in processing plants as a real-time method to assess beef quality attributes. However, no studies have explored the reliability of REIMS measures on beef matrices collected on different days, instruments, or labs. Understanding the reliability of results to data variations due to the aforementioned factors will increase confidence in this technology, facilitating its implementation and opening to new studies.

CHAPTER II: ABILITY OF RAPID EVAPORATIVE IONIZATION MASS-SPECTROMETRY TO PREDICT TENDERNESS, JUICINESS, AND FLAVOR ATTRIBUTES OF TWO BEEF MUSCLES AT DIFFERENT AGING TIMES

Introduction

Beef palatability depends on tenderness, juiciness, and flavor. Tenderness has been previously defined as the most important quality trait for consumer satisfaction when consuming beef (Wheeler et al., 1990; Miller et al., 1995). However, when tenderness is acceptable, beef flavor plays a major role in the overall eating experience (Goodson et al., 2002; Behrends et al., 2005; Legako et al., 2015). In addition, juiciness contributes 7.4% to the overall palatability of beef (O'Quinn et al., 2018). The USDA quality grade attempts to predict the three palatability traits of beef based on carcass traits, including sex class, lean texture/firmness, marbling score, and animal maturity, although marbling score is the main determinant. The USDA quality grade segregates carcasses by the probability of having a positive eating experience (O'Quinn et al., 2018). However, instrumental marbling score explains less than 45% variability in juiciness, 40% variability in tenderness, and 61% overall palatability of *longissimus* muscle (Smith et al., 1985; Emerson et al., 2013). Therefore, the current grading system does not account for other significant sources of variation in tenderness and juiciness.

Real-time assessment of beef quality traits (flavor, tenderness, and juiciness) at current production speeds is challenging, and implementation of current techniques is impractical. Large commercial beef facilities, on average, can process 6,000 heads of beef cattle per day. Current methods to assess beef quality attributes are slow compared to the production rate of commercial facilities. Mechanical methods for measuring beef tenderness, such as slice shear force (SSF) or

Warner-Bratzler shear force (WBF), have been suggested to be implemented in-line (Shackelford et al., 1999). However, mechanical methods are not popular among producers because 1-inch ribeye steak is wasted during the analysis, and the time is required to complete such analyses.

Non-destructive techniques that can predict beef palatability traits (flavor, tenderness, and juiciness) in real-time could improve the current grading system. Rapid Evaporative Ionization Mass-Spectrometry (REIMS) is an ambient mass spectrometry technique that provides rapid chemometric profiling of biological tissues in-situ, without sample preparation (Balog et al., 2016). Since this method does not require sample preparation, data collection takes a few seconds, making REIMS a good alternative for assessing quality attributes in production lines. Previously, REIMS has been used to classify beef striploins samples aged 14 days into "tough" and "tender" classes defined by a cutoff of 20kg of SSF for tenderness and reported that REIMS could classify beef samples into the two tenderness categories with 91% accuracy (Gredell et al., 2019). However, REIMS data for this study were collected from steaks aged 14 days, so the potential of REIMS to be used to predict quality during grading is still unknown. Therefore, the goal of this study was to evaluate the ability of REIMS as a real-time method to predict meat tenderness, juiciness, and flavor using data collected at the grading time.

Materials and Methods

Product collection, aging, and measure of sensory attributes were performed at Texas Tech University (TTU), while REIMS and data analysis of REIMS data were performed at Colorado State University (CSU).

Product Collection and Aging

Forty-two (n=42) USDA Select and forty-two USDA upper 2/3 of the Choice grade were selected from a commercial beef processing facility over three production days. Slivers from both

sides of the carcasses were collected at the grading time (around 36 h post-mortem; GR samples), frozen with liquid nitrogen, and stored at -80°C until further analysis. After grading, whole striploins (*Logissiumus medius*; LM) and tenderloins (*Psoas major*; PM) were collected from both sides of the carcasses and transported in coolers (2-4°C) to the meat laboratory at TTU. Upon arrival, striploins (from both sides of the carcasses) were fabricated into six 6-cm sections. The *gluteus medius* from the posterior part of the LM was excluded. Similarly, heavy connective tissue was removed from the PM from both sides and fabricated into six 9-cm sections. Each of the six sections per muscle was vacuum sealed and randomly assigned to one of the six aging periods (3, 14, 28, 42, 56, and 70 days). However, only samples aged 3, 14, and 28 days were used in this study to simulate industry practices. Portions were aged at 0-2°C for the corresponding aging period and were stored at -80°C until further analysis. Frozen LM sections were fabricated into two 2.54-cm steaks using a band saw, and each steak was randomly assigned to shear force and sensory panels. Likewise, frozen PM sections were fabricated into three 2.54-cm steaks, and two steaks were randomly assigned to sensory panels and the remaining steak to shear force.

Cooking procedure

Steaks for sensory panels and shear force evaluation were cooked using the same procedure. Frozen steaks were thawed at 0-2°C for 24-48 h to attain raw internal temperatures of 0-2°C before cooking. Steaks were cooked in an oven (Model SCC WE 61 E; Rational, Landsberg am Lech, Germany) at 204°C and 0% relative humidity to a final internal temperature of 71°C. Peak internal temperatures were recorded in the geometric center of the steaks using a thermometer (AccuTuff 34032, Cooper-Atkins Corporation, Middlefield, CT).

Trained Sensory Panel

The LM and PM sensory panels were performed similarly but separately. Panels were performed in 44 sessions/muscle, with a maximum of 2 sessions/day, a resting time of at least 7h between sessions, a maximum of 12 samples/session, and at least 8 trained panelists/session. All samples (n=504 per USDA grade) were randomly assigned to each session, ensuring that both USDA quality grades and all aging treatments were present in each session. Immediately after cooking, external fat and connective tissue were removed, and the steaks were cut into small cubes. Two to three cubes were served to the panelists in individual booths equipped with a red incandescent light. Panelists evaluated tenderness, juiciness, and 12 flavor descriptors from the Beef Lexicon (Adhikari et al., 2011) described in Table 1 on a 100-mm unstructured line scale. The left end of the line scale was anchored with "not present," "extremely dry," or "extremely tough," and the right end represented "extremely intensity," "extremely juicy," or "extremely tender." A warm-up sample (USDA Low Choice strip steak) was served at the beginning of each session to calibrate panelists. Consensus on the attributes of the warm-up sample was reached before moving to the study samples.

Shear Force

Tenderness of both muscles was evaluated using Warner-Bratzler shear force (WBF) and slice shear force (SSF) as described by Shackelford et al. (1999) and Lorenzen et al. (2010). Samples were cooked, and internal peak temperatures were recorded following the protocol described above. Within 2-3 min after recording peak temperature, the lateral ends of the steak were removed, and a 1 x 5 cm slice was cut parallel to the muscle fiber. SSF was measured by shearing the slice perpendicular to muscle fibers with a universal testing machine (Instron Corp., Canton, MA) equipped with a flat, blunt-end blade (500 mm/min crosshead speed and 100 kg load

capacity). The remaining part of the sample was cooled down to room temperature or below, and 2 to 6 cores (1.2-cm diameter) were removed parallel to the muscle fibers orientation. The WBF was measured by shearing each core perpendicular to the fiber with the universal testing machine equipped with a WBF blade (74.2 mm wide x 138.7 mm long x 0.99 mm thick, 200 mm/min crosshead speed, and 100kg load cell capacity).

Rapid Evaporative Ionization Mass Spectrometry (REIMS)

Metabolomic profiles of the GR samples were obtained using a quadrupole time-of-flight mass spectrometer (Synapt G2-Si Q-ToF, Waters Corporation, Milford, MA) equipped with a REIMS source (Waters Corporation). An electronic probe (Waters Corporation) powered by an electrosurgical generator (Erbe VIO 50C, Erbe Elektromedizin GmbH, Tubingen, Germany) was used as a sampling device. The electrosurgical generator was set to dry cut mode and maximum cutting power of 40W. Mass spectra from 100-1,000 m/z were acquired in negative ion mode, with cone voltage at 40V and heater bias at 60V. Samples were thawed at 0-4°C for 16-24 h and randomly sorted for REIMS analysis. The sampling device was used to generate at least 5 burns over the surface of individual samples. Burns were made in the 4 corners, and the middle of a 2.5 x 2.5 cm square from the surface of the sample, and each burn lasted approximately 1s. Extra burns were made in a random location equidistant to the last burn when additional burns were required due to low signal intensity. A 40µg/L of leucine-enkephalin/isopropanol solution was injected directly into the REIMS sources (flow rate: 200µg/min) for lock mass correction. REIMS data was pre-processed with AMX Recognition software (version 1.0.2184.0, Waters Corporation). This process included lock mass correction using leucine-enkephalin (554.2615m/z), background subtraction using standard Masslynx pre-processing algorithms, total ion current normalization, peak binning in 0.5m/z intervals, and exclusion of bins in the range of 550-600m/z.

Data analysis

The statistical analyses described here were performed in R statistical computing program (version 4.1.0, 2018).

Tenderness, juiciness, and flavor classification

Cutoffs for juiciness and tenderness, including panel tenderness (PT), SSF, and WBF, were specified close to the means of each attribute to divide the samples of 3, 14, and 28 days into relatively balanced classes. For juiciness, LM and PM samples with values lower than or equal to 55 were classified as "dry" and the remaining as "juicy." LM samples with PT higher or equal to 55 were classified "tender" and the remaining as "tough," while PM samples with PT higher or equal to 77 were classified "tender" and the remaining as "tough." SSF cutoffs were 14.0 and 12.0 kg, and WBF cutoffs were 3.1 and 2.9 kg for LM and PM, respectively. Samples with higher values of SSF and WBF than the cutoffs were classified as "tough" and lower than the cutoff were classified as "tender."

Hierarchical clustering was used to group samples into classes with similar flavor performance. Briefly, a Principal Component Analysis (PCA) of the 13 flavor attributes was performed using the *PCA* function from the *FactoMineR* package (Lê et al., 2008). Then, the *HCPC* function from the *FactoMineR* package was used to group the samples with similar flavor performance based on the first five principal components (PC). Samples aligned to positive flavor attributes were assigned to the "acceptable" class, and those aligned to the negative ones to the "unacceptable" class.

Predictive model building

Classification models were built using REIMS data of GR samples to predict quality attributes of PM and LM samples on days 3, 14, and 28. Models were built using three

dimensionality reduction methods and ten machine learning algorithms. PCA, Feature selection (FS), and both methods combined (PCA-FS) were used to reduce the dimensionality of the REIMS preprocessed data. PCA was performed with the *PCA* function from the *FactoMineR* package, and the number of principal components was selected with a scree plot. In FS, highly correlated pairs ($|r| > 0.9$) were identified and removed from the data; then, a recursive feature elimination was performed using the *rfe* function from the *caret* package (Kuhn, 2008). The PCA-FS was performed with PCA, followed by recursive feature elimination using the aforementioned functions. Fifteen machine learning algorithms (Table 2) from the *caret* package repository (Kuhn, 2008) were used to build predictive models for each combination of muscle, aging period, response variable, and dimensionality reduction method. Models were trained and tuned with the *train* function of the *caret* package (Kuhn, 2008) using 10-fold cross-validation. The performance of the models was analyzed with the accuracy, sensitivity, and specificity of the top-performing models. Sensitivity results by dividing the true positives by the actual positives, while specificity is the fraction of true negatives by the actual negatives. "Tough," "dry," and "unacceptable" samples were treated as positive for sensitivity calculation of tenderness, juiciness, and flavor models, respectively. While "Tender," "juicy," and "acceptable" were treated as negative for specificities calculation.

Results and discussion

Tenderness and juiciness classes

LM and PM samples were classified into "tough" and "tender" classes for tenderness based on three measurement methods (SSF, WBF, and PT) and into "dry" and "juicy" classes for juiciness. The numbers of samples per class and aging period for LM and PM are presented in Table 3 and 4Table 4, respectively. As expected, the number of tough samples decreased with the

aging time, and the number of tender ones increased with aging in all tenderness classifications and both muscles. It is known that tenderness increases with aging due to proteolysis of structural proteins of the muscles (Koohmaraie, 1996). Juiciness classes of LM samples were similar in all aging periods since juiciness did not change with aging for these muscles ($P = 0.24$). Lepper-Blilie et al. (2016) also found similar results when comparing the juiciness of LM steaks aged 14 to 49 days and did not find differences in juiciness. However, the number of samples classified as dry with PM increased with time. Juiciness of PM samples aged 3 days was higher than those aged more than 42 days ($P < 0.01$), which could explain the increase of samples in the dry class with aging.

Flavor classes

LM and PM samples aged 3, 14, and 28 days were classified using PCA of the ten flavor attributes evaluated in the sensory panels (Figures 1 and 2). Loading plots of the PCA of both muscles showed that brown, beefy, and roasted flavors were positively correlated, and these attributes were negatively associated with livery flavor. As expected, fat-like and buttery flavors were associated since the intensity of both increases with the intramuscular fat (Legako et al., 2016; O'Quinn et al., 2016). However, they were more closely related to umami than to other flavors. Foraker et al. (2020) found similar results when performing a discriminant function analysis of aged beef LM sensory attributes. They found that positive flavors such as roasted, beef identity, fat-like, and browned were positively associated with each other but negatively associated with off-flavors (sour, liver-like, oxidized, and metallic). Gredell (2018) also found similar results when performing a PCA of beef samples from different USDA grades, cattle breeds, and production systems. PCA analysis by Gredell showed that browned and beef ID attributes and buttery and fat-like flavors were associated.

The hierarchical clustering of the sample resulted in two classes that were defined as "acceptable" and "unacceptable" (Figures 1 and 2). Samples classified as acceptable were mostly located in the positive quadrant of dimension 1 (dim1), suggesting higher intensity on the positive attributes and lower values of the negative ones. Classifications were mostly driven by the PCA's dim1, and the major contributor to the dim1 was the beef identity. Foraker et al. (2020) and Gredell (2018) obtained similar results by performing a multivariable analysis of flavor attributes. Both studies concluded that beef identity was the major contributor to separating samples with positive from negative performance.

Classification models of tenderness, juiciness, and flavor of striploins

Three dimensionality reduction methods combined with different machine learning algorithms were trained using 10-fold cross-validation. The accuracy, sensitivity, and specificity (based on 10-fold cross-validation) of the best models for each attribute and aging time of LM are reported in Table 5.

Tenderness

For tenderness models, the SSF and WBF of 28 days (d28) and PT of 3 days (d3) showed the best accuracies. PLSDA plots of tenderness classes (Figures 3-5) showed overlapped classes. These results were expected since tenderness cutoffs were close to the means, and tenderness is a numerical attribute. The PLSDA plot of SSF and WBF of d28 and PT of d3 showed the smallest overlap between those tenderness classes (Figures 3-4), which explains the highest accuracy of those models. SSF d2 and PT d3 also showed high accuracy, but the sensitivity of the SSF model to classify tough samples was very low, and the specificity of the PT model was low (Table 5). The best performing model for tenderness was the WBF d28, showing an accuracy of 94.8% with 84.2% sensitivity and 96.7% specificity (Table 5). Specificity of all tenderness models increased

with aging, and sensitivity decreased with aging in all tenderness models, except for WBF models. The ability of REIMS to classify tough samples decreased with aging time, and the ability to sort tender samples increased with aging time. This observation could be attributed to the data structure used in this study. Data of d3 and d28 were unbalanced because most d3 observations were tough, while most d28 were tender. Machine learning algorithms perform poorly when data is unbalanced because the models favor the larger class (Cieslak and Chawla, 2008). In addition, overlap between classes and a small number of observations in the minority group contribute to the low performance of predictive models (Batista et al., 2004). This phenomenon was observed when 50 random iterations of 10-fold cross-validation were performed, and the WBF, SSF, and PT values versus the proportion of times specific samples were misclassified were plotted (Figures 8-10). Samples in the minority classes were more frequently misclassified than samples in the majority classes. Compared to other studies, the accuracies of tenderness models obtained in this study were similar to Gredell et al. (2019). These researchers reported 90.8% accuracy of a classification model that differentiated tough from tender samples of LM aged 14 days using SSF= 20.0kg as a cutoff.

The limitation of REIMS to analyze high molecular weight compounds could limit its ability to predict tenderness with high accuracy. Studies have shown that a large part of the variability in beef tenderness depends on muscle proteins and intramuscular fat (Platter et al., 2003; Emerson et al., 2013; Picard et al., 2014; Gagaoua et al., 2018). During the first few days of post-mortem, the cross-sectional area of muscle fiber and sarcomere length are large contributors to beef tenderness, while background toughness becomes more relevant after 14 days of aging (Dubost et al., 2013). Those factors depend on contractile proteins, cytoskeletal and stromal proteins found in muscle tissues. Because REIMS requires that the compounds found in the samples be vaporized, only compounds with molecular weight between 1000 and 1500Da reach

the mass spectrometer. Therefore, REIMS can analyze chemical compounds, including fatty acids, sugars, phospholipids, and small peptides, but cannot detect large peptides or proteins (Ross et al., 2021). REIMS could provide information on metabolites identified as possible indicators of beef tenderness (for example, malic acid, glucose, glucose, glucose-6-phosphate) and compounds related to intramuscular fat (King et al., 2019). However, REIMS may not be able to analyze intact proteins (e.g., collagen, desmin) that also influence tenderness variability.

Flavor

The low performance of flavor may be due to the complexity of the response variable (Table 5). Flavor perception results from taste and odor sensations and can be defined by multiple flavor descriptors (for example, browned, sourness, or beef identity). Specific flavor attributes have been associated with numerous chemical compounds or groups of compounds. Elaborating a composite score that condenses ten flavor attributes (Figure 1) into two classes might have simplified beef flavor interpretation. However, clustering multiple flavor attributes represents a more difficult problem in building classification models based on the chemical fingerprint. Mathematically, when the number of dimensions (attributes) increases, the distance between observations increases, making it more difficult to find nearby observations in the hyperplanes (Aremu et al., 2020). In other words, samples in one class could be chemically different because the classification results from a combination of chemically unrelated attributes. Since machine learning algorithms use similarities in the training dataset to predict classes of the test data, more data will be required to cover a broader range of flavor profiles.

Gredell (2018) evaluated the ability of REIMS to predict flavor classes defined with multivariate analysis. The accuracy of his model for binary classification was 73.8%. However, in that study, beef samples were from different USDA grades, cattle breeds, and production systems,

which could increase variability in the fingerprint across classes and increase the error. Additionally, Gredell (2018) only used PLSDA algorithms to build the model, whereas in the current study, multiple algorithms were evaluated, and the best one was reported. Our results using PLSDA algorithms for flavor classification ranged from 55.0 to 88.6% (Table 5). Verplanken et al. (2017) also evaluated the ability of REIMS to predict pork flavor. They demonstrated that REIMS could classify pork carcasses with untainted and tainted boar flavor with perfect accuracy using an orthogonal PLSDA model. However, identifying boar taint could be less complex than identifying beef flavor profiles because boar taint is a specific flavor with a physiological origin. Boar taint is an odor occurring in the meat of non-castrated male pigs and is associated with high levels of skatole and androstenone (Verplanken et al., 2017). In contrast, beef flavor profile evaluated in the present study results from a combination of attributes that could come from different sources of variation.

Juiciness

PLSDA plots corresponding to juiciness classes (Figure 7) show an overlap between the two classes in all aging times, which could explain the low performance of the three juiciness models. In the PLSDA plot, the overlap implies similarity in the REIMS data of both classes. These results are not surprising since juiciness classes were defined with a cutoff close to the average juiciness of all the samples. There is not a gap that could differentiate juicy from dry samples chemically. Samples corresponding to different classes but with similar values of juiciness are probably more similar than samples in the same class with different values. For example, a sample considered as juicy with a juiciness value of 55.1 is perhaps more similar to a sample with a juiciness value of 54.7 classified as dry than a sample with a juiciness value of 63 classified as juicy. This could explain why after 50 random iterations of 10-fold cross-validation, samples that

were more frequently misclassified had a juiciness value close to the cutoff on d3 and d28 (Figure 11). However, the reason why the same pattern is not present in 14 days (d14) models is unknown. Similar to tenderness, juiciness is also influenced by intramuscular fat and protein composition of muscle tissues (Dubost et al., 2013). Therefore, the inability of REIMS to evaluate proteins could contribute to low prediction accuracy for this attribute.

Another source of error in all models could be the inconsistency in the cooking temperatures of the samples used for sensory evaluation. Even following the same cooking protocol, it is very common to have minor variations in the cooking temperature due to uncontrolled variables during the cooking process (e.g., steak size or location in the oven). The three sensory attributes evaluated are highly dependent on the cooking temperature. Steaks cooked with a higher internal temperature are less tender and less juicy than steaks cooked with a lower internal temperature (Cross et al., 1976; Savell et al., 1999; McKillip et al., 2017). Beef flavor is produced by chemical reactions catalyzed by temperature; therefore, cooking temperature plays a major role in developing beef flavor compounds (Kerth and Miller, 2015). Consequently, using REIMS on GR samples does not account for variations in sensory attributes affected by extrinsic factors during storage, transportation, and cooking.

Classification models of tenderness, juiciness, and flavor of tenderloins

The accuracy, sensitivity, and specificity of the best models for PM are reported in Table 6. In all cases, except for the model of flavor of d3 and PT of d28, the accuracy of LM models was higher than PM models. Similar to LM models, the specificity of SSF and WBF increased, and sensitivity decreased with aging time due to unbalanced data. The number of tender samples increases with aging, enhancing the ability of the model to predict the majority class.

Since REIMS was collected on GR samples from LM muscle, the ability of REIMS to predict sensory attributes of PM muscle was expected to be lower. These two muscles are metabolically different, with distinct biochemical properties (Abraham et al., 2017). While LM is predominantly glycolytic, PM is primarily oxidative (Kim et al., 2021). LM contains more muscle fiber type IIX than type I and IIA, while PM comprises more type I fibers than type IIX and IIA fibers (Kim et al., 2021). Several proteins degrade at different rates in both muscles suggesting that proteolysis mechanisms related to quality attributes are muscle-dependent (Kim et al., 2021). Proteome analysis of LM and PM revealed that sarcoplasmic proteins, including metabolic enzymes, antioxidants, and chaperones proteins, differed in both muscles (Joseph et al., 2012). Previous authors have reported that sensory attributes change differently in both muscles due to metabolic differences. Tenderness of the LM improves until 21 days post-mortem, while the PM improves until day 7 with no additional improvement observed with extended aging (Nair et al., 2019). Additionally, factors that affect tenderness are specific to the muscle. Gagaoua et al. (2018) found that the tenderness variability of LM depends on carcass characteristics, while the *semitendinous* on fattening period-related factors (e.g., average daily gain, feed efficiency). Therefore, even if particular fingerprints of the LM are found to classify carcasses on tenderness with perfect accuracy, it is expected that the prediction accuracy of the LM muscle to be lower.

Conclusions

Non-destructive techniques that can predict meat quality attributes in real-time would be a great asset for the meat industry. Consumer satisfaction in eating beef depends on the tenderness, juiciness, and flavor. Therefore, techniques that allow predicting these attributes and can be implemented in the current grading system would improve the consumer experience, increasing consumer trust and demand for beef. This study demonstrated that REIMS could potentially be

used as a real-time, in-situ technique to classify beef carcasses into flavor, juiciness, and tenderness classes at different aging times. However, the ability of REIMS to predict meat juiciness and tenderness may have been limited by its inability to analyze all of the proteins related to these attributes. Moreover, working with a composite score for flavor classes improves the interpretation of this complex attribute but increases model complexity. Developing a predictive model of beef flavor that encompasses a wide range of sensory profiles will require a large sample size to cover a wide range of flavor profiles. All classes evaluated were overlapped and were occasionally unbalanced, which affected the efficacy of REIMS in classifying the samples with high accuracy. Therefore, further exploring the true potential of REIMS will require increasing the sample size and developing a sampling method that allows increasing the gap between sensory evaluations.

Table 1. Definition and reference standards for beef sensory attributes and their intensities based on Adhikari et al. (2011), where 0 = none and 100 = extremely intense

Attribute	Definition	Reference
Tenderness	The overall tenderness of the sample	Strip steak to 160°F = 60 Tenderloin to 160°F = 95
Juiciness	The amount of perceived juice that is released from the product during mastication	Carrot = 55 Strip steak cooked to 175°F = 60 Strip steak cooked to 135°F = 75 Watermelon = 95
Beef Flavor ID	Amount of beef flavor identity in the sample	Swanson's Beef Broth = 35 Beef brisket (160°F) = 80
Bitter	The fundamental taste factor associated with a caffeine solution	0.01% caffeine solution = 15 0.02% caffeine solution = 25
Browned	Aromatic associated with the outside of grilled or broiled meat; seared but not blackened or burnt	Beef suet (broiled) = 55
Buttery	Sweet, dairy-like aromatic associated with natural butter	Land O'Lakes unsalted butter = 45
Fat-Like	The aromatics associated with cooked animal fat	Hillshire Farm Beef Lit'l Smokies = 45 Beef suet = 80
Liver-Like	The aromatics associated with cooked organ meat/liver	Beef liver = 50
Metallic	The impression of slightly oxidized metal, such as iron, copper, and silver spoons	0.10% potassium chloride solution = 10 Select striploin steak (60°C internal) = 25 Dole canned pineapple juice = 40
Roasted	Aromatic associated with roasted meat	80% lean ground chuck = 65
Sour	The fundamental taste factor associated with citric acid	0.015% citric acid solution = 10 0.050% citric acid solution = 25
Umami	Flat, salty, somewhat brothy. The taste of glutamate, salts of amino acids, and other molecules called nucleotides	0.035% Accent Flavor Enhancer solution = 50

Table 2. R machine learning algorithms, functions, and libraries used for building classification models

Algorithm	Function	Libraries
Stochastic gradient boosting	gbm	gbm, plyr
Support vector machine radial kernel	svmRadial	kernlab
Support vector machine linear kernel	svmLinear	kernlab
Support vector machine polynomial kernel	svmPoly	kernlab
Linear discriminant analysis	lda	MASS
eXtreme gradient boosting	xgbTree	xgboost, plyr
Penalized discriminant analysis	pda	mda
Boosted logistic regression	LogitBoost	caTools
Random forest	rf	randomForest
Generalized linear model	glm	stats
Lasso and elastic-net regularized generalized linear models	glmnet	glmnet, Matrix
K-nearest neighbors	knn	class
Recursive partitioning tree	rpart	rpart
Bagged classification tree	treebag	ipred, plyr, e1071

Table 3. Number of observations (percent) of striploins (LM) per class of slice shear force (SSF), Warner-Bratzler shear force (WBF), sensory panel tenderness (PT), flavor, and juiciness

Attribute[1]	Classes[2]	3 days	14 days	28 days
SSF	Tender (SSF≤14.0kg)	21 (25.3)	44 (55.0)	66 (82.5)
	Tough (SSF>14.0kg)	62 (74.7)	36 (45.0)	14 (17.5)
WBF	Tender (WBF≤3.1kg)	17 (20.5)	44 (55.0)	60 (75.0)
	Tough (WBF>3.1kg)	66 (79.5)	36 (45.0)	19 (23.8)
PT	Tender (PT≥55)	12 (14.5)	28 (35.4)	57 (71.3)
	Tough (PT<55)	71 (85.5)	51 (64.6)	23 (28.8)
Flavor[2]	Acceptable	53 (63.9)	46 (58.2)	35 (43.8)
	Unacceptable	30 (36.1)	33 (41.8)	45 (56.3)
Juiciness	Dry (juiciness≤55)	37 (44.6)	39 (49.4)	40 (50.0)
	Juicy (juiciness>55)	46 (55.4)	40 (50.6)	40 (50.0)

[1] SSF, slice shear force; WBF, Warner-Bratzler shear force; PT, sensory panel tenderness
[2] Flavor classes based on hierarchical clustering of the samples using PCA

Table 4. Number of observations (percent) of tenderloins (PM) per class of slice shear force (SSF), Warner-Bratzler shear force (WBF), sensory panel tenderness (PT), flavor, and juiciness

Attribute[1]	Classes	3 days	14 days	28 days
SSF	Tender (SSF≤12.0kg)	32 (38.1)	47 (57.3)	50 (61)
	Tough (SSF>12.0kg)	52 (61.9)	35 (42.7)	32 (39)
WBF	Tender (WBF≤2.9kg)	34 (40.5)	39 (47.6)	52 (63.4)
	Tough (WBF>2.9kg)	50 (59.5)	43 (52.4)	30 (36.6)
PT	Tender (PT≥77)	54 (64.3)	41 (48.8)	45 (54.2)
	Tough (PT<77)	30 (35.7)	43 (51.2)	38 (45.8)
Flavor	Acceptable	43 (51.2)	43 (51.2)	23 (27.7)
	Unacceptable	41 (48.8)	41 (48.8)	60 (72.3)
Juiciness	Dry (juiciness≥55)	31 (36.9)	39 (46.4)	43 (51.8)
	Juicy (juiciness<55)	53 (63.1)	45 (53.6)	40 (48.2)

[1] SSF, slice shear force; WBF, Warner-Bratzler shear force; PT, sensory panel tenderness

[2] Flavor classes based on hierarchical clustering of the samples using principal component analysis (PCA)

Table 5. Top machine learnings prediction accuracies (based on 10-fold cross-validation, percent) for striploin (LM) tenderness, juiciness, and flavor based on REIMS of grading (GR) samples

Models[1]	Aging period[2]	Top model[3]	N of predictors	Accuracy	Sensitivity	Specificity
SSF	d3	FS/GLMNET	155	85.8	95.2	57.1
	d14	PCA-FS/XGBoost	12	84.8	80.6	88.6
	d28	PCA-FS/XGBoost	3	91.5	50.0	100.0
WBF	d3	FS/XGBoost	14	86.5	95.5	52.9
	d14	FS/RF	325	82.5	77.8	86.4
	d28	PCA-FS/SVM Poly	44	94.8	84.2	96.7
PT	d3	FS/XGBoost	23	93.9	100.0	58.3
	d14	FS/XGBoost	46	87.4	78.6	92.2
	d28	FS/PLSDA	600	82.5	59.5	95.2
Flavor	d3	FS/GBM	181	86.7	76.7	92.5
	d14	FS/PLSDA	265	88.6	92.5	78.3
	d28	FS/XGBoost	17	81.3	86.7	74.3
Juiciness	d3	FS/XGBoost	84	84.5	89.1	78.4
	d14	FS/XGBoost	26	83.6	84.6	82.5
	d28	FS/XGBoost	61	82.5	85.0	80.0

[1] SSF, slice shear force; WBF, Warner-Bratzler shear force; PT, sensory panel tenderness

[2] d3, 3 days; d14, 14 days; d28, 28 days of aging

[3] FS, Feature selection; PCA-FS, principal component analysis - feature selection; GLMNET, Lasso and elastic-net regularized generalized linear models; XGBoost, extreme gradient boosting; RF, random forest; SVM, support vector machine; PLSDA, partial least square discriminant analysis; GLM, generalized linear model.

Table 6. Top machine learnings prediction accuracies (based on 10-fold cross-validation, percent) for tenderloin (PM) tenderness, juiciness, and flavor based on REIMS of grading (GR) samples

Models[1]	Aging period[2]	Top model[3]	N of predictors	Max accuracy	Sensitivity	Specificity
	d3	PCA-FS/SVM Poly	59	82.1	88.5	71.9
SSF	d14	PCA-FS/XGBoost	51	73.4	68.6	76.6
	d28	FS/PLSDA	38	89.0	75.0	98.0
	d3	FS/SVM Poly	205	78.6	88.0	64.7
WBF	d14	FS/RF	4	77.7	81.4	74.4
	d28	FS/SVM Poly	32	81.5	76.7	84.6
	d3	PCA-FS/XGBoost	34	83.8	73.3	88.9
PT	d14	FS/XGBoost	80	81.0	81.4	80.5
	d28	PCA-FS/SVM	46	84.3	78.9	88.9
	d3	PCA-FS/XGBoost	29	83.3	75.6	90.7
Flavor	d14	FS/XGBoost	4	73.8	75.6	72.1
	d28	PCA-FS/SVM Poly	35	84.3	95.0	56.5
	d3	FS/XGBoost	30	81.1	90.6	64.5
Juiciness	d14	PCA-FS/SVM Poly	65	79.5	84.4	74.4
	d28	PCA-FS/SVM Poly	59	79.6	80.0	79.1

[1]SSF, slice shear force; WBF, Warner-Bratzler shear force; PT, sensory panel tenderness

[2]d3, 3 days; d14, 14 days; d28, 28 days of aging

[3]FS, Feature selection; PCA-FS, principal component analysis - feature selection; GLMNET, Lasso and elastic-net regularized generalized linear models; XGBoost, extreme gradient boosting; RF, random forest; SVM, support vector machine; PLSDA, partial least square discriminant analysis; GLM, generalized linear model.

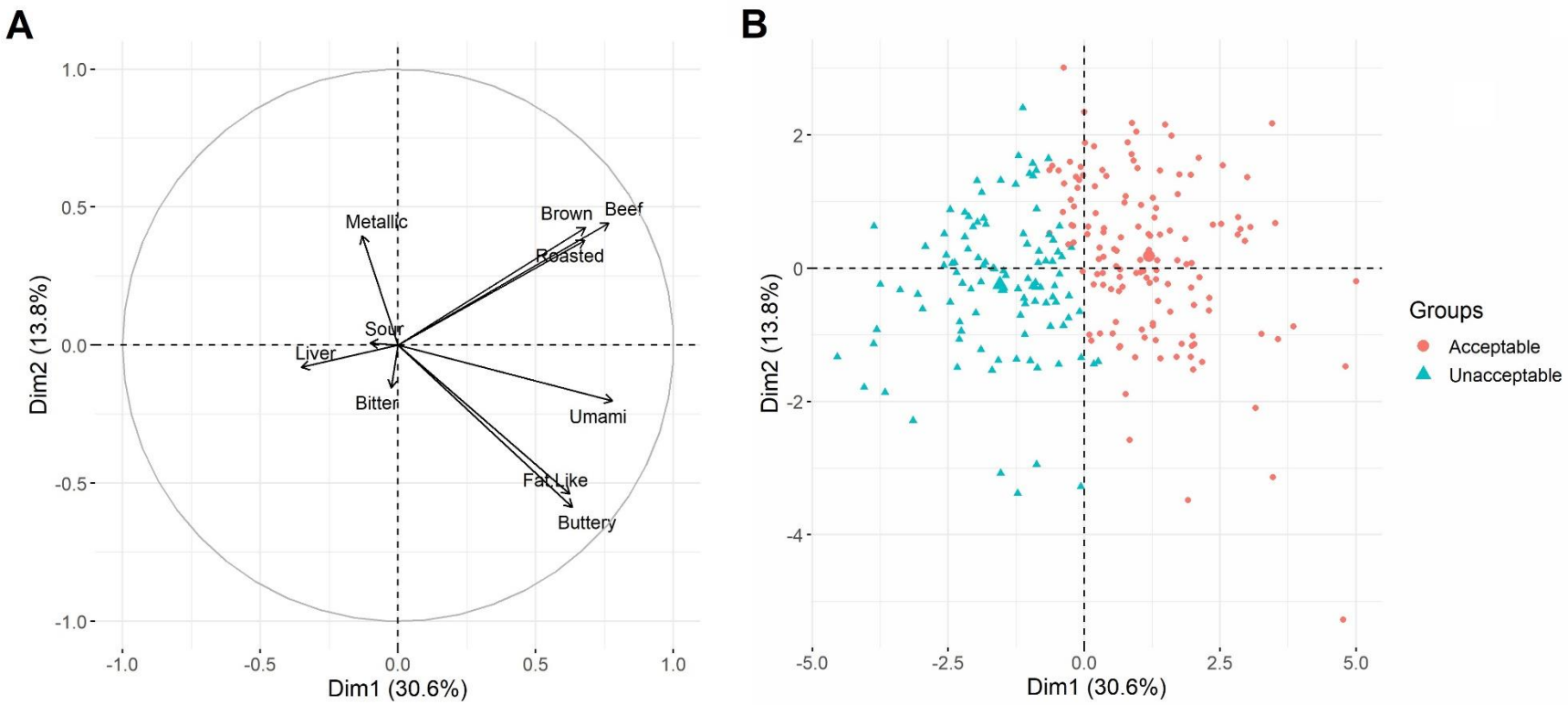

Figure 1. Principal component analysis (PCA) of 10 flavor attributes of beef striploins (LM; n=242) aged 3, 14, and 28 days. (A) Loading plot of PCA and (B) flavor classes based on hierarchical clustering of the samples in the PCA.

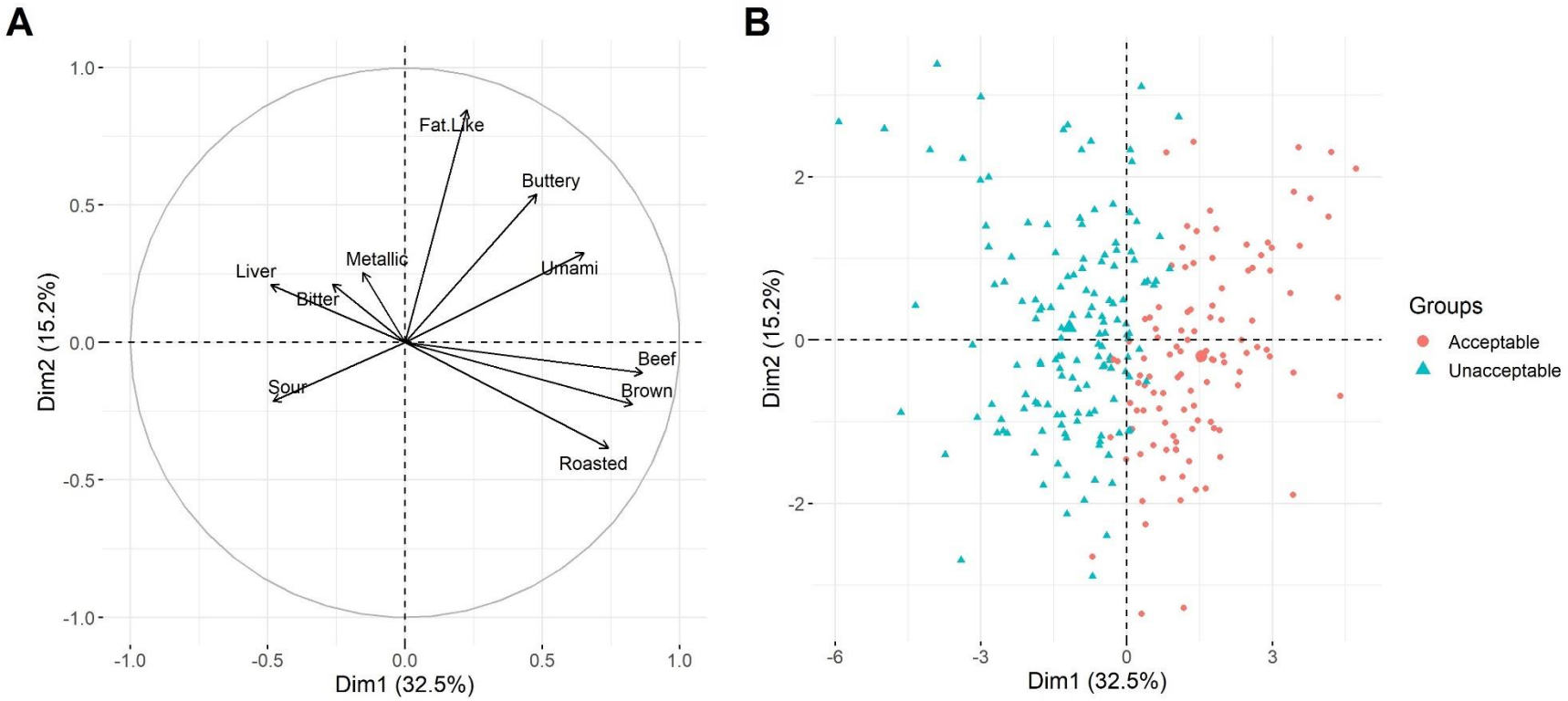

Figure 2. Principal component analysis (PCA) of 10 flavor attributes of beef tenderloins (PM; n=251) aged 3, 14, and 28 days. (A) Loading plot of PCA and (B) flavor classes based on hierarchical clustering of the samples in the PCA.

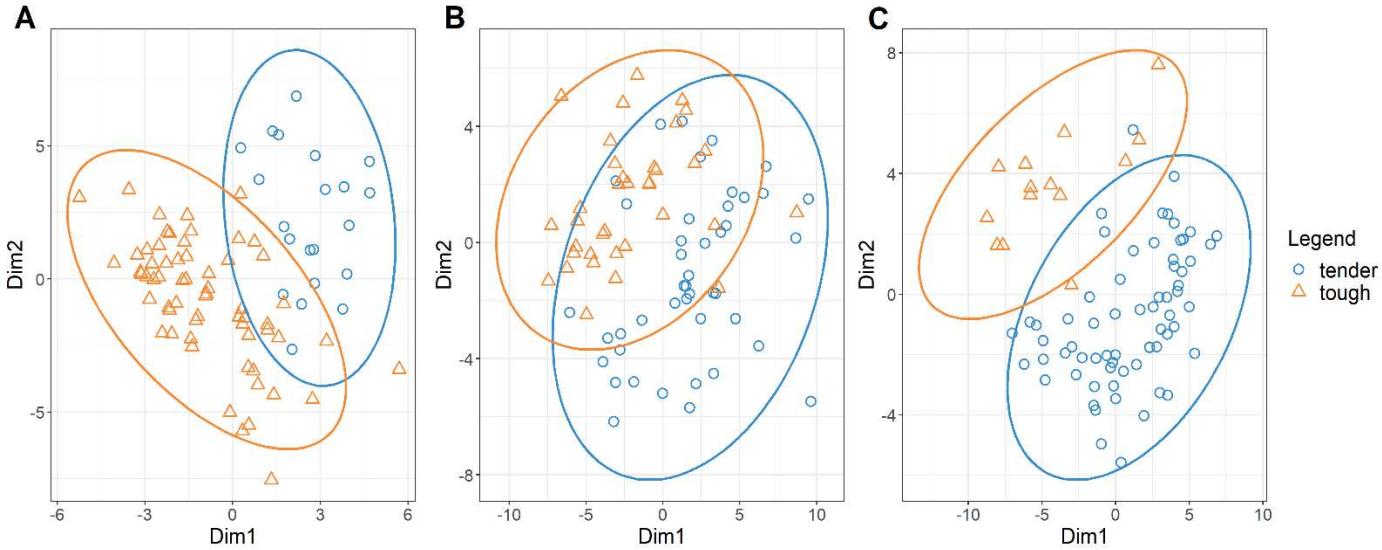

Figure 3. Partial least squares-discriminant analysis (PLSDA) plot of slice shear force of striploin (LM) classes corresponding to (A) 3, (B), 14, and (C) 28 days of aging

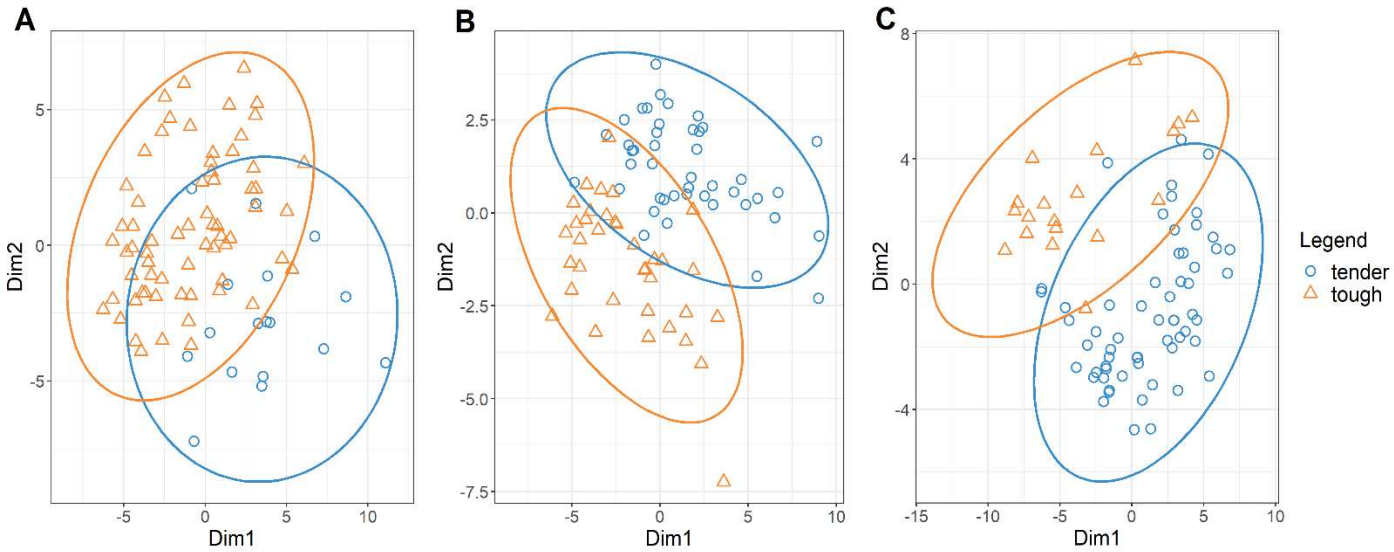

Figure 4. Partial least squares-discriminant analysis (PLSDA) plot of Warner-Bratzler shear force of striploin (LM) classes corresponding to (A) 3, (B), 14, and (C) 28 days of aging

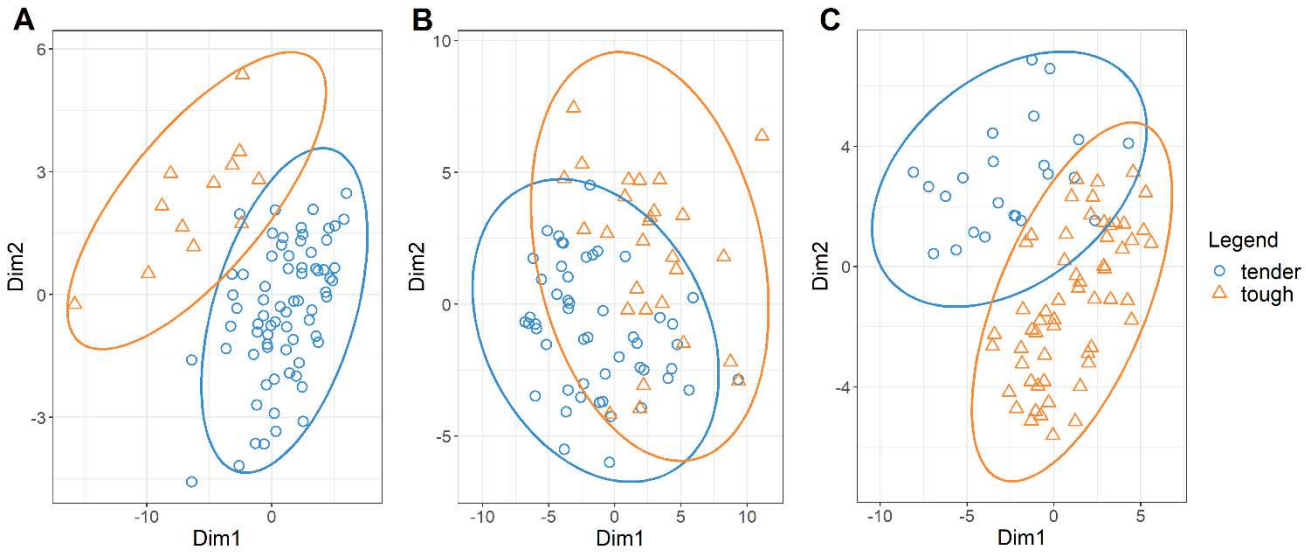

Figure 5. Partial least squares-discriminant analysis (PLSDA) plot of sensory panel tenderness of striploin (LM) classes corresponding to (A) 3, (B), 14, and (C) 28 days of aging

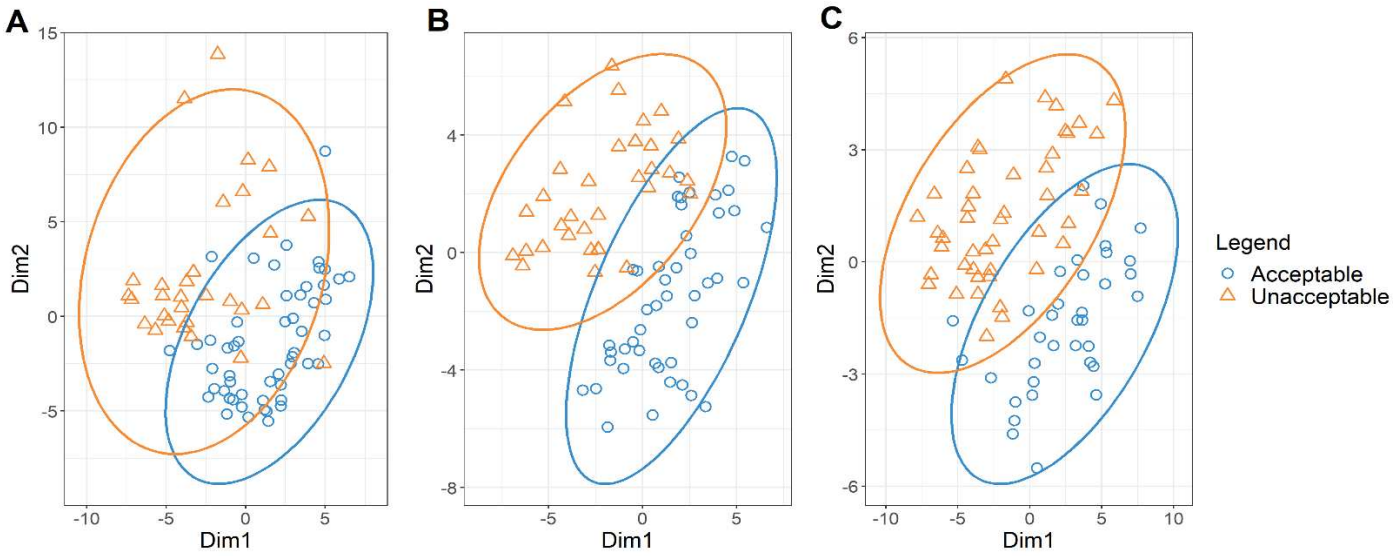

Figure 6. Partial least squares-discriminant analysis (PLSDA) plot of flavor of striploin (LM) classes corresponding to (A) 3, (B), 14, and (C) 28 days of aging

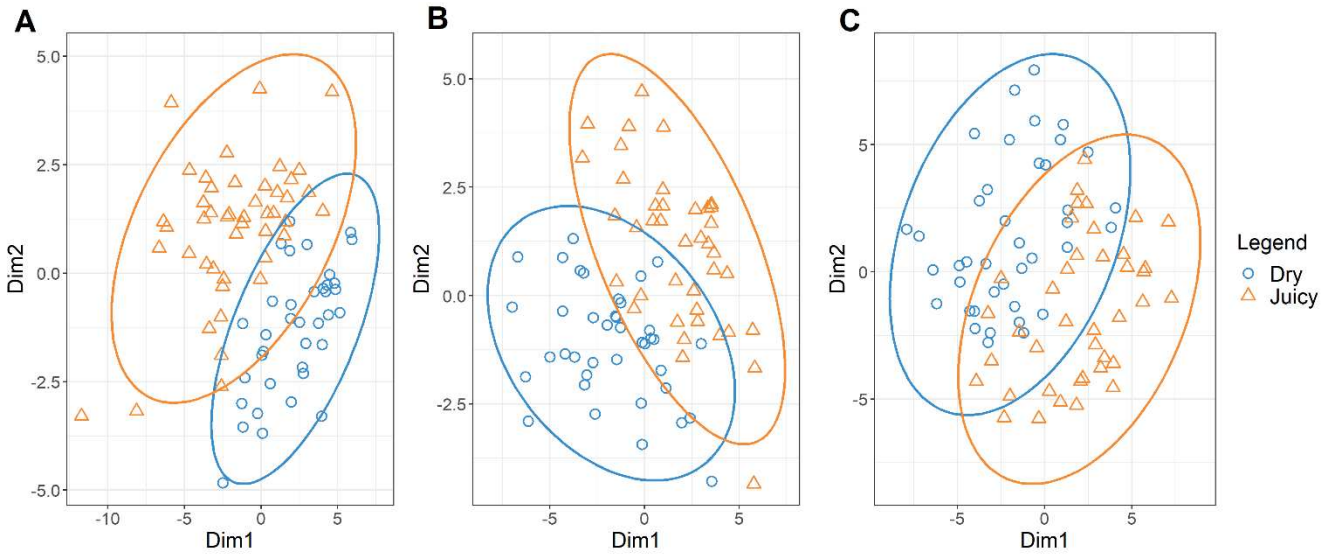

Figure 7. Partial least squares-discriminant analysis (PLSDA) plot of juiciness of striploin (LM) classes corresponding to (A) 3, (B), 14, and (C) 28 days of aging

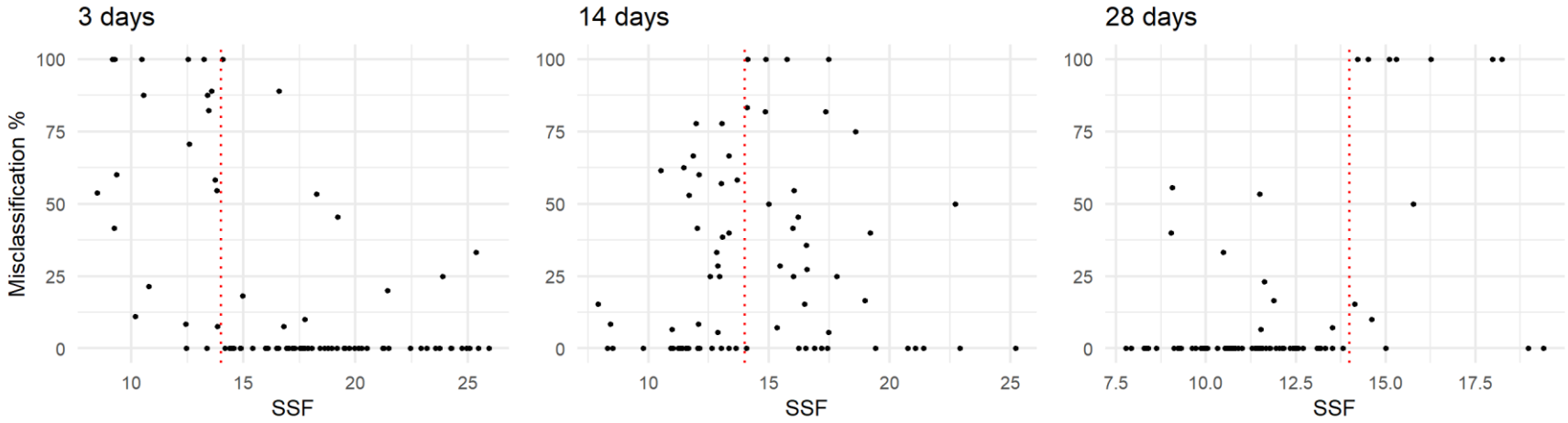

Figure 8. Slice shear force (SSF) versus the proportion of misclassification of each observation after performing 50 random iterations of training/testing (10-fold cross-validation) using the best tuning parameters of the top predictive models of striploins (LM).

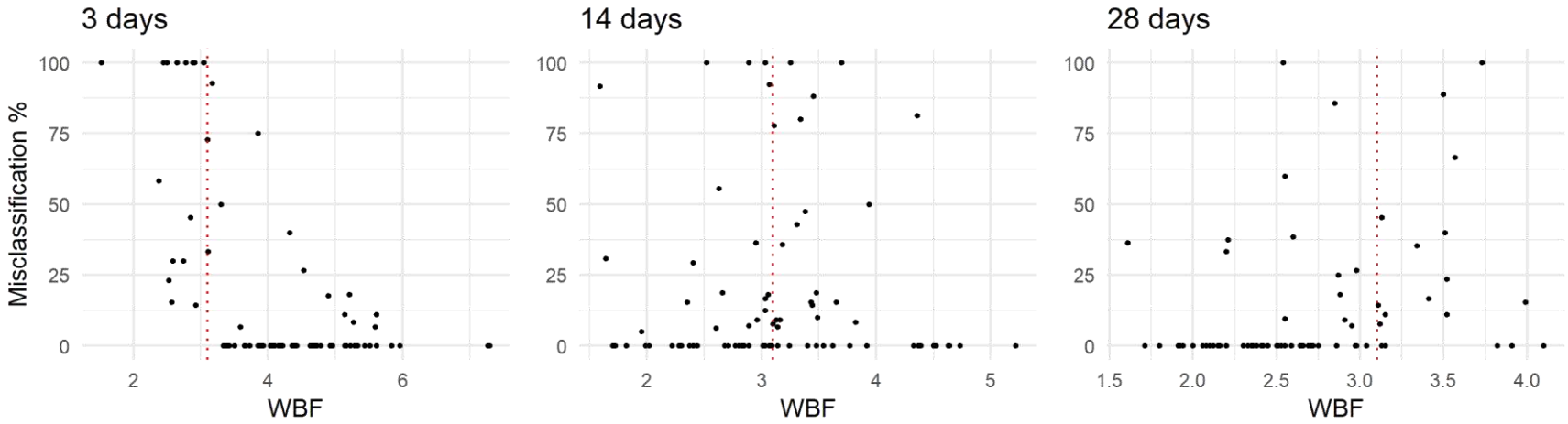

Figure 9. Warner-Bratzler shear force (WBF) versus the proportion of misclassification of each observation after performing 50 random iterations of training/testing (10-fold cross-validation) using the best tuning parameters of the top predictive models of striploins (LM).

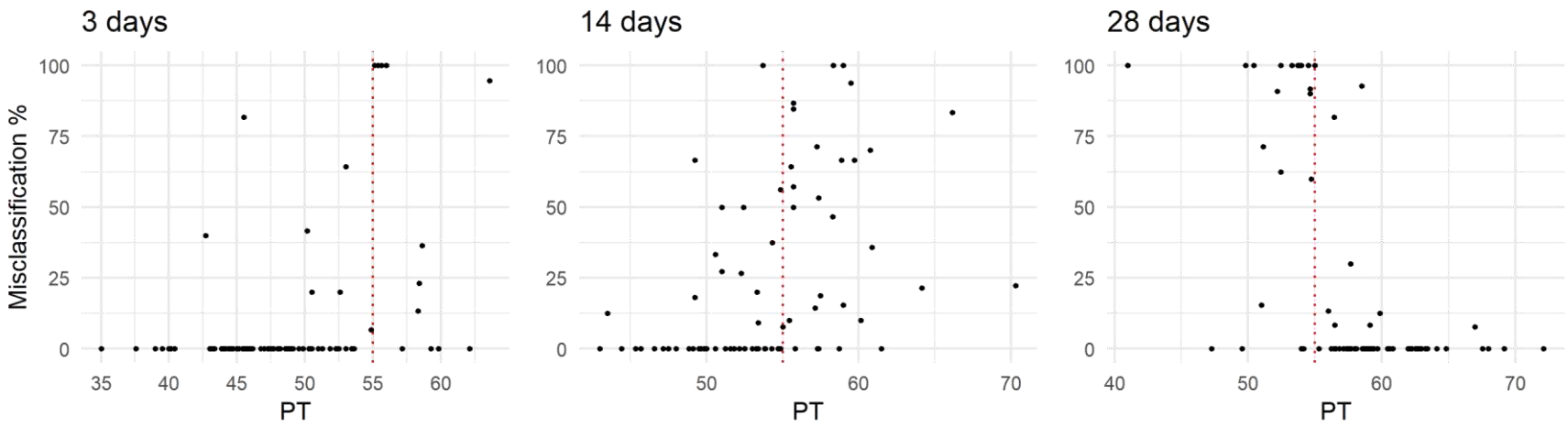

Figure 10. Sensory panel tenderness (PT) versus the proportion of misclassification of each observation after performing 50 random iterations of training/testing (10-fold cross-validation) using the best tuning parameters of the top predictive models of striploins (LM).

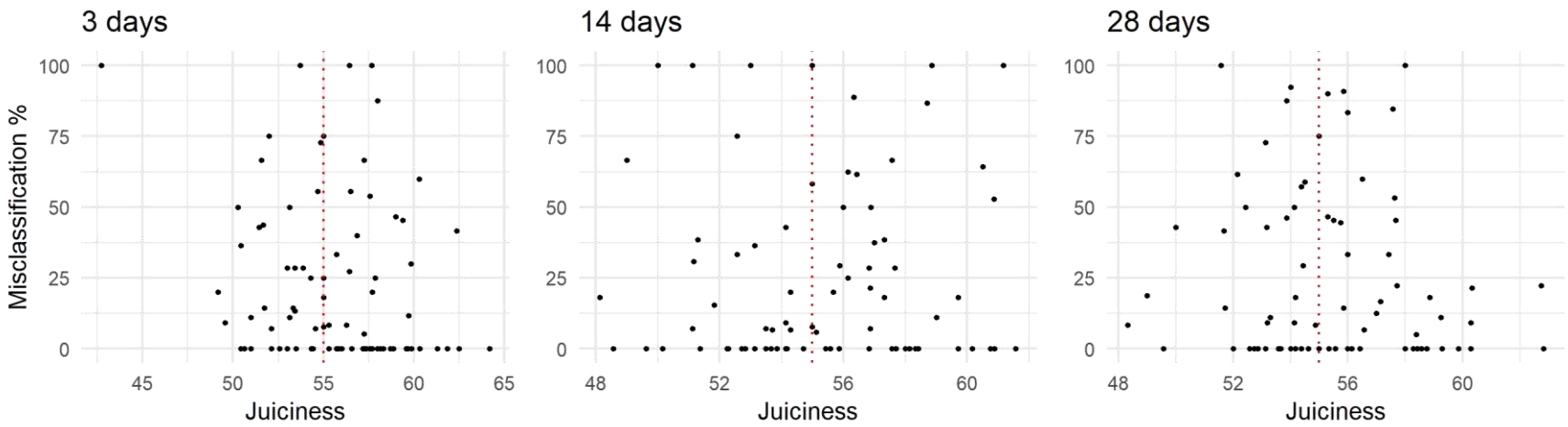

Figure 11. Juiciness score versus the proportion of misclassification of each observation after performing 50 random iterations of training/testing (10-fold cross-validation) using the best tuning parameters of the top predictive models of striploins (LM).

CHAPTER III: REPEATABILITY AND REPRODUCIBILITY OF RAPID EVAPORATIVE IONIZATION MASS SPECTROMETRY FOR BEEF SAMPLES

Introduction

Rapid Evaporative Ionization Mass Spectrometry (REIMS) is an ambient mass spectrometry (MS) technique that can generate a mass spectral fingerprint of biological tissues in real-time and in situ (Balog et al., 2010). Compared to conventional MS methods (e.g., LC-MS/MS or GC-MS/MS), REIMS does not require sample preparation which enables rapid collection of valuable chemical information at a lower cost and with a low level of technical expertise (Verplanken et al., 2017). These benefits have raised interest in evaluating REIMS for multiple applications in the beef industry. Studies have demonstrated that REIMS has the potential to be used as a real-time tool for applications in food provenance and fraud, processing systems, and meat quality. For example, the chemometric fingerprint obtained with REIMS can differentiate between beef, horse, and venison meat with almost perfect accuracy and can detect adulteration of ground beef with horse meat or offals to levels of 5% (Balog et al., 2016; Black et al., 2019). Predictive models using metabolomic profiling can classify beef carcasses into tenderness categories with 91% accuracy (Gredell et al., 2019). REIMS can also be implemented as a verification process of marketing programs based on the breed of animals or production systems. Chemometric profiling can differentiate at breed level with up to 97% accuracy and between organic production systems from conventional with 84% accuracy (Balog et al., 2016; Gredell et al., 2019; Robson et al., 2022). Since REIMS is a non-destructive technique that provides rapid chemical information, it is a viable alternative as a control point to verify beef attributes.

Repeatability, reproducibility, and robustness are necessary attributes for all new techniques to cross over from bench-top experiments to plant implementation. A measurement method is repeatable when it can provide consistent measurements on the same sample measured with the same instrument and operator and within the same day. Reproducibility represents the variations in measurements obtained with different devices and operators in different locations and times (Tabb et al., 2010). Robustness represents the resilience of an analytical technique to variation in measurement conditions without affecting the output (Ferreira et al., 2017). While robust methods are desirable for laboratory applications, robust processes are almost indispensable for industrial applications. In general, experiments carried out in laboratories have controlled conditions, and sources of variability are reduced. On the other hand, controlling the source of variation without affecting the efficiency of production processes is generally challenging, so robustness is an important attribute to consider. Reproducing results obtained in laboratory experiments with REIMS into industrial applications will depend on the stability of the data over time with different operators and equipment. Furthermore, understanding the reliability of results to data variations due to the factors mentioned above will increase confidence in this technology, facilitating its implementation and opening to new studies. Therefore, this study aims to evaluate the stability of REIMS data within and between days, the reproducibility of the data collected with different instruments, sampling methods and operators, and its reliability in reproducing results with data variation in beef applications.

Materials and Methods

Beef samples and quality control samples

The repeatability and reproducibility of REIMS in beef applications were analyzed using REIMS data of samples from the procedure described in Chapter II and quality control samples.

Briefly, 42 carcasses USDA Select (marbling scores Slight^{00} – Slight^{99}), and 42 upper 2/3rd Choice (marbling scores Modest^{00} - Moderate^{99}, n = 42, N=84) were randomly selected in a commercial facility for this study. Both striploins (LM) and tenderloins (PM) were collected around 36 h post-mortem, cut into portions, and each section was randomly assigned to 6 aging times (3, 14, 28, 42, 56, and 70 days). In addition, two sets of slivers from both sides of the *longissimus dorsi* muscle between the 12-13th rib (GR samples) were collected at the grading time to be analyzed by two laboratories (Colorado State University [CSU] and Texas Tech University [TTU]).

QC samples were made from ground beef (80% lean, 20% fat) acquired from the CSU Meat Lab facility. Ground beef was homogenized for 15s using a food processor (Blixer 6V, Robot Coupe, MS), mixed manually, and homogenized for a second time with the blender. Then, the homogenized product was formed into patties. Each patty was individually vacuum-packed and stored at -20°C until further analysis.

REIMS data collection

Chemometric profiles of the samples were obtained using a quadrupole time-of-flight (ToF) mass spectrometer (Synapt G2-Si Q-ToF, Waters Corporation, UK) equipped with a REIMS source (Waters Corporation, Milford, MA). An electronic probe (MedRes Medical Research Engineering Ltd., Budapest, Hungary) powered by an electrosurgical generator (Erbe VIO 50C, Erbe Elektromedizin GmbH, Tubingen, Germany) was used as the sampling device. The electrosurgical generator was set to dry cut mode and maximum cutting power of 40W. Mass spectra from 50-1,500 m/z were acquired in negative ion mode, with cone voltage at 40V heater bias at 60V. Samples were thawed at 0-4°C for 16-24 h and randomly sorted for REIMS analysis. The sampling device was used to generate at least five burns over the surface of individual samples. Burns were made in the four corners, and the middle of a 2.5x2.5cm square from the surface of

the sample, and each burn lasted approximately 1s. A 40µg/L of leucine-enkephalin/isopropanol solution was injected directly into the REIMS sources (flow rate: 200µg/min) for lock mass correction.

REIMS data were pre-processed with AMX Recognition software (version 1.0.2184.0, Waters Corporation). This process included lock mass correction using leucine-enkephalin (554.2615m/z), background subtraction using standard Masslynx pre-processing algorithms, total ion current normalization, peak binning in 0.5 or 0.1m/z intervals, and exclusion of bins in the range of 550-600m/z.

REIMS repeatability and reproducibility

All samples, including GR (n=168), LM and PM samples of all aging treatments (n=1008), and QC samples (n=29), were analyzed at CSU on 5 continuous days, including two batches per day. The ion transfer device of the REIMS instrument was cleaned every day before analysis, and the venturi and transfer capillary was cleaned between batches. The GR samples were analyzed during day 1, while the LM and PM samples were randomly analyzed during the remaining days. QC samples were analyzed at the beginning, middle, and end of each batch.

The evaluation of reproducibility of REIMS between labs, instruments, technicians, and sampling devices was performed with the two sets of GR samples. One set of samples was analyzed during day 1 at CSU following the protocol described before, and the second set was analyzed at TTU using the same protocol used at CSU but with a different MS analyzer brand (Xevo G2-S quadrupole ToF, Waters Corp., Wilmslow, UK), technician, and using a monopolar electrosurgical as sampling device (iKnife, Waters Corporation, Milford, MA). Both sets of samples were analyzed in both labs in one day to eliminate a possible day effect of the instrument. Since the optimal cutting power of each sampling device is different due to different geometries,

the cutting power used in TTU was 50W to get an optimal ionization. Aged LM and PM samples were randomly analyzed from day 2 to day 5 to evaluate the day effect and robustness of REIMS. QC samples were analyzed at the beginning, middle, and end of each batch to assess the stability of REIMS. The measurement stability represents the instrument's ability to provide consistent measures or with low variability over time.

Statistical analysis

All data analyses were performed using R statistical computing software (RStudio Team, version 2022.2.1.461, 2022). REIMS data of QC samples were used to evaluate the stability of the data across multiple batches and days. Repeatability and reproducibility were evaluated using a random-effects model of the top 5 most abundant peaks fitted with the *lmer* function from the *lmer4* R package (Bates et al., 2015). Models included the relative abundance of the greatest mass bins as response variables with batches nested within collection days as random effects. Best linear unbiased predictions were calculated with the *ranef* function from the *lmer4* R package. The *rand* function from the *lmerTest* R package (Kuznetsova et al., 2017) was used to test the null hypothesis of variance equal to 0 using likelihood ratio tests for random effects. Coefficients of variation (CV) between burns of all mass bins in the range 50-1500m/z with a binning of 0.1m/z were calculated across the 29 QC samples to analyze the variability of the data.

Interlab reproducibility of REIMS was evaluated with principal component analysis (PCA) of the GR data from the two labs. The *PCA* function from the *FactoMineR* package (Lê et al., 2008) with a Boolean normalization of the data was used to perform PCA. The function *fviz_pca_ind* from the *factoextra* package (version 1.0.7, 2020) was used to visualize the PCA and build the 95% confidence ellipses. The robustness of REIMS to differences in the data between labs and the reproducibility of multiple machine learning algorithms were evaluated with

predictive models for juiciness, tenderness, and flavor on days 3, 14, and 28 of aging. Models were trained with REIMS data of the GR collected at CSU using three dimensionality reduction methods (PCA, feature selection [FS], and PCA-FS) and fifteen machine learning algorithms. Models were trained and tuned with the *train* function from the *caret* R package (Kuhn, 2008) using 10-fold cross-validation. Then, tuning parameters of the top CSU models of each attribute and aging time were used to re-train the models with CSU data and test them with CSU and TTU data using 10-fold cross-validation. The last step was performed manually since the *train* function does not allow cross-validation between different data sets. The same random seed was fixed to train and tune the models the first time and to train the model the second time. A random seed is a pseudorandom generator that ensures consistent results of machine learning algorithms (Gundersen et al., 2022b). Since splitting the data into training/test sets and, in some cases, steps of the machine learning algorithms (e.g., random feature selection in decision tree-based algorithms) depend on random generators, fixing a random seed ensures exact results if the data and procedure are equal (Gundersen et al., 2022a). In this analysis, even when the same random seed was used to train the first and second models with CSU data, there is no guarantee that the results of the first and second ones are the same. This occurs because, in the first step, the models were trained and tuned, and the random generator changes with each iteration of the tuning process, while in the second step, the models were only trained once. Tracing back the random generator through the tuning process to use the same random seed on the second step was not straightforward, so the results of the first and second models were reported.

LM and PM data were used to analyze the day effect of REIMS. PCA evaluated how data were clustered by collection days, muscle type, and aging period. The robustness of REIMS to the day effect of the instrument was assessed by building predictive models for muscle type and aging

period. For this purpose, sparse partial least squared discriminant analysis (sPLSDA) models were trained using the *tune.splsda* function from the *mixOmics* package with 10-fold cross-validation.

Results and discussion

Stability of REIMS data

Variability of the mass bins (bin size = 0.1m/z) representing 90% of the total ion current (TIC) in a range of 50-1500m/z of the QC samples (n=29) is presented in Figure 12. The CV of mass bins ranged from 0.07 to 0.98, with CV of the most abundant mass bins lower than 0.30 (Figure 12B). Features with low variability between burns were mostly located in the 50-1000m/z range, while most mass bins between 1000-1500m/z had CV greater than 0.30 (Figure 12C). The number of features that accounted for 90% of the TIC with CV less than 0.30 was 1460, while the number of features with CV less than 0.20 was 466. None of the mass bins representing the remaining 10% of the TIC had CV less than 0.20, and only 28 had CV less than 0.30. Ross et al. (2021) found similar results evaluating the CV between burns of 175 beef samples. They found that low-intensity mass bins predominantly have lower measurement stability. In contrast, they found that the most stable features were between 50-1600m/z, while in the current study, the more stable features were between 50-1000m/z, but these differences could be due to different sampling devices. Previous research reported that peaks between 100-350 tentatively correspond to fatty acids and peaks between 600-1100mz correspond to glycerophospholipid and triglyceride in meat samples (Verplanken et al., 2017; Genangeli et al., 2019; Ross et al., 2021). Consequently, it could be interpreted that REIMS is more suitable for providing chemical information on lipids found in beef samples.

Measurement stability of the top 5 abundance mass bins across ten batches and five collection days are presented in Table 7 and 8. The top most abundant mass bins in QC samples

were 281.25, 279.25, 726.55, 742.55, and 283.25, corresponding to exact mass peaks 281.2399, 279.2256, 726.5475, 742.5462, and 283.2558. Putative identification of those peaks using Lipidmaps (www.lipidmaps.org) suggests that 279.2256, 281.2399, and 283.2558 might correspond to the fatty acids and conjugated class, probably deprotonated fatty acids C18:2, C18:1, C18:0, respectively. Ions 726.5475 and 742.5462 might be compounds in the family of glycerophosphoethanolamines. Other authors have previously identified similar mass bins as relevant in the chemometric profiling of beef samples (Balog et al., 2016; He et al., 2021; Zhang et al., 2021). The likelihood ratio test for both random effects (Tables 7 and 8) demonstrated agreement between measures taken in different batches and days for all top mass bins (P<0.05).

REIMS interday reproducibility

The PCA of LM and PM samples (n=504) corresponding to two cuts and six aging periods (3, 14, 28, 42, 56, and 70 days) are presented in Figure 13. The data were clustered by the collection days on the first two dimensions of the PCA, which represents 52% of the data variability (Figure 13A). Data collected during days 2 and 3 (d2 and d3) seem more dispersed than data collected on days 4 and 5 (d4 and d5), suggesting higher variability on the first days than the later ones. Clustering by muscle type was observed in dimension 4 of the PCA (Figure 13D), while clustering by aging times was not observed in the first ten dimensions. These observations in the PCA suggest that collection day has a stronger effect on the variability of REIMS data than muscle types and aging times.

Higher variability (greater dispersion of the data in the PCA plot) of d2 and d3 compared with d4 and d5 may be due to the replacement of the heated impactor or an improvement in the sampling technique (Figure 13A). Almost at the end of d3, the heated impactor of the MS analyzer failed, and it was replaced for d4 and d5. The heated impactor is an electronic component of the

MS instrument source responsible for the ionization of the molecular species. Therefore, it is possible that when the heated impactor was about to fail, it induced variability in the mass spectra, which is a possible scenario when working with this equipment. Higher variability on the first days could also be due to improvement in the sampling technique. Since aerosol generation during sampling is a thermally dependent process, changes that influence heat transfer impact the vapor generation. He et al. (2021) demonstrated that changing the cutting speed from 1mm/s to 5mm/s could increase the peak intensity of some m/z up to 3.5 times. Bodai et al. (2018) evaluate multiples electrode geometries to optimize the sampling technique for species-level classification models for microorganisms. In their study, changes in the electrode geometries and heating power impacted the cross-validation accuracy of the models radically, from 77% to 93% accuracy in negative mode and 33 to 91% in positive mode. While attempts were made to minimize variation in sampling techniques, non-controlled changes in the techniques, such as pressure applied during sampling or the sampling device's holding angle, may contribute to the variability between the first and last days. Therefore, it is important to consider these factors when developing models using REIMS data.

Although REIMS data showed a cluster by collection day (Figure 13A), the performance of classification models for muscle type was not affected by the day structure of the data. The ability of REIMS to segregate samples by muscle types and aging periods across collection days is reported in Table 9. The accuracy of the model of muscle type across all aging times and collection days was 98.1% (Table 9). Performances of muscle type models for specific collection days (d2 and d5) or aging periods were similar to the full model showing accuracies between 95.8 to 98.4%. These results were expected since the LM and PM data were clustered by muscle type in dimension 4 of the PCA (Figure 13D). The LM muscle has less oxidative fiber, lipid content,

and different fatty acid composition than the PM muscle, explaining the differences in metabolic profile observed in the PCA (Pavan and Duckett, 2013; Kim et al., 2021). Abraham et al. (2017) also found similar results when comparing metabolites found in the two muscles aged 0, 3, and 7 days using GC-MS analysis. The metabolic profiles of the two muscles were different and clustered by muscle when a PCA of all the metabolites was performed.

The performance of predictive models of aging time depends on the differences in metabolic fingerprints of the classes and the model complexity. Model of aging including all the samples, only LM, and only PM have poor performances (48.5%, 54.9%, and 45.4% accuracy, Table 9). However, models to predict the two most distant aging periods (3 vs. 70 days) across the two muscle types, LM, and PM, showed almost perfect performances with 99.9%, 100%, and 99.3% accuracies, respectively. Models to classify samples into 3 vs. 28 days of aging classes also showed acceptable accuracies, but the accuracy of the models dropped when aging classes were closer (for example, classification models for 3 vs. 14 days). The differences in the metabolic profiles of both muscles increase with time (Abraham et al., 2017). Therefore, increasing the time between aging classes increases the accuracy of the model. Even when classification models for 3 vs. 28 days and 3 vs. 70 days of aging had high performance, the model performance dropped with model complexity when an extra class was included. Multiclass models are prone to higher errors than binary classifications since the classifiers need to separate data into more categories, increasing error chances (Lorena et al., 2009). In addition, most machine learning algorithms are designed for binary classification, and multicategory classification results from decomposing the classes into binary subproblems. Fusion of those algorithms is challenging and can contribute to error (Rifkin and Klautau, 2004; García-Pedrajas and Ortiz-Boyer, 2011).

Even when REIMS data vary with collection days, it can still be used to build acceptable predictive models if classes are different. Previous research demonstrated that REIMS is a robust technology that can classify samples independently when the data is collected. In the medical field, REIMS data collected over multiple days have been used to classify cancer cells from normal cells with almost perfect accuracies (Balog et al., 2013; Phelps et al., 2018). Similarly, Verplanken et al. (2017) found that classification models of tainted and untainted boars based on REIMS from neck fat were not affected by the collection day. However, they found that the accuracy of models with three classes (tainted, untainted boars, and sows) was reduced from 99 to 89% when data were collected in one day versus collected in three days and variations in cone voltage.

REIMS interlab reproducibility and statistical reproducibility of predictive models

Figure 14 shows the PCA of the two sets of GR samples analyzed by two different labs, brands of MS analyzer, operators, and types of sampling devices. The first two dimensions of the PCA represent 53.8% of the variability. The PCA plot showed that the REIMS data collected at CSU differed from the data collected at TTU. Predictive models of tenderness, juiciness, and flavor were built using CSU data and tested with TTU data to evaluate the robustness of REIMS against differences in the data between labs. Predictive models were trained and tuned with 10-fold cross-validation using only CSU data, and the maximum accuracies were reported in Table 10. Then, top models and their tuning parameters were used to re-train the models with CSU data and test them with CSU and TTU data. Models were validated with 10-fold cross-validation, and accuracies, sensitivities, and specificities of re-tested CSU data and test of TTU data are reported in Table 10. FS as dimensionality method yielded the best performance in all the models, except for some SSF and WBF models. None of the best models included PCA, and only SSF of 14 and 28 days and WBF of 28 days had PCA-FS as the best dimensionality reduction method.

The accuracies of the first models obtained from the tunning process (max accuracy, Table 10) were not reproducible after re-training the model with CSU data (accuracy CSU, Table 10). These results were expected due to the stochastic nature of machine learning algorithms and since the top CSU models and the re-tested models used different random seeds for splitting the data into train/test data sets or random selection of some algorithms changed. A random seed is a pseudorandom generator that allows getting the exact outcome of the same model using the same data by reproducing the generation of random numbers in the machine learning algorithm (Gundersen et al., 2022b). Only accuracies of SSF d28 and WBF d14 models were higher in the re-test, while the performance of PT d28 was maintained. Controlling the reproducibility of machine learning algorithms has been a concern due to the great variability of results that can lead to different conclusions (Islam et al., 2017; Pineau et al., 2020; Gundersen et al., 2022b). Machine learning algorithms involve algorithm, implementation, and observational factors that depend on randomness and impact results reproducibility (Gundersen et al., 2022a). Reproducibility of the results of the algorithms used for this study that were subject to randomness, including splitting of the data into training and test sets, hyperparameters optimization, and feature selection in the particular cases of extreme gradient boosting (XGBoost) and random forest (RF). Therefore, even when the optimized parameters of the first models were used to re-test the same data, the results were not reproducible because of random splitting data and feature selection on XGBoost and RF models. Reproducibility of statistical models does not apply only to REIMS but to studies where machine learning is used. Some authors suggest that setting random seeds and sharing data, code, and results (Beam et al., 2020; Pineau et al., 2020; Heil et al., 2021) will help to improve reproducibility. However, seeding a random generator would lead to a generalized conclusion of a specific random seed (Bouthillier et al., 2019). Therefore, other approaches recommend reporting

error rates with confidence intervals or including sensitivity analysis for model variations (Gardner et al., 2018; Henderson et al., 2018). Implementing these practices for reporting REIMS results could increase the transparency and interpretability of the results.

The models built with CSU data could not predict sensory classes of TTU data. All CSU models performed better when tested with CSU data than tested with TTU data, except for the WBF d28 model (Table 10). SSF d3 classified all TTU samples as tender and could not recognize tough samples, while SSF d28 and PT d3 models classified all samples as tough and could not identify tender samples. The only model that showed a certain ability to segregate TTU samples correctly was the WBF d28, with an accuracy of 81.3% but displayed 42% specificity. The remaining models resulted in low accuracy, corroborating the findings of Figure 14 that REIMS data collected at CSU differed from TTU data.

It is unclear if the data differences are solely due to different MS analyzers, technicians, sampling devices, or a combination of these factors. Previous studies have reported that REIMS is reproducible across MS instruments. Strittmatter et al. (2014) evaluated the inter-instrument reproducibility of REIMS to classify nine strains of microorganisms. They found that the classification accuracy of the model built from data collected with an Orbitrap instrument was 100%, while the accuracy of the model from a Xevo Q-TOF analyzer was 97.8%. They also showed that samples in the recursive maximum margin criterion (RMMC) plot was dominated by microorganisms strains even when the data were collected with different instruments. The last findings could suggest that variation between CSU data and TTU might be due to the differences in devices or technicians since different analyzers yielded similar results. However, RMMC is a supervised algorithm that maximizes interclass separability (Veselkov et al., 2014), explaining

why samples were clustered by strains and not by instruments. In addition, model cross-validation was performed within MS analyzers and not between instruments as in the present study.

Other studies have demonstrated that the cutting speed, cutting power, and geometry of the sampling device blade influence REIMS chemical profiling (Bodai et al., 2018; He et al., 2021). The ionization process is heat sensitive, and therefore, differences in techniques by technicians and different ionization devices may generate differences in the REIMS data between labs. At the same time that a spray of molecules is produced, chemical reactions of compounds present in the samples occur. Therefore, differences in exposure time and temperature can cause variations in heat-sensitive compounds or heat-catalyzed reactions, resulting in variations in MS readings. Genangeli et al. (2019) collected REIMS data from multiple animal tissues with a diathermic knife and CO_2 laser and evaluated species and animal age classification models with cross-validation between instruments. In contrast to the present study, they found that data collected with one sampling device can be used to predict species and age classes from data generated from the other one. However, they compared two sampling devices with the same MS analyzer, and models were between well-differentiated classes (for example, cow vs. calf or cow vs. chicken), while in the present study, different MS analyzers were used, and models were between the same species with similar ages.

Conclusions

REIMS is a rapid technique that can provide valuable chemical information in-situ of meat tissues. It has the potential to be implemented as a real-time verification method for multiple purposes if it can provide consistent and reliable measures. This study has demonstrated that REIMS is a repeatable technique that provides chemical information of beef samples with good stability of the measures within days, especially for lipids compounds. Although there was an

evident cluster of REIMS data by collection day, it does not limit REIMS capacity to produce accurate classification models for beef cuts and aging time when classes are distant.

The data obtained from two instruments, technicians, and types of sampling devices were different. The construction of a universal predictive model that allows evaluating meat quality attributes (tenderness, juiciness, and flavor) with data from different instruments, operators, and sampling devices was not possible. However, it was not possible to determine if REIMS was not reproducible due to multiple factors or an individual factor because they were not evaluated separately. Further studies are required to determine the individual effect of instruments, technicians, and sampling devices on REIMS reproducibility to develop strategies to improve the interlab reproducibility of this technology.

The reliability of REIMS-based predictive models to interday variations depends on the differences between classes and the complexity of the models. Since this study of the LM and PM samples was performed on five continuous days, these results are limited to short periods. The robustness of REIMS to interday variations is limited to the models evaluated. However, these results generate inputs on the importance of evaluating REIMS models with data collected on different days. Understanding the interday reproducibility of REIMS in industrial applications will require evaluating this technology for longer periods of continuous operation to simulate industrial practices.

Table 7. Best linear unbiased predictions of the relative abundance (percent) of the top 5 more abundant peaks found in REIMS data of quality control samples corresponding to each batch

Mass bins	B1[1]	B2	B3	B4	B5	B6	B7	B8	B9	B10	σ_{batch}^2	p-value
281.25	5.27	5.90	5.53	5.74	5.68	5.11	5.69	5.59	6.09	5.73	<0.01	0.63
279.25	2.19	2.19	2.19	2.19	2.19	2.19	2.19	2.19	2.19	2.19	<0.01	1.00
726.55	1.50	1.57	1.55	1.53	1.54	1.55	1.56	1.55	1.55	1.56	<0.01	0.83
742.55	1.18	1.20	1.19	1.19	1.19	1.19	1.20	1.20	1.19	1.20	<0.01	0.93
283.25	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	<0.01	1.00

[1]B1-B10 correspond to the batches

Table 8. Best linear unbiased predictions of the relative abundance (percent) of the top 5 more abundant peaks found in REIMS data of quality control samples corresponding to each collection day

Mass bins	d1*	d2	d3	d4	d5	σ_{day}^2	p-value
281.25	5.63	5.63	5.63	5.63	5.63	<0.01	1.00
279.25	2.19	2.19	2.19	2.19	2.19	<0.01	1.00
726.55	1.34	1.50	1.52	1.71	1.66	<0.01	0.24
742.55	1.10	1.17	1.18	1.28	1.24	<0.01	0.43
283.25	1.10	1.10	1.10	1.10	1.10	<0.01	1.00

*d1 to d5 correspond to the collection days of the samples

[a-c] Least-squares means in the same row lacking a common superscript differ ($P < 0.05$)

Table 9. Accuracy of 10-fold cross-validation of sparse partial least squares-discriminant analysis (sPLSDA) classification models for cut type or aging period across combinations of cut type, aging period, and REIMS collection day.

Classification model	Muscle type	Aging[1]	Collection day[2]	N. samples	Accuracy
Muscle	All	All	All	1001	98.1%
Muscle	All	3 days	All	167	97.7%
Muscle	All	28 days	All	165	98.1%
Muscle	All	70 days	All	168	95.8%
Muscle	All	3 vs. 14 days	All	335	95.6%
Muscle	All	3 vs. 28 days	All	332	97.5%
Muscle	All	3 vs. 70 days	All	335	96.0%
Muscle	All	All	d2	240	98.4%
Muscle	All	All	d5	215	97.9%
Aging	All	All	All	1001	48.5%
Aging	All	3 vs.14 days	All	335	73.8%
Aging	All	3 vs. 28 days	All	332	97.8%
Aging	All	3 vs. 70 days	All	335	99.9%
Aging	All	3 vs. 28 vs. 70 days	All	500	81.1%
Aging	Striploin	All	All	499	54.9%
Aging	Striploin	3 vs.14 days	All	168	81.5%
Aging	Striploin	3 vs. 28 days	All	165	98.7%
Aging	Striploin	3 vs. 70 days	All	168	100%
Aging	Striploin	3 vs. 28 vs. 70 days	All	249	85.7%
Aging	Tenderloin	All	All	502	45.4%
Aging	Tenderloin	3 vs.14 days	All	167	71.4%
Aging	Tenderloin	3 vs. 28 days	All	167	91.0%
Aging	Tenderloin	3 vs. 70 days	All	167	99.3%
Aging	Tenderloin	3 vs. 28 vs. 70 days	All	251	80.1%

[1] Aging period contains six levels 3, 14, 28, 42, 56, and 70 days.

[2] Collection day contains four levels d2, d3, d4, and d5, corresponding to each collection day

Table 10. Prediction accuracy of the top classification models of REIMS predictive models for beef tenderness, juiciness, and flavor trained with 10-fold cross-validation from Colorado State University (CSU) data and tested with CSU and Texas Tech University (TTU) data

Models[1]	Aging period[2]	Top model[3]	N of predictors	Max accuracy	Accuracy		Sensitivity		Specificity	
					CSU	TTU	CSU	TTU	CSU	TTU
SSF	d3	FS/GLMNET	155	85.8	82.5	70.0	63.3	0.0	95.0	100.0
	d14	PCA-FS/XGBoost	12	84.8	76.3	60.0	80.0	78.1	76.5	46.9
	d28	PCA-FS/XGBoost	3	91.5	91.7	75.0	100.0	90.5	66.7	0.0
WBF	d3	FS/XGBoost	14	86.5	76.8	51.8	14.3	23.8	93.1	57.3
	d14	FS/RF	325	82.4	85.7	49.0	64.3	64.3	95.2	42.9
	d28	PCA-FS/SVM Poly	44	94.8	78.4	81.3	81.6	90.7	67.1	42.0
PT	d3	FS/XGBoost	23	93.9	83.3	70.8	95.2	100.0	66.67	0.0
	d14	FS/XGBoost	46	87.4	81.4	44.3	91.3	26.8	61.50	65.8
	d28	FS/PLSDA	600	82.5	82.5	60.0	59.5	37.0	95.24	72.0
Flavor	d3	FS/GBM	181	86.7	83.8	33.8	89.6	11.4	78.8	73.8
	d14	FS/PLSDA	265	88.6	80.0	40.0	87.3	19.5	76.5	75.7
	d28	FS/XGBoost	17	81.0	76.3	42.5	73.2	64.3	82.1	28.0
Juiciness	d3	FS/XGBoost	18	85.0	75.0	57.5	73.0	64.0	79.8	50.8
	d14	FS/XGBoost	26	83.6	77.1	55.7	70.2	53.8	81.8	55.3
	d28	FS/XGBoost	61	86.3	76.3	53.8	74.2	80.7	79.3	31.7

[1] SSF, slice shear force; WBF, Warner-Bratzler shear force; PT, sensory panel tenderness

[2] d3, 3 days; d14, 14 days; d28, 28 days of aging

[3] FS, Feature selection; PCA-FS, principal component analysis - feature selection; GLMNET, Lasso and elastic-net regularized generalized linear models; XGBoost, extreme gradient boosting; RF, random forest; SVM, support vector machine; PLSDA, partial least square discriminant analysis; GLM, generalized linear model.

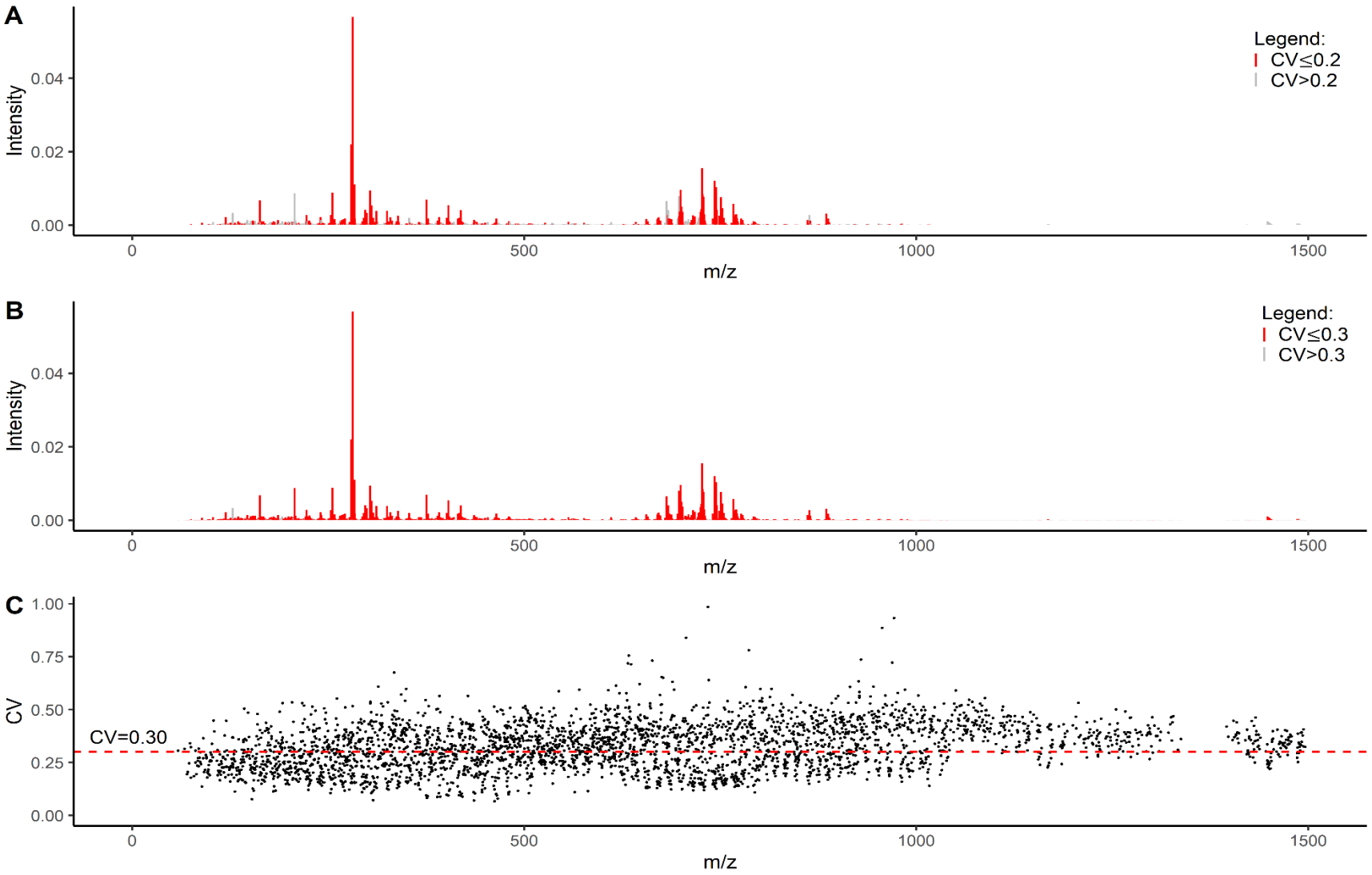

Figure 12. Variability of REIMS data mass bins representing 90% of the total ion current (TIC) over 50–1500 m/z with a bin size of 0.1 m/z of 29 ground beef samples. (A) Identification of mass bins with a coefficient of variation (CV) ≤ 0.2, (B) identification of mass bins with CV ≤ 0.30, and (C) Median CV between burns of mass bins representing 90% of total ion current (TIC)

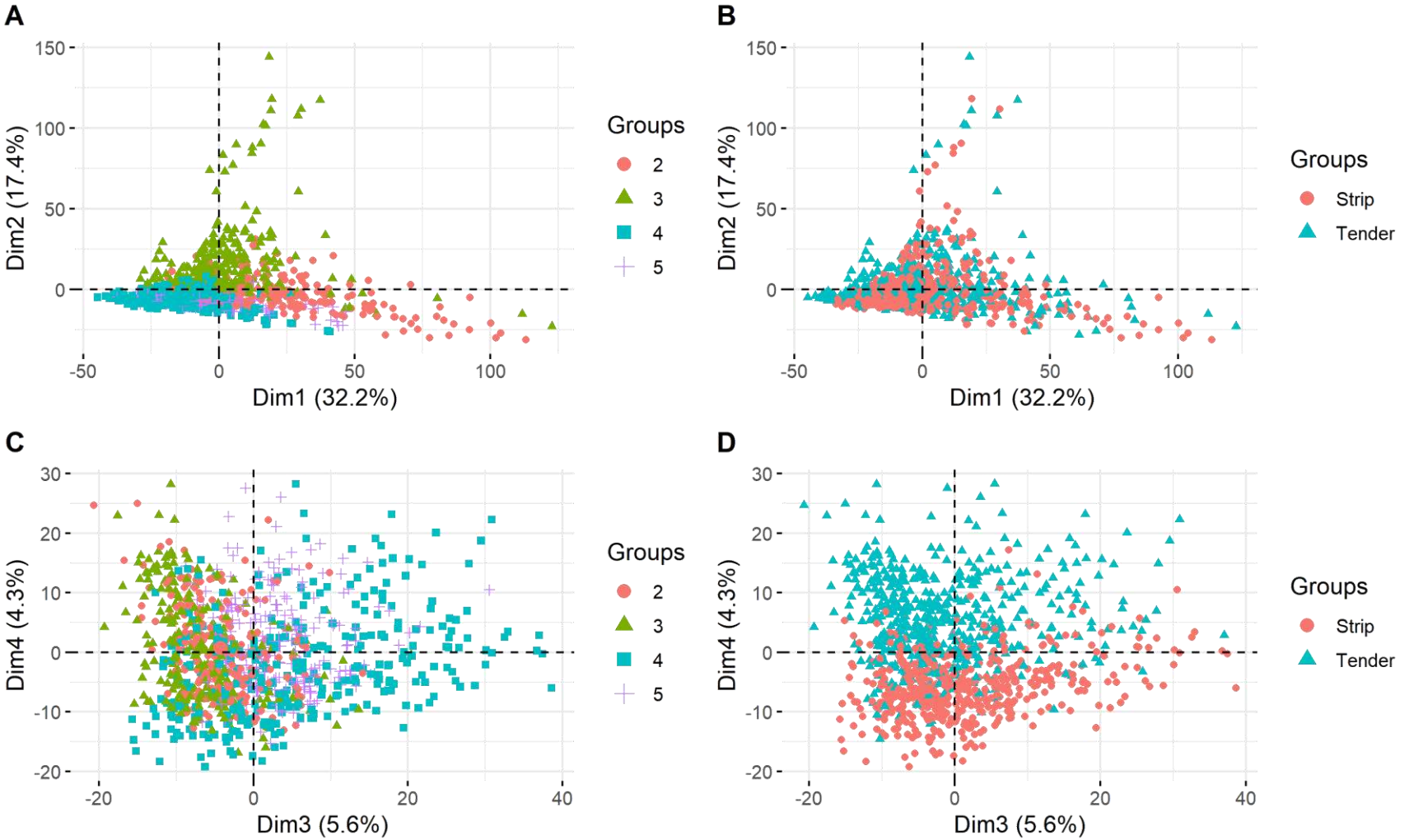

Figure 13. Principal component analysis (PCA) of REIMS data of striploin and tenderloin samples aged 3, 14, 28, 42, 56, and 70 days collected in 4 running days of REIMS instrumentation. (A) Dimensions 1 and 2 of the PCA colored by day, (B) dimensions 1 and 2 of the PCA color

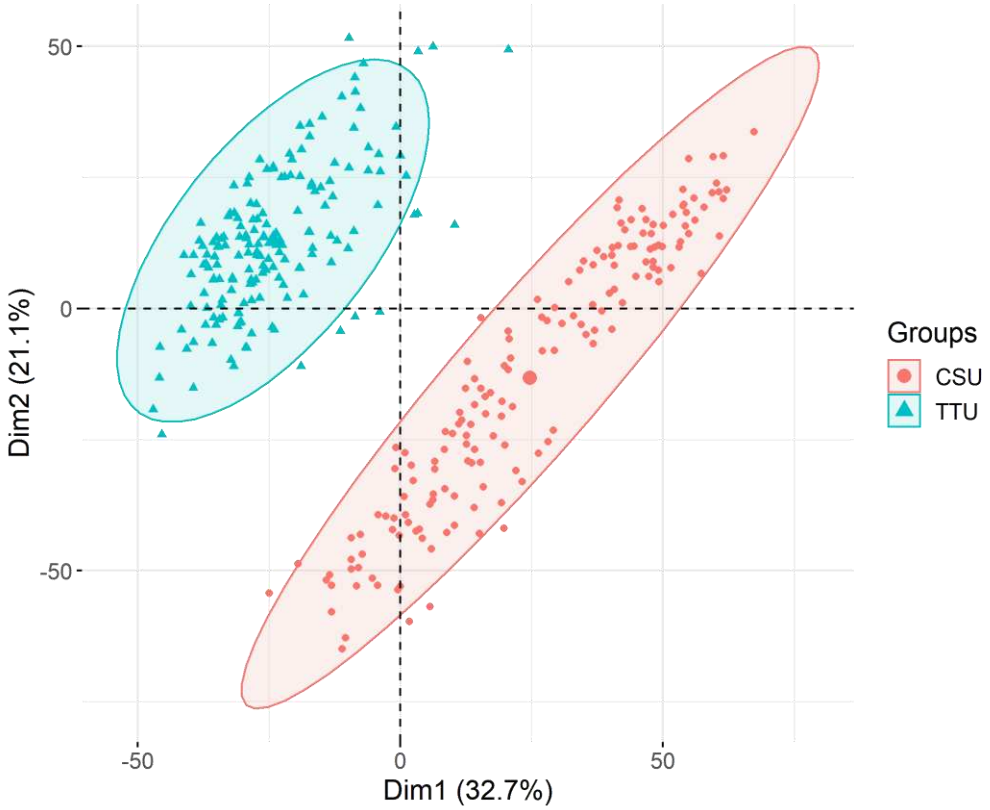

Figure 14. Principal component analysis of REIMS data of beef *longissimus dorsi* samples taken at the grading time (GR) collected at Colorado State University (CSU) with an electronic probe and a Synapt G2-Si mass spectrometer (MS) analyzer and collected at Texas Tech University (TTU) with an iKnife as sampling device and Xevo G2-S as mass spectrometer analyzer

REFERENCES

Abraham, A., J. W. Dillwith, G. G. Mafi, D. L. VanOverbeke, and R. Ramanathan. 2017. Metabolite profile differences between beef *Longissimus* and *Psoas* muscles during display. Meat Muscle Biol. 1. doi:10.22175/mmb2016.12.0007. Available from: https://www.iastatedigitalpress.com/mmb/article/id/9429/

Acheson, R. J., D. R. Woerner, and J. D. Tatum. 2014. Effects of USDA carcass maturity on sensory attributes of beef produced by grain-finished steers and heifers classified as less than 30 months old using dentition. J. Anim. Sci. 92:1792–1799. doi:10.2527/jas.2013-7553.

Adhikari, K., E. Chambers IV, R. Miller, L. Vázquez-Araújo, N. Bhumiratana, and C. Philip. 2011. Development of a lexicon for beef flavor in intact muscle. J. Sens. Stud. 26:413–420. doi:10.1111/j.1745-459X.2011.00356.x.

AMS. 2018a. Quality assessment division - fiscal year 2018 summary. Agricultural Marketing Service, USDA.

AMS. 2018b. United States standards for grades of carcass beef.

Aremu, O. O., D. Hyland-Wood, and P. R. McAree. 2020. A machine learning approach to circumventing the curse of dimensionality in discontinuous time series machine data. Reliab. Eng. Syst. 195:106706. doi:10.1016/j.ress.2019.106706.

ASTM. 2011. Standard specification for tenderness marketing claims associated with meat cuts derived from beef.

Balog, J., D. Perenyi, C. Guallar-Hoyas, A. Egri, S. D. Pringle, S. Stead, O. P. Chevallier, C. T. Elliott, and Z. Takats. 2016. Identification of the species of origin for meat products by rapid evaporative ionization mass spectrometry. J Agric Food Chem. 64:4793–4800. doi:10.1021/acs.jafc.6b01041.

Balog, J., L. Sasi-Szabó, J. Kinross, M. R. Lewis, L. J. Muirhead, K. Veselkov, R. Mirnezami, B. Dezső, L. Damjanovich, A. Darzi, J. K. Nicholson, and Z. Takáts. 2013. Intraoperative tissue identification using rapid evaporative ionization mass spectrometry. Sci. Transl. Med. 5:194ra93-194ra93. doi:10.1126/scitranslmed.3005623.

Balog, J., T. Szaniszlo, K.-C. Schaefer, J. Denes, A. Lopata, L. Godorhazy, D. Szalay, L. Balogh, L. Sasi-Szabo, M. Toth, and Z. Takats. 2010. Identification of biological tissues by rapid evaporative ionization mass spectrometry. Anal. Chem. 82:7343–7350. doi:10.1021/ac101283x.

Barlow, R. S., A. G. Fitzgerald, J. M. Hughes, K. E. McMillan, S. C. Moore, A. L. Sikes, A. B. Tobin, and P. J. Watkins. 2021. Rapid evaporative ionization mass spectrometry: a review on its application to the red meat industry with an Australian context. Metabolites. 11:171. doi:10.3390/metabo11030171.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. Fitting linear mixed-effects models esing lme4. J. Stat. Softw. 67:1–48. doi:10.18637/jss.v067.i01.

Batista, G. E. A. P. A., R. C. Prati, and M. C. Monard. 2004. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor. Newsl. 6:20–29. doi:10.1145/1007730.1007735.

Beam, A. L., A. K. Manrai, and M. Ghassemi. 2020. Challenges to the reproducibility of machine learning models in health care. JAMA. 323:305–306. doi:10.1001/jama.2019.20866.

Behrends, J. M., K. J. Goodson, M. Koohmaraie, S. D. Shackelford, T. L. Wheeler, W. W. Morgan, J. O. Reagan, B. L. Gwartney, J. W. Wise, and J. W. Savell. 2005. Beef customer satisfaction: Factors affecting consumer evaluations of calcium chloride-injected top sirloin steaks when given instructions for preparation1. Anim. Sci. J. 83:2869–2875. doi:10.2527/2005.83122869x.

Belk, K. E., M. H. George, J. D. Tatum, G. G. Hilton, R. K. Miller, M. Koohmaraie, J. O. Reagan, and G. C. Smith. 2001. Evaluation of the Tendertec beef grading instrument to predict the tenderness of steaks from beef carcasses. J. Anim. Sci. 79:688–697. doi:10.2527/2001.793688x.

Bhat, Z. F., J. D. Morton, S. L. Mason, and A. E.-D. A. Bekhit. 2018. Role of calpain system in meat tenderness: a review. Food Sci. Hum. 7:196–204. doi:10.1016/j.fshw.2018.08.002.

Birse, N., O. Chevallier, V. Hrbek, V. Kosek, J. Hajŝlová, and C. Elliott. 2021. Ambient mass spectrometry as a tool to determine poultry production system history: A comparison of rapid evaporative ionisation mass spectrometry (REIMS) and direct analysis in real time (DART) ambient mass spectrometry platforms. Food Control. 123:107740. doi:10.1016/j.foodcont.2020.107740.

Black, C., O. P. Chevallier, K. M. Cooper, S. A. Haughey, J. Balog, Z. Takats, C. T. Elliott, and C. Cavin. 2019. Rapid detection and specific identification of offals within minced beef samples utilising ambient mass spectrometry. Sci Rep. 9:6295. doi:10.1038/s41598-019-42796-5.

Black, C., O. P. Chevallier, S. A. Haughey, J. Balog, S. Stead, S. D. Pringle, M. V. Riina, F. Martucci, P. L. Acutis, M. Morris, D. S. Nikolopoulos, Z. Takats, and C. T. Elliott. 2017. A real time metabolomic profiling approach to detecting fish fraud using rapid evaporative ionisation mass spectrometry. Metabolomics. 13:153. doi:10.1007/s11306-017-1291-y.

Bodai, Z., S. Cameron, F. Bolt, D. Simon, R. Schaffer, T. Karancsi, J. Balog, T. Rickards, A. Burke, K. Hardiman, J. Abda, M. Rebec, and Z. Takats. 2018. Effect of electrode geometry on the classification performance of rapid evaporative ionization mass spectrometric (REIMS) bacterial identification. J Am Soc Mass Spectrom. 29:26–33. doi:10.1007/s13361-017-1818-5.

Bouthillier, X., C. Laurent, and P. Vincent. 2019. Unreproducible Research is Reproducible. In: Proceedings of the 36th International Conference on Machine Learning. Vol. 97. PMLR. p. 725–734. Available from: https://proceedings.mlr.press/v97/bouthillier19a.html

Boykin, C. A., L. C. Eastwood, M. K. Harris, D. S. Hale, C. R. Kerth, D. B. Griffin, A. N. Arnold, J. D. Hasty, K. E. Belk, D. R. Woerner, R. J. Delmore Jr., J. N. Martin, D. L. VanOverbeke, G. G.

Mafi, M. M. Pfeiffer, T. E. Lawrence, T. J. McEvers, T. B. Schmidt, R. J. Maddock, D. D. Johnson, C. C. Carr, J. M. Scheffler, T. D. Pringle, A. M. Stelzleni, J. Gottlieb, and J. W. Savell. 2017. National Beef Quality Audit–2016: in-plant survey of carcass characteristics related to quality, quantity, and value of fed steers and heifers. J. Anim. Sci. 95:2993–3002. doi:10.2527/jas.2017.1543.

Careri, M., F. Bianchi, and C. Corradini. 2002. Recent advances in the application of mass spectrometry in food-related analysis. J. Chromatogr. A. 970:3–64. doi:10.1016/S0021-9673(02)00903-2.

Casas, E., S. N. White, T. L. Wheeler, S. D. Shackelford, M. Koohmaraie, D. G. Riley, C. C. Chase Jr., D. D. Johnson, and T. P. L. Smith. 2006. Effects of calpastatin and μ-calpain markers in beef cattle on tenderness traits. J. Anim. Sci. 84:520–525. doi:10.2527/2006.843520x.

Christensen, C. M. 1984. Food texture perception. In: C. O. Chichester, E. M. Mrak, and B. S. Schweigert, editors. Advances in Food Research. Vol. 29. Academic Press. p. 159–199. Available from: https://www.sciencedirect.com/science/article/pii/S0065262808600579

Cieslak, D. A., and N. V. Chawla. 2008. Learning Decision Trees for Unbalanced Data. In: W. Daelemans, B. Goethals, and K. Morik, editors. Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg. p. 241–256.

Cross, H. R., M. S. Stanfield, and E. J. Koch. 1976. Beef palatability as affected by cooking rate and final internal temperature. J. Anim. Sci. 43:114–121. doi:10.2527/jas1976.431114x.

Di Rosa, A. R., F. Leone, F. Cheli, and V. Chiofalo. 2017. Fusion of electronic nose, electronic tongue and computer vision for animal source food authentication and quality assessment – a review. J. Food Eng. 210:62–75. doi:10.1016/j.jfoodeng.2017.04.024.

Dubost, A., D. Micol, B. Picard, C. Lethias, D. Andueza, D. Bauchart, and A. Listrat. 2013. Structural and biochemical characteristics of bovine intramuscular connective tissue and beef quality. Meat Sci. 95:555–561. doi:10.1016/j.meatsci.2013.05.040.

Ellis, M., M. s. Brewer, D. s. Sutton, H.-Y. Lan, R. c. Johnson, and F. k. Mckeith. 1998. Aging and cooking effects on sensory traits of pork from pigs of different breed lines. J. Muscle Foods. 9:281–291. doi:10.1111/j.1745-4573.1998.tb00661.x.

Emerson, M. R., D. R. Woerner, K. E. Belk, and J. D. Tatum. 2013. Effectiveness of USDA instrument-based marbling measurements for categorizing beef carcasses according to differences in longissimus muscle sensory attributes1. J. Anim. Sci. 91:1024–1034. doi:10.2527/jas.2012-5514.

Ferreira, S. L. C., A. O. Caires, T. da S. Borges, A. M. D. S. Lima, L. O. B. Silva, and W. N. L. dos Santos. 2017. Robustness evaluation in analytical methods optimized using experimental designs. Microchem. J. 131:163–169. doi:10.1016/j.microc.2016.12.004.

Foraker, B. A., D. A. Gredell, J. F. Legako, R. D. Stevens, J. D. Tatum, K. E. Belk, and D. R. Woerner. 2020. Flavor, Tenderness, and Related Chemical Changes of Aged Beef Strip Loins.

Meat and Muscle Biology. 4. doi:10.22175/mmb.11115. Available from: https://www.iastatedigitalpress.com/mmb/article/id/11115/

Gagaoua, M., V. Monteils, and B. Picard. 2018. Data from the farmgate-to-meat continuum including omics-based biomarkers to better understand the variability of beef tenderness: an integromics approach. J. Agric. Food Chem. 66:13552–13563. doi:10.1021/acs.jafc.8b05744.

García-Pedrajas, N., and D. Ortiz-Boyer. 2011. An empirical study of binary classifier fusion methods for multiclass classification. Inf Fusion. 12:111–130. doi:10.1016/j.inffus.2010.06.010.

Gardner, J., Y. Yang, R. Baker, and C. Brooks. 2018. Enabling end-to-end machine learning replicability: a case study in educational data mining. arXiv. Available from: http://arxiv.org/abs/1806.05208

Gareth, J., D. Witten, T. Hastie, and R. Tibshirani. 2021a. Statistical learning. In: An Introduction to Statistical Learning. Second. NY, USA. p. 15–52. Available from: https://hastie.su.domains/ISLR2/ISLRv2_website.pdf

Gareth, J., D. Witten, T. Hastie, and R. Tibshirani. 2021b. Linear model selection and regularization. In: An Introduction to Statistical Learning. Second. NY, USA. p. 225–321. Available from: https://hastie.su.domains/ISLR2/ISLRv2_website.pdf

Gareth, J., D. Witten, T. Hastie, and R. Tibshirani. 2021c. Resampling Methods. In: An Introduction to Statistical Learning. Second. NY, USA. p. 197–219. Available from: https://hastie.su.domains/ISLR2/ISLRv2_website.pdf

Genangeli, M., R. M. A. Heeren, and T. Porta Siegel. 2019. Tissue classification by rapid evaporative ionization mass spectrometry (REIMS): comparison between a diathermic knife and CO2 laser sampling on classification performance. Anal Bioanal Chem. 411:7943–7955. doi:10.1007/s00216-019-02148-8.

Goodson, K. J., W. W. Morgan, J. O. Reagan, B. L. Gwartney, S. M. Courington, J. W. Wise, and J. W. Savell. 2002. Beef customer satisfaction: factors affecting consumer evaluations of clod steaks. Anim. Sci. J. 80:401–408. doi:10.2527/2002.802401x.

Grayson, A. L., S. D. Shackelford, D. A. King, R. O. McKeith, R. K. Miller, and T. L. Wheeler. 2016. The effects of degree of dark cutting on tenderness and sensory attributes of beef. J. Anim. Sci. 94:2583–2591. doi:10.2527/jas.2016-0388.

Gredell, D. 2018. Assessment of rapid evaporative ionization mass spectrometry (REIMS) to characterize beef quality and the impact of oven temperature and relative humdity on beef. Colorado State University, CO, USA.

Gredell, D. A., A. R. Schroeder, K. E. Belk, C. D. Broeckling, A. L. Heuberger, S.-Y. Kim, D. A. King, S. D. Shackelford, J. L. Sharp, T. L. Wheeler, D. R. Woerner, and J. E. Prenni. 2019. Comparison of machine learning algorithms for predictive modeling of beef attributes using rapid evaporative ionization mass spectrometry (REIMS) data. Sci Rep. 9:5721. doi:10.1038/s41598-019-40927-6.

Guitton, Y., G. Dervilly-Pinel, R. Jandova, S. Stead, Z. Takats, and B. Le Bizec. 2018. Rapid evaporative ionisation mass spectrometry and chemometrics for high-throughput screening of growth promoters in meat producing animals. Food Addit Contam Part A. 35:900–910. doi:10.1080/19440049.2017.1421778.

Gundersen, O. E., K. Coakley, and C. Kirkpatrick. 2022a. Sources of irreproducibility in machine learning: a Review. arXiv. Available from: http://arxiv.org/abs/2204.07610

Gundersen, O. E., S. Shamsaliei, and R. J. Isdahl. 2022b. Do machine learning platforms provide out-of-the-box reproducibility? Future Gener Comput Syst. 126:34–47. doi:10.1016/j.future.2021.06.014.

Haggard, P., and L. de Boer. 2014. Oral somatosensory awareness. Neurosci Biobehav Rev. 47:469–484. doi:10.1016/j.neubiorev.2014.09.015.

Hauskrecht, M., R. Pelikan, M. Valko, and J. Lyons-Weiler. 2007. Feature selection and dimensionality reduction in genomics and proteomics. In: W. Dubitzky, M. Granzow, and D. Berrar, editors. Fundamentals of Data Mining in Genomics and Proteomics. Springer US, Boston, MA. p. 149–172. Available from: https://doi.org/10.1007/978-0-387-47509-7_7

He, Q., M. Yang, X. Chen, X. Yan, Y. Li, M. He, T. Liu, F. Chen, and F. Zhang. 2021. Differentiation between fresh and frozen-thawed meat using rapid evaporative ionization mass spectrometry: the case of beef muscle. J Agric Food Chem. 69:5709–5724. doi:10.1021/acs.jafc.0c07942.

Heil, B. J., M. M. Hoffman, F. Markowetz, S.-I. Lee, C. S. Greene, and S. C. Hicks. 2021. Reproducibility standards for machine learning in the life sciences. Nat Methods. 18:1132–1135. doi:10.1038/s41592-021-01256-7.

Henderson, P., R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger. 2018. Deep reinforcement learning that matters. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. p. 3207–3214. Available from: https://ojs.aaai.org/index.php/AAAI/article/view/11694

Hofmann, T., and P. Schieberle. 2000. Formation of aroma-active Strecker-aldehydes by a direct oxidative degradation of Amadori compounds. J. Agric. Food Chem. 48:4301–4305. doi:10.1021/jf000076e.

Howard, S., D. Goldberg, J. Tatum, D. R. Woerner, G. Smith, and K. E. Belk. 2010. Beef research - high resolution imaging and tenderness. Colorado State University and Tenera Technologies. Available from: https://www.beefresearch.org/resources/product-quality/project-summaries/2009/high-resolution-imaging-and-tenderness

Hunt, M. R. 2013. Consumer assessment and objective measures of beef flavor of four beef muscles from USDA choice and select graded carcasses [Thesis]. Texas Tech University. Available from: https://ttu-ir.tdl.org/handle/2346/50296

Islam, R., P. Henderson, M. Gomrokchi, and D. Precup. 2017. Reproducibility of benchmarked deep reinforcement learning tasks for continuous control. arXiv. Available from: http://arxiv.org/abs/1708.04133

Ismail, I., Y.-H. Hwang, and S.-T. Joo. 2020. Low-temperature and long-time heating regimes on non-volatile compound and taste traits of beef assessed by the electronic tongue system. Food Chem. 320:126656. doi:10.1016/j.foodchem.2020.126656.

Jones, E. A., D. Simon, T. Karancsi, J. Balog, S. D. Pringle, and Z. Takats. 2019. Matrix assisted rapid evaporative ionization mass spectrometry. Anal. Chem. 91:9784–9791. doi:10.1021/acs.analchem.9b01441.

Joseph, P., S. P. Suman, G. Rentfrow, S. Li, and C. M. Beach. 2012. Proteomics of muscle-specific beef color stability. J. Agric. Food Chem. 60:3196–3203. doi:10.1021/jf204188v.

Juárez, M., N. Aldai, Ó. López-Campos, M. E. R. Dugan, B. Uttaro, and J. L. Aalhus. 2012. Beef texture and juiciness. In: Handbook of meat and meat processing. 2nd ed. CRC Press. p. 178–195.

Kemsley, E. K. 1996. Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods. Chemometr Intell Lab Syst. 33:47–61. doi:10.1016/0169-7439(95)00090-9.

Kerth, C. R., and R. K. Miller. 2015. Beef flavor: a review from chemistry to consumer. J. Sci. Food Agric. 95:2783–2798. doi:10.1002/jsfa.7204.

Kettaneh, N., A. Berglund, and S. Wold. 2005. PCA and PLS with very large data sets. CSDA. 48:69–85. doi:10.1016/j.csda.2003.11.027.

Kim, G.-D., S. Yun Lee, E.-Y. Jung, S. Song, and S. Jin Hur. 2021. Quantitative changes in peptides derived from proteins in beef tenderloin (psoas major muscle) and striploin (longissimus lumborum muscle) during cold storage. Food Chem. 338:128029. doi:10.1016/j.foodchem.2020.128029.

King, D. A., S. D. Shackelford, C. D. Broeckling, J. E. Prenni, K. E. Belk, and T. L. Wheeler. 2019. Metabolomic investigation of tenderness and aging response in beef longissimus steaks. Meat Muscle Biol. 3. doi:10.22175/mmb2018.09.0027. Available from: https://www.iastatedigitalpress.com/mmb/article/id/9107/

Kodogiannis, V. 2017. A rapid detection of meat spoilage using an electronic nose and fuzzy-wavelet systems. In: Y. Bi, S. Kapoor, and R. Bhatia, editors. Lecture Notes in Networks and Systems. Vol. 15. Springer, London. Available from: https://doi.org/10.1007/978-3-319-56994-9_36

Kohavi, R. 1995. A study of cross validation and bootstrap for accuracy estimation and model selection. Available from: http://ai.stanford.edu/~ronnyk/accEst.pdf

Konda Naganathan, G., K. Cluff, A. Samal, C. R. Calkins, D. D. Jones, G. E. Meyer, and J. Subbiah. 2016. Three dimensional chemometric analyses of hyperspectral images for beef tenderness forecasting. J. Food Eng. 169:309–320. doi:10.1016/j.jfoodeng.2015.09.001.

Konda Naganathan, G., L. M. Grimes, J. Subbiah, C. R. Calkins, A. Samal, and G. E. Meyer. 2008. Partial least squares analysis of near-infrared hyperspectral images for beef tenderness prediction. Sens. & Instrumen. Food Qual. 2:178–188. doi:10.1007/s11694-008-9051-3.

Koohmaraie, M. 1996. Biochemical factors regulating the toughening and tenderization processes of meat. Meat Sci. 43:193–201. doi:10.1016/0309-1740(96)00065-4.

Kuhn, M. 2008. Building predictive models in *R* using the caret package. J. Stat. Soft. 28:1–26. doi:10.18637/jss.v028.i05.

Kuznetsova, A., P. Brockhoff, and R. Christensen. 2017. lmerTest package: tests in linear mixed effects models. J. Stat. Softw. 83:1–26. doi:doi: 10.18637/jss.v082.i13.

Lê, S., J. Josse, and F. Husson. 2008. FactoMineR: An *R* Package for Multivariate Analysis. J. Stat. Soft. 25. doi:10.18637/jss.v025.i01. Available from: http://www.jstatsoft.org/v25/i01/

Legako, J. F., J. C. Brooks, T. G. O'Quinn, T. D. J. Hagan, R. Polkinghorne, L. J. Farmer, and M. F. Miller. 2015. Consumer palatability scores and volatile beef flavor compounds of five USDA quality grades and four muscles. Meat Sci. 100:291–300. doi:10.1016/j.meatsci.2014.10.026.

Legako, J. F., T. T. N. Dinh, M. F. Miller, K. Adhikari, and J. C. Brooks. 2016. Consumer palatability scores, sensory descriptive attributes, and volatile compounds of grilled beef steaks from three USDA Quality Grades. Meat Sci. 112:77–85. doi:10.1016/j.meatsci.2015.10.018.

Lepper-Blilie, A. N., E. P. Berg, D. S. Buchanan, and P. T. Berg. 2016. Effects of post-mortem aging time and type of aging on palatability of low marbled beef loins. Meat Sci. 112:63–68. doi:10.1016/j.meatsci.2015.10.017.

Liu, J., M.-P. Ellies-Oury, S. Chriki, I. Legrand, G. Pogorzelski, J. Wierzbicki, L. Farmer, D. Troy, R. Polkinghorne, and J.-F. Hocquette. 2020. Contributions of tenderness, juiciness and flavor liking to overall liking of beef in Europe. Meat Sci. 168:108190. doi:10.1016/j.meatsci.2020.108190.

Lorena, A. C., A. C. P. L. F. de Carvalho, and J. M. P. Gama. 2009. A review on the combination of binary classifiers in multiclass problems. Artif Intell Rev. 30:19. doi:10.1007/s10462-009-9114-9.

Lorenzen, C. L., C. R. Calkins, M. D. Green, R. K. Miller, J. B. Morgan, and B. E. Wasser. 2010. Efficacy of performing Warner-Bratzler and slice shear force on the same beef steak following rapid cooking. Meat Sci. 85:792–794. doi:10.1016/j.meatsci.2010.03.030.

van der Maaten, L., E. Postma, and J. Van den Herik. 2009. Dimensionality reduction: a comparative review.

Machiels, D., L. Istasse, and S. M. van Ruth. 2004. Gas chromatography-olfactometry analysis of beef meat originating from differently fed Belgian Blue, Limousin and Aberdeen Angus bulls. Food Chem. 86:377–383. doi:10.1016/j.foodchem.2003.09.011.

Macleod, G. 1994. The flavour of beef. In: F. Shahidi, editor. Flavor of meat and meat products. Springer US, BO, USA. p. 4–37. Available from: https://doi.org/10.1007/978-1-4615-2177-8_2

Malik, A. K., C. Blasco, and Y. Picó. 2010. Liquid chromatography–mass spectrometry in food safety. J. Chromatogr. A. 1217:4018–4040. doi:10.1016/j.chroma.2010.03.015.

Martinez, H. A., A. N. Arnold, J. C. Brooks, C. C. Carr, K. B. Gehring, D. B. Griffin, D. S. Hale, G. G. Mafi, D. D. Johnson, C. L. Lorenzen, R. J. Maddock, R. K. Miller, D. L. VanOverbeke, B. E. Wasser, and J. W. Savell. 2017. National Beef Tenderness Survey—2015: Palatability and shear force assessments of retail and foodservice beef. Meat Muscle Biol. 1. doi:10.22175/mmb2017.05.0028. Available from: https://iastatedigitalpress.com/mmb/article/id/9443/

McKillip, K. V., A. K. Wilfong, J. M. Gonzalez, T. A. Houser, J. A. Unruh, E. A. E. Boyle, and T. G. O'Quinn. 2017. Sensory evaluation of enhanced beef strip loin steaks cooked to 3 degrees of doneness. Meat Muscle Biol. 1. doi:10.22175/mmb2017.06.0033. Available from: https://www.iastatedigitalpress.com/mmb/article/id/9452/

Miller, M. F., M. A. Carr, C. B. Ramsey, K. L. Crockett, and L. C. Hoover. 2001. Consumer thresholds for establishing the value of beef tenderness. J. Anim. Sci. 79:3062–3068. doi:10.2527/2001.79123062x.

Miller, M. f., L. c. Hoover, K. d. Cook, A. l. Guerra, K. l. Huffman, Ks. Tinney, C. b. Ramsey, H. c. Brittin, and L. m. Huffman. 1995. Consumer acceptability of beef steak tenderness in the home and restaurant. J. Food Sci. 60:963–965. doi:10.1111/j.1365-2621.1995.tb06271.x.

Mitsumoto, M., S. Maeda, T. Mitsuhashi, and S. Ozawa. 1991. Near-infrared spectroscopy determination of physical and chemical characteristics in beef cuts. J. Food Sci. 56:1493–1496. doi:10.1111/j.1365-2621.1991.tb08623.x.

Mohammed, H. H. H., L. He, A. Nawaz, G. Jin, X. Huang, M. Ma, W. S. Abdegadir, E. A. Elgasim, and I. Khalifa. 2021. Effect of frozen and refrozen storage of beef and chicken meats on inoculated microorganisms and meat quality. Meat Science. 175:108453. doi:10.1016/j.meatsci.2021.108453.

Nair, M. N., A. C. V. C. S. Canto, G. Rentfrow, and S. P. Suman. 2019. Muscle-specific effect of aging on beef tenderness. LWT. 100:250–252. doi:10.1016/j.lwt.2018.10.038.

Nishimura, T., A. Hattori, and K. Takahashi. 1999. Structural changes in intramuscular connective tissue during the fattening of Japanese black cattle: effect of marbling on beef tenderization. J. Anim. Sci. 77:93–104. doi:10.2527/1999.77193x.

O'Quinn, T. G., J. F. Legako, J. C. Brooks, and M. F. Miller. 2018. Evaluation of the contribution of tenderness, juiciness, and flavor to the overall consumer beef eating experience. Transl. Anim. Sci. 2:26–36. doi:10.1093/tas/txx008.

O'Quinn, T. G., D. R. Woerner, T. E. Engle, P. L. Chapman, J. F. Legako, J. C. Brooks, K. E. Belk, and J. D. Tatum. 2016. Identifying consumer preferences for specific beef flavor characteristics in relation to cattle production and postmortem processing parameters. Meat Sci. 112:90–102. doi:10.1016/j.meatsci.2015.11.001.

Pavan, E., and S. K. Duckett. 2013. Fatty acid composition and interrelationships among eight retail cuts of grass-feed beef. Meat Sci. 93:371–377. doi:10.1016/j.meatsci.2012.09.021.

Phelps, D. L., J. Balog, L. F. Gildea, Z. Bodai, A. Savage, M. A. El-Bahrawy, A. V. Speller, F. Rosini, H. Kudo, J. S. McKenzie, R. Brown, Z. Takáts, and S. Ghaem-Maghami. 2018. The surgical intelligent knife distinguishes normal, borderline and malignant gynaecological tissues using rapid evaporative ionisation mass spectrometry (REIMS). Br J Cancer. 118:1349–1358. doi:10.1038/s41416-018-0048-3.

Picard, B., M. Gagaoua, D. Micol, I. Cassar-Malek, J.-F. Hocquette, and C. E. M. Terlouw. 2014. Inverse relationships between biomarkers and beef tenderness according to contractile and metabolic properties of the muscle. J. Agric. Food Chem. 62:9808–9818. doi:10.1021/jf501528s.

Pineau, J., P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and H. Larochelle. 2020. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 Reproducibility Program). J Mach Learn Res. 20:1–20. doi:10.48550/arXiv.2003.12206.

Platter, W. J., J. D. Tatum, K. E. Belk, P. L. Chapman, J. A. Scanga, and G. C. Smith. 2003. Relationships of consumer sensory ratings, marbling score, and shear force value to consumer acceptance of beef strip loin steaks. J. Anim. Sci. 81:2741–2750. doi:10.2527/2003.81112741x.

Purchas, R. W. 2014. Tenderness measurement. In: M. Dikeman and C. Devine, editors. Encyclopedia of meat sciences. 2nd ed. Academic Press, Oxford. p. 452–459. Available from: https://www.sciencedirect.com/science/article/pii/B9780123847317001902

Raes, K., A. Balcaen, P. Dirinck, A. De Winne, E. Claeys, D. Demeyer, and S. De Smet. 2003. Meat quality, fatty acid composition and flavour analysis in Belgian retail beef. Meat Sci. 65:1237–1246. doi:10.1016/S0309-1740(03)00031-7.

Rifkin, R., and A. Klautau. 2004. In defense of one-vs-all classification. J. Mach. Learn. Res. 5:101–141.

Ringnér, M. 2008. What is principal component analysis? Nat Biotechnol. 26:303–304. doi:10.1038/nbt0308-303.

Robson, K., N. Birse, O. Chevallier, and C. Elliott. 2022. Metabolomic profiling to detect different forms of beef fraud using rapid evaporative ionisation mass spectrometry (REIMS). npj Sci Food. 6:9. doi:10.1038/s41538-022-00125-7.

Rosenvold, K., E. Micklander, P. W. Hansen, R. Burling-Claridge, M. Challies, C. Devine, and M. North. 2009. Temporal, biochemical and structural factors that influence beef quality measurement using near infrared spectroscopy. Meat Sci. 82:379–388. doi:10.1016/j.meatsci.2009.02.010.

Ross, A., C. Brunius, O. Chevallier, G. Dervilly, C. Elliott, Y. Guitton, J. E. Prenni, O. Savolainen, L. Hemeryck, N. H. Vidkjær, N. Scollan, S. L. Stead, R. Zhang, and L. Vanhaecke. 2021. Making complex measurements of meat composition fast: application of rapid evaporative ionisation mass spectrometry to measuring meat quality and fraud. Meat Sci. 181:108333. doi:10.1016/j.meatsci.2020.108333.

Savell, J. w., R. e. Branson, H. r. Cross, D. m. Stiffler, J. w. Wise, D. b. Griffin, and G. c. Smith. 1987. National Consumer Retail Beef Study: palatability evaluations of beef loin steaks that differed in marbling. J. Food Sc. 52:517–519. doi:10.1111/j.1365-2621.1987.tb06664.x.

Savell, J. W., C. L. Lorenzen, T. R. Neely, R. K. Miller, J. D. Tatum, J. W. Wise, J. F. Taylor, M. J. Buyck, and J. O. Reagan. 1999. Beef customer satisfaction: cooking method and degree of doneness effects on the top sirloin steak. J. Anim. Sci. 77:645–652. doi:10.2527/1999.773645x.

Schäfer, K.-C., J. Dénes, K. Albrecht, T. Szaniszló, J. Balog, R. Skoumal, M. Katona, M. Tóth, L. Balogh, and Z. Takáts. 2009. In vivo, in situ tissue analysis using rapid evaporative ionization mass spectrometry. Angew. Chem. Int. Ed. 48:8240–8242. doi:10.1002/anie.200902546.

Semler, M. L., D. R. Woerner, K. E. Belk, K. J. Enns, and J. D. Tatum. 2016. Effects of United States Department of Agriculture carcass maturity on sensory attributes of steaks produced by cattle representing two dental age classes. J Anim Sci. 94:2207–2217. doi:10.2527/jas.2016-0382.

Shackelford, S. D., T. L. Wheeler, and M. Koohmaraie. 1999. Evaluation of slice shear force as an objective method of assessing beef longissimus tenderness. Journal of Animal Science. 77:2693. doi:10.2527/1999.77102693x.

Shahidi, F. 1994. Flavor of meat and meat products—an overview. In: Fereidoon Shahidi, editor. Flavor of meat and meat products. Springer US, BO, USA. p. 1–3. Available from: https://doi.org/10.1007/978-1-4615-2177-8_1

Smith, G. C., Z. L. Carpenter, H. R. Cross, C. E. Murphey, H. C. Abraham, J. W. Savell, G. W. Davis, B. W. Berry, and F. C. J. Parrish. 1985. Relationship of USDA marbling groups to palatability of cooked beef. J. Food Qual. 7:289–308.

Smulders, F. J. M., B. B. Marsh, D. R. Swartz, R. L. Russell, and M. E. Hoenecke. 1990. Beef tenderness and sarcomere length. Meat Sci. 28:349–363. doi:10.1016/0309-1740(90)90048-B.

Spence, C., and B. Piqueras-Fiszman. 2016. Oral-somatosensory contributions to flavor perception and the appreciation of food and drink. In: B. Piqueras-Fiszman and C. Spence, editors. Multisensory flavor perception. Woodhead Publishing. p. 59–79. Available from: https://www.sciencedirect.com/science/article/pii/B9780081003503000043

St John, E. R., J. Balog, J. S. McKenzie, M. Rossi, A. Covington, L. Muirhead, Z. Bodai, F. Rosini, A. V. M. Speller, S. Shousha, R. Ramakrishnan, A. Darzi, Z. Takats, and D. R. Leff. 2017. Rapid evaporative ionisation mass spectrometry of electrosurgical vapours for the identification of breast pathology: towards an intelligent knife for breast cancer surgery. Breast Cancer Res. 19:59. doi:10.1186/s13058-017-0845-2.

Strittmatter, N., M. Rebec, E. A. Jones, O. Golf, A. Abdolrasouli, J. Balog, V. Behrends, K. A. Veselkov, and Z. Takats. 2014. Characterization and identification of clinically relevant microorganisms using rapid evaporative ionization mass spectrometry. Anal. Chem. 86:6555–6562. doi:10.1021/ac501075f.

Tabb, D. L., L. Vega-Montoto, P. A. Rudnick, A. M. Variyath, A.-J. L. Ham, D. M. Bunk, L. E. Kilpatrick, D. D. Billheimer, R. K. Blackman, H. L. Cardasis, S. A. Carr, K. R. Clauser, J. D. Jaffe, K. A. Kowalski, T. A. Neubert, F. E. Regnier, B. Schilling, T. J. Tegeler, M. Wang, P. Wang, J. R. Whiteaker, L. J. Zimmerman, S. J. Fisher, B. W. Gibson, C. R. Kinsinger, M. Mesri, H. Rodriguez, S. E. Stein, P. Tempst, A. G. Paulovich, D. C. Liebler, and C. Spiegelman. 2010. Repeatability and reproducibility in proteomic identifications by liquid chromatography−tandem mass spectrometry. J. Proteome Res. 9:761–776. doi:10.1021/pr9006365.

USDA. 2017. United States Standards for Grades of Carcass Beef.

USDA. 2019. What's Your Beef – Prime, Choice or Select? Available from: https://www.usda.gov/media/blog/2013/01/28/whats-your-beef-prime-choice-or-select

Van Oeckel, M. J., N. Warnants, and Ch. V. Boucqué. 1999. Comparison of different methods for measuring water holding capacity and juiciness of pork versus on-line screening methods. Meat Sci. 51:313–320. doi:10.1016/S0309-1740(98)00123-5.

Verplanken, K., S. Stead, R. Jandova, C. V. Poucke, J. Claereboudt, J. V. Bussche, S. D. Saeger, Z. Takats, J. Wauters, and L. Vanhaecke. 2017. Rapid evaporative ionization mass spectrometry for high-throughput screening in food analysis: the case of boar taint. Talanta. 169:30–36. doi:10.1016/j.talanta.2017.03.056.

Veselkov, K. A., R. Mirnezami, N. Strittmatter, R. D. Goldin, J. Kinross, A. V. M. Speller, T. Abramov, E. A. Jones, A. Darzi, E. Holmes, J. K. Nicholson, and Z. Takats. 2014. Chemo-informatic strategy for imaging mass spectrometry-based hyperspectral profiling of lipid signatures in colorectal cancer | PNAS. PNAS. 111:1216–1221. doi:https://doi.org/10.1073/pnas.1310524111.

Volmer, D., and L. L. Jessome. 2006. Ion Suppression: A Major Concern in Mass Spectrometry. LCGC North America. 24:498–510.

Vote, D. J., K. E. Belk, J. D. Tatum, J. A. Scanga, and G. C. Smith. 2003. Online prediction of beef tenderness using a computer vision system equipped with a BeefCam module. J. Anim. Sci. 81:457–465. doi:10.2527/2003.812457x.

Ward, M. 2021. Beef quality grades explained. Beef Magazine. Available from: https://www.beefmagazine.com/beef-quality/beef-quality-grades-explained

Wasserman, A. E. 1972. Thermally produced flavor components in the aroma of meat and poultry. J. Agric. Food Chem. 20:737–741. doi:10.1021/jf60182a015.

Webb, N. B., O. J. Kahlenberg, and H. D. Naumann. 1964. Factors influencing beef tenderness. J. Anim. Sci. 23:1027–1031. doi:10.2527/jas1964.2341027x.

Wheeler, T. L., L. V. Cundiff, and R. M. Koch. 1994. Effect of marbling degree on beef palatability in *Bos taurus* and *Bos indicus* cattle. J. Anim. Sci. 72:3145–3151. doi:10.2527/1994.72123145x.

Wheeler, T. l., R. k. Miller, J. w. Savell, and H. r. Cross. 1990. Palatability of chilled and frozen beef steaks. J. Food Sci. 55:301–304. doi:10.1111/j.1365-2621.1990.tb06748.x.

Wheeler, T. L., L. S. Papadopoulos, R. Miller, K. E. Belk, M. Dikeman, C. R. Calkins, A. King, M. F. Miller, S. D. Shackelford, B. E. Wasser, and L. D. Yates. 2015. Research guidelines for cookery, sensory evaluation, and instrumental tenderness measurements of meat.

Wheeler, T. L., S. D. Shackelford, L. P. Johnson, M. F. Miller, R. K. Miller, and M. Koohmaraie. 1997. A comparison of Warner-Bratzler shear force assessment within and among institutions. J. Anim. Sci. 75:2423–2432. doi:10.2527/1997.7592423x.

Wheeler, T. L., S. D. Shackelford, and M. Koohmaraie. 1996. Sampling, cooking, and coring effects on Warner-Bratzler shear force values in beef. J Anim Sci. 74:1553–1562. doi:10.2527/1996.7471553x.

Wheeler, T. L., S. D. Shackelford, and M. Koohmaraie. 2005. Shear force procedures for meat tenderness measurement. Available from: www.ars.usda.gov/ARSUserFiles/30400510/protocols/ShearForceProcedures.pdf

Wheeler, T. L., D. Vote, J. M. Leheska, S. D. Shackelford, K. E. Belk, D. M. Wulf, B. L. Gwartney, and M. Koohmaraie. 2002. The efficacy of three objective systems for identifying beef cuts that can be guaranteed tender. J Anim Sci. 80:3315–3327. doi:10.2527/2002.80123315x.

Wyle, A. M., D. J. Vote, D. L. Roeber, R. C. Cannell, K. E. Belk, J. A. Scanga, M. Goldberg, J. D. Tatum, and G. C. Smith. 2003. Effectiveness of the SmartMV prototype BeefCam system to sort beef carcasses into expected palatability groups. J. Anim. Sci. 81:441–448. doi:10.2527/2003.812441x.

Yancey, J. W. S., J. K. Apple, J.-F. Meullenet, and J. T. Sawyer. 2010. Consumer responses for tenderness and overall impression can be predicted by visible and near-infrared spectroscopy, Meullenet–Owens razor shear, and Warner–Bratzler shear force. Meat Sci. 85:487–492. doi:10.1016/j.meatsci.2010.02.020.

Zhang, R., A. B. Ross, M. J. Y. Yoo, and M. M. Farouk. 2021. Use of rapid evaporative ionisation mass spectrometry fingerprinting to determine the metabolic changes to dry-aged lean beef due to different ageing regimes. Meat Sci. 181:108438. doi:10.1016/j.meatsci.2021.108438.

Zhang, X., Y. Zhang, Q. Meng, N. Li, and L. Ren. 2015. Evaluation of beef by electronic tongue system TS-5000Z: flavor assessment, recognition and chemical compositions according to its correlation with flavor. PLOS ONE. 10:e0137807. doi:10.1371/journal.pone.0137807.

Table 11. Classification accuracies (based on 10-fold cross-validation, percent) of slice shear force models of striploins aged 3 (d3), 14 (d14), and 28 (d28) days.

Machine learning[1]	d3			d14			d28		
	FS[2]	PCA[3]	PCA-FS[4]	FS	PCA	PCA-FS	FS	PCA	PCA-FS
SVM	80.6	73.8	74.8	80.2	52.0	80.9	85.2	81.4	86.7
SVM Linear	83.3	74.7	79.4	78.5	59.0	76.1	76.6	82.7	82.8
LDA	56.8	73.6	75.4	45.1	60.1	76.2	58.3	79.8	83.1
Logit Boost	75.8	72.4	70.0	73.6	49.6	79.5	80.0	78.5	90.1
SVM Poly	84.4	75.8	79.2	84.1	61.7	79.8	83.7	82.6	90.6
RF	74.8	68.3	74.7	77.2	52.4	83.7	86.7	78.9	87.5
XGBoost	82.0	74.5	76.9	83.5	62.4	84.8	87.8	82.9	91.5
PLSDA	83.8	73.8	68.8	81.3	60.0	77.5	78.8	82.5	85.0
GBM	79.1	65.3	73.5	75.1	48.2	81.5	89.0	80.9	90.1
PDA	76.5	72.5	70.1	77.9	58.5	74.1	61.7	80.2	86.2
GLM	50.4	72.2	72.2	59.1	61.2	61.2	52.1	80.7	80.7
GLMNET	85.8	74.7	74.7	73.7	58.9	58.9	82.7	82.8	82.8
KNN	79.4	73.5	73.5	66.5	54.7	54.7	86.4	82.6	82.6
RPART	73.6	60.1	60.1	50.9	63.9	63.9	68.6	76.8	76.8
TREEBAG	71.4	68.6	68.6	70.9	47.1	47.1	81.6	72.2	72.2

[1] SVM, support vector machine; LDA, linear discriminant analysis; RF, random forest; XGBoost, extreme gradient boosting; PLSDA, partial least square discriminant analysis; GBM, gradient boosting machine; PDA, penalized discriminant analysis; GLM, generalized linear model; KNN, k-nearest neighbors; RPART, recursive partitioning

[2] FS, feature selection

[3] PCA, principal component analysis

[4] PCA-FS, principal component analysis followed by feature selection

Table 12. Classification accuracies (based on 10-fold cross-validation, percent) of Warner-Bratzler shear force models of striploins aged 3 (d3), 14 (d14), and 28 (d28) days.

Machine learning[1]	d3			d14			d28		
	FS[2]	PCA[3]	PCA-FS[4]	FS	PCA	PCA-FS	FS	PCA	PCA-FS
SVM	80.7	78.1	79.7	74.1	47.3	56.0	86.1	74.8	88.6
SVM Linear	76.5	79.7	77.0	64.2	63.6	64.1	86.3	76.1	88.6
LDA	68.9	79.3	74.2	68.1	58.2	63.9	60.9	73.6	87.5
Logit Boost	80.6	76.9	77.1	77.2	57.4	58.9	77.1	76.1	81.1
SVM Poly	82.3	82.1	79.7	76.3	62.7	62.5	87.3	76.1	94.8
RF	82.5	77.9	80.9	82.4	51.9	57.0	83.8	72.0	79.8
XGBoost	86.5	80.3	80.6	80.2	57.4	66.4	87.3	77.5	84.6
PLSDA	82.5	82.5	80.0	80.0	57.5	62.5	85.7	75.7	85.7
GBM	84.1	73.0	77.1	77.7	51.3	60.8	83.6	71.7	85.9
PDA	71.0	77.2	77.0	64.9	61.9	60.1	79.8	73.6	85.9
GLM	67.6	79.5	79.5	63.8	62.3	62.3	53.2	74.8	74.8
GLMNET	80.8	82.2	82.2	74.8	62.3	62.3	82.1	76.1	76.1
KNN	80.7	79.8	79.8	66.8	56.1	56.1	79.8	72.1	72.1
RPART	67.4	69.0	69.0	63.0	66.0	66.0	72.1	68.6	68.6
TREEBAG	76.2	74.4	74.4	76.4	44.4	44.4	77.0	73.4	73.4

[1] SVM, support vector machine; LDA, linear discriminant analysis; RF, random forest; XGBoost, extreme gradient boosting; PLSDA, partial least square discriminant analysis; GBM, gradient boosting machine; PDA, penalized discriminant analysis; GLM, generalized linear model; KNN, k-nearest neighbors; RPART, recursive partitioning

[2] FS, feature selection

[3] PCA, principal component analysis

[4] PCA-FS, principal component analysis followed by feature selection

Table 13. Classification accuracies (based on 10-fold cross-validation, percent) of tenderness panel models of striploins aged 3 (d3), 14 (d14), and 28 (d28) days.

Machine learning[1]	d3			d14			d28		
	FS[2]	PCA[3]	PCA-FS[4]	FS	PCA	PCA-FS	FS	PCA	PCA-FS
SVM	93.1	85.4	86.9	83.6	62.9	64.6	72.5	70.5	72.0
SVM Linear	85.4	85.7	88.0	76.3	60.9	64.7	77.3	71.4	73.7
LDA	80.8	84.2	86.8	69.6	65.6	53.1	55.2	70.1	73.6
Logit Boost	93.1	87.1	89.1	79.1	69.6	63.3	72.5	65.5	70.9
SVM Poly	91.4	86.9	89.3	86.1	67.5	64.7	78.1	71.4	76.7
RF	92.9	87.6	86.8	81.0	58.6	64.7	76.2	69.9	76.0
XGBoost	93.9	89.3	91.7	87.4	72.2	71.3	79.7	71.3	77.5
PLSDA	90.0	87.5	86.3	80.0	62.9	60.0	75.0	70.0	67.5
GBM	91.5	85.8	89.2	82.6	68.5	63.7	76.0	65.8	76.2
PDA	83.1	87.1	87.8	66.5	68.5	65.8	77.3	70.1	71.8
GLM	79.4	84.4	84.4	66.7	65.0	65.0	59.3	70.0	70.0
GLMNET	91.7	87.9	87.9	82.1	70.8	70.8	74.1	71.3	71.3
KNN	91.7	86.8	86.8	72.7	67.1	67.1	73.9	76.3	76.3
RPART	83.3	80.8	80.8	62.0	68.0	68.0	66.4	69.4	69.4
TREEBAG	88.2	83.1	83.1	80.8	64.2	64.2	71.1	72.3	72.3

[1] SVM, support vector machine; LDA, linear discriminant analysis; RF, random forest; XGBoost, extreme gradient boosting; PLSDA, partial least square discriminant analysis; GBM, gradient boosting machine; PDA, penalized discriminant analysis; GLM, generalized linear model; KNN, k-nearest neighbors; RPART, recursive partitioning

[2] FS, feature selection

[3] PCA, principal component analysis

[4] PCA-FS, principal component analysis followed by feature selection

Table 14. Classification accuracies (based on 10-fold cross-validation, percent) of juiciness models of striploins aged 3 (d3), 14 (d14), and 28 (d28) days.

Machine learning[1]	d3			d14			d28		
	FS[2]	PCA[3]	PCA-FS[4]	FS	PCA	PCA-FS	FS	PCA	PCA-FS
SVM	79.1	64.8	67.9	76.1	56.1	68.8	77.5	59.8	72.5
SVM Linear	70.9	67.4	66.4	70.7	63.0	64.8	68.8	62.5	68.8
LDA	48.0	64.7	69.0	69.6	60.7	64.6	52.5	57.5	75.0
Logit Boost	83.3	57.9	74.5	77.1	62.1	67.1	73.8	71.3	63.8
SVM Poly	80.1	69.9	71.0	78.2	64.5	71.3	81.3	66.3	73.8
RF	82.5	68.2	69.3	77.3	54.3	64.8	80.0	58.8	73.8
XGBoost	85.0	65.2	76.1	83.6	65.9	69.3	86.3	70.0	77.5
PLSDA	70.0	63.8	66.3	77.1	60.0	65.7	73.8	62.5	72.5
GBM	84.5	62.7	73.7	82.5	57.5	75.0	80.0	62.3	77.5
PDA	55.7	65.0	63.2	64.5	60.9	66.1	60.0	60.0	73.8
GLM	56.7	66.8	66.8	67.3	59.8	59.8	55.0	62.5	62.5
GLMNET	78.2	67.0	67.0	75.7	67.1	67.1	80.0	62.5	62.5
KNN	78.7	66.3	66.3	73.4	72.1	72.1	75.0	60.0	60.0
RPART	68.0	61.0	61.0	60.5	58.2	58.2	68.0	68.8	68.8
TREEBAG	84.6	61.6	61.6	73.2	54.3	54.3	80.0	57.5	57.5

[1] SVM, support vector machine; LDA, linear discriminant analysis; RF, random forest; XGBoost, extreme gradient boosting; PLSDA, partial least square discriminant analysis; GBM, gradient boosting machine; PDA, penalized discriminant analysis; GLM, generalized linear model; KNN, k-nearest neighbors; RPART, recursive partitioning

[2] FS, feature selection

[3] PCA, principal component analysis

[4] PCA-FS, principal component analysis followed by feature selection

Table 15. Classification accuracies (based on 10-fold cross-validation, percent) of flavor models of striploins aged 3 (d3), 14 (d14), and 28 (d28) days.

Machine learning[1]	d3			d14			d28		
	FS[2]	PCA[3]	PCA-FS[4]	FS	PCA	PCA-FS	FS	PCA	PCA-FS
SVM	78.6	62.8	76.3	85.7	62.5	63.4	76.3	61.4	56.3
SVM Linear	75.0	68.2	83.1	77.1	60.7	62.7	65.5	61.9	55.2
LDA	49.0	72.1	84.3	44.0	61.1	52.5	66.3	63.9	53.3
Logit Boost	81.3	69.6	79.4	70.1	67.7	60.9	69.9	54.1	64.0
SVM Poly	80.6	74.7	83.5	86.8	68.6	71.3	80.2	71.4	56.3
RF	84.6	64.7	78.6	74.5	63.3	64.4	73.8	58.0	70.1
XGBoost	85.7	72.2	86.7	82.2	73.1	74.6	81.0	69.7	69.8
PLSDA	81.3	72.5	83.8	88.6	60.0	64.3	72.5	60.0	55.0
GBM	86.7	62.3	84.3	76.2	61.0	67.2	77.1	52.8	66.8
PDA	79.4	73.2	82.9	82.2	60.9	59.9	67.4	60.2	61.6
GLM	49.3	73.5	73.5	46.1	61.2	61.2	66.2	65.6	65.6
GLMNET	81.0	72.2	72.2	81.3	68.5	68.5	76.0	62.4	62.4
KNN	76.8	70.8	70.8	73.0	63.5	63.5	78.8	65.7	65.7
RPART	58.2	70.0	70.0	53.0	67.7	67.7	67.7	58.0	58.0
TREEBAG	80.4	64.3	64.3	63.0	67.2	67.2	75.6	56.4	56.4

[1] SVM, support vector machine; LDA, linear discriminant analysis; RF, random forest; XGBoost, extreme gradient boosting; PLSDA, partial least square discriminant analysis; GBM, gradient boosting machine; PDA, penalized discriminant analysis; GLM, generalized linear model; KNN, k-nearest neighbors; RPART, recursive partitioning

[2] FS, feature selection

[3] PCA, principal component analysis

[4] PCA-FS, principal component analysis followed by feature selection

Table 16. Classification accuracies (based on 10-fold cross-validation, percent) of slice shear force models of tenderloins aged 3 (d3), 14 (d14), and 28 (d28) days.

Machine learning[1]	d3		d14			d28		
	FS[2]	PCA-FS[4]	FS	PCA	PCA-FS	FS	PCA	PCA-FS
SVM	64.3	75.1	66.3	57.4	69.6	84.3	61.1	61.1
SVM Linear	64.0	78.3	52.2	53.5	63.6	72.2	61.1	59.9
LDA	58.5	74.9	54.3	54.3	70.0	79.3	59.7	53.3
Logit Boost	74.0	56.9	67.9	65.0	65.2	84.0	52.4	64.6
SVM Poly	68.2	82.1	66.7	57.4	70.2	84.3	61.1	61.1
RF	70.6	64.4	71.8	59.3	66.4	85.8	44.3	63.6
XGBoost	74.9	69.6	72.5	66.1	73.4	89.0	57.1	68.5
PLSDA	67.5	77.5	61.3	55.0	65.0	81.3	56.3	60.0
GBM	73.4	63.7	68.6	64.9	68.4	85.7	53.3	61.1
PDA	59.6	76.3	44.8	58.6	62.8	76.8	59.6	53.8
GLM	63.3	68.6	54.1	53.6	63.5	70.8	50.8	61.9
GLMNET	62.5	79.9	65.9	55.6	66.2	86.8	59.6	69.9
KNN	69.3	64.3	68.5	56.3	55.9	77.1	51.3	59.9
RPART	57.5	45.6	52.4	51.2	69.6	81.9	52.8	62.2
TREEBAG	61.5	64.3	69.6	52.9	62.3	80.4	42.6	62.2

[1] SVM, support vector machine; LDA, linear discriminant analysis; RF, random forest; XGBoost, extreme gradient boosting; PLSDA, partial least square discriminant analysis; GBM, gradient boosting machine; PDA, penalized discriminant analysis; GLM, generalized linear model; KNN, k-nearest neighbors; RPART, recursive partitioning

[2] FS, feature selection

[3] PCA, principal component analysis

[4] PCA-FS, principal component analysis followed by feature selection

Table 17. Classification accuracies (based on 10-fold cross-validation, percent) of Warner-Bratzler shear force models of tenderloins aged 3 (d3), 14 (d14), and 28 (d28) days.

Machine learning[1]	d3			d14			d28		
	FS[2]	PCA[3]	PCA-FS[4]	FS	PCA	PCA-FS	FS	PCA	PCA-FS
SVM	71.4	61.1	70.3	69.6	57.7	68.5	76.8	64.3	77.8
SVM Linear	59.6	59.6	67.4	66.6	60.0	64.2	80.7	63.3	71.0
LDA	49.6	58.2	68.8	65.4	55.3	54.7	70.4	63.3	56.5
Logit Boost	61.1	55.0	65.0	65.8	60.5	55.5	64.7	65.4	69.7
SVM Poly	78.6	61.0	77.2	75.7	61.3	68.3	81.5	63.3	72.2
RF	70.6	52.4	67.9	77.7	60.0	61.0	79.3	63.5	68.2
XGBoost	77.6	70.4	71.3	74.2	69.4	66.0	77.9	66.9	74.4
PLSDA	71.3	57.5	73.8	62.5	50.0	77.5	80.0	66.3	71.3
GBM	77.6	62.1	69.0	74.3	63.3	59.4	71.5	63.2	72.9
PDA	54.9	59.6	69.0	64.9	52.8	60.7	76.8	64.7	63.8
GLM	49.6	60.7	66.9	65.1	53.5	52.5	70.7	60.7	64.6
GLMNET	71.8	61.7	73.9	65.6	57.1	60.0	78.1	68.5	71.8
KNN	67.1	56.1	64.4	59.6	56.5	61.4	74.4	64.4	65.6
RPART	55.0	65.3	54.6	62.1	58.1	50.3	63.3	63.6	67.1
TREEBAG	63.6	50.4	60.6	67.2	55.9	53.8	74.7	61.1	72.6

[1] SVM, support vector machine; LDA, linear discriminant analysis; RF, random forest; XGBoost, extreme gradient boosting; PLSDA, partial least square discriminant analysis; GBM, gradient boosting machine; PDA, penalized discriminant analysis; GLM, generalized linear model; KNN, k-nearest neighbors; RPART, recursive partitioning

[2] FS, feature selection

[3] PCA, principal component analysis

[4] PCA-FS, principal component analysis followed by feature selection

Table 18. Classification accuracies (based on 10-fold cross-validation, percent) of tenderness panel models of tenderloins aged 3 (d3), 14 (d14), and 28 (d28) days.

Machine learning[1]	d3			d14			d28		
	FS[2]	PCA[3]	PCA-FS[4]	FS	PCA	PCA-FS	FS	PCA	PCA-FS
SVM	74.9	64.2	80.0	74.5	45.9	61.8	71.3	69.0	84.0
SVM Linear	75.3	64.2	71.3	70.8	53.6	47.2	75.2	65.3	69.4
LDA	54.6	58.2	79.9	46.4	45.4	49.7	72.4	65.5	76.0
Logit Boost	79.9	64.0	78.5	76.3	50.8	59.3	78.1	64.9	76.7
SVM Poly	79.9	64.2	80.4	77.3	58.5	62.4	77.5	72.0	79.0
RF	71.5	52.9	66.5	74.3	44.0	59.1	74.0	63.0	80.3
XGBoost	80.8	64.2	83.8	80.7	51.4	75.0	83.1	69.6	78.1
PLSDA	80.0	62.5	75.0	73.8	55.0	51.3	73.8	63.8	77.5
GBM	78.6	60.6	83.5	78.2	45.4	60.7	77.1	65.0	75.8
PDA	66.4	58.3	78.8	69.2	54.0	56.9	70.0	66.1	74.7
GLM	46.0	60.7	69.9	44.9	47.2	54.8	72.4	62.5	77.4
GLMNET	81.0	64.2	77.1	77.6	48.8	54.3	75.2	70.3	80.1
KNN	71.4	60.4	68.1	69.0	47.2	57.2	76.9	68.4	72.3
RPART	63.2	53.3	62.5	55.4	48.8	58.3	72.1	64.9	60.0
TREEBAG	72.6	47.5	68.9	68.2	42.9	58.2	72.2	60.7	69.0

[1] SVM, support vector machine; LDA, linear discriminant analysis; RF, random forest; XGBoost, extreme gradient boosting; PLSDA, partial least square discriminant analysis; GBM, gradient boosting machine; PDA, penalized discriminant analysis; GLM, generalized linear model; KNN, k-nearest neighbors; RPART, recursive partitioning

[2] FS, feature selection

[3] PCA, principal component analysis

[4] PCA-FS, principal component analysis followed by feature selection

Table 19. Classification accuracies (based on 10-fold cross-validation, percent) of juiciness models of tenderloins aged 3 (d3), 14 (d14), and 28 (d28) days.

Machine learning[1]	d3			d14			d28		
	FS[2]	PCA[3]	PCA-FS[4]	FS	PCA	PCA-FS	FS	PCA	PCA-FS
SVM	71.7	63.1	72.7	62.4	62.1	66.1	70.3	67.5	68.5
SVM Linear	56.4	63.1	67.2	52.5	65.3	75.0	63.3	60.1	65.3
LDA	59.3	57.1	73.7	59.8	60.9	64.8	66.8	61.5	64.7
Logit Boost	62.9	65.4	74.0	61.8	54.3	55.0	72.6	57.8	69.9
SVM Poly	73.0	63.8	71.4	66.5	61.9	79.5	74.6	61.4	79.6
RF	76.9	76.0	76.0	67.6	58.1	64.4	75.0	57.2	69.6
XGBoost	81.1	74.1	79.7	73.9	65.0	67.4	79.4	63.8	74.6
PLSDA	63.8	55.0	72.5	62.5	55.0	72.5	70.0	61.3	68.8
GBM	71.6	65.6	73.6	73.8	64.0	61.8	76.9	59.9	73.3
PDA	66.3	58.4	73.9	51.1	63.1	69.9	62.8	57.9	64.0
GLM	45.4	61.8	69.8	46.8	62.0	73.9	69.6	61.5	61.3
GLMNET	67.8	61.8	74.3	66.5	61.8	74.7	70.1	64.0	70.6
KNN	69.7	62.8	67.5	67.0	56.2	49.7	70.4	51.0	58.3
RPART	63.1	69.0	61.0	57.1	56.6	63.9	61.1	56.5	60.4
TREEBAG	73.9	71.1	65.7	66.8	57.2	61.8	64.7	59.4	63.8

[1] SVM, support vector machine; LDA, linear discriminant analysis; RF, random forest; XGBoost, extreme gradient boosting; PLSDA, partial least square discriminant analysis; GBM, gradient boosting machine; PDA, penalized discriminant analysis; GLM, generalized linear model; KNN, k-nearest neighbors; RPART, recursive partitioning

[2] FS, feature selection

[3] PCA, principal component analysis

[4] PCA-FS, principal component analysis followed by feature selection

Table 20. Classification accuracies (based on 10-fold cross-validation, percent) of flavor models of tenderloins aged 3 (d3), 14 (d14), and 28 (d28) days.

Machine learning[1]	d3			d14			d28		
	FS[2]	PCA[3]	PCA-FS[4]	FS	PCA	PCA-FS	FS	PCA	PCA-FS
SVM	60.4	58.2	80.3	60.6	52.8	69.6	81.0	72.5	82.4
SVM Linear	54.9	59.3	67.2	65.8	44.2	63.5	77.2	72.5	80.8
LDA	53.4	55.1	68.8	64.8	47.4	54.1	80.8	71.3	74.6
Logit Boost	59.6	54.0	76.4	64.0	54.9	53.5	79.6	65.1	76.1
SVM Poly	60.4	59.3	77.4	67.4	58.5	66.9	79.7	74.7	84.3
RF	61.7	51.8	74.9	69.0	41.9	56.1	82.1	64.2	73.6
XGBoost	63.3	58.3	83.3	74.0	48.1	68.2	81.8	68.9	79.4
PLSDA	58.8	57.5	76.3	65.0	50.0	62.5	77.5	71.3	78.8
GBM	57.6	59.6	81.0	72.8	46.7	63.3	79.3	67.8	79.3
PDA	58.9	53.3	67.6	70.1	51.0	65.7	77.2	70.0	77.1
GLM	52.2	55.6	74.4	63.4	48.2	62.5	76.3	71.3	71.4
GLMNET	63.3	56.9	75.4	63.1	55.8	65.6	81.9	72.5	79.4
KNN	56.7	52.1	59.4	66.5	38.7	56.6	78.2	70.1	71.4
RPART	53.9	46.1	69.3	68.8	49.4	46.7	73.6	68.6	63.9
TREEBAG	56.8	54.4	66.5	71.5	37.9	48.4	78.5	58.8	72.8

[1] SVM, support vector machine; LDA, linear discriminant analysis; RF, random forest; XGBoost, extreme gradient boosting; PLSDA, partial least square discriminant analysis; GBM, gradient boosting machine; PDA, penalized discriminant analysis; GLM, generalized linear model; KNN, k-nearest neighbors; RPART, recursive partitioning

[2] FS, feature selection

[3] PCA, principal component analysis

[4] PCA-FS, principal component analysis followed by feature selection