Gastrograph AI

WHITE PAPER

Gastrograph AI Perception and Preference Model Validation vs Central Location Test (CLT)

Jason Cohen | Francis Aubin | Ryan Ahn

CONTENTS

Abstract

Independent validation of new technology is necessary to gain trust and adoption from the industry. In this study, Analytical Flavor Systems (AFS), the company behind Gastrograph AI, partnered with Ajinomoto to conduct a validation of Gastrograph AI's predictability of consumer sensory perception and preference. A joint-panel of experts from AFS and Ajinomoto designed a double blinded study.

A central location test (CLT) consumer panel was hosted in China (N = 242) to measure overall liking and 6 attribute intensity scores of 9 commercial products. Gastrograph AI sampled the same products and predicted the matching sensory results before the CLT was run. For perception predictions, an 85% agreement was found between the CLT measurement and Gastrograph AI's predictions.

No significant difference was found in preference distributions between the two methods, while a 0.77 coefficient of determination was found between the mean preference scores in the two methods. The strong agreement shows that Gastrograph AI as an effective sensory method to accurately translate perceptions across demographics and predict consumer preferences.

1. Introduction

Traditional sensory methods rely on empirical testing of the products under consideration with individual panelists drawn from the population being modeled. As a standard industry practice, central location tests (CLT) are often used to collect consumer perception and preference data on food and beverage products. Overall liking scores and attribute intensities are empirically measured to evaluate differences among samples. While empirical data are readily interpretable, it can suffer from the lack of statistical power, high data requirements, effects of experimental design, and high rates of statistical error from non-representative sampling of panelists.

The recent advancements in artificial intelligence (AI) have led to successful applications in musical compositions, games (super-human performance), text analysis (topic analysis, knowledge gathering, and structured parsing), and text generation (GPT-3). In the food and beverage industry, Analytical Flavor Systems (AFS), the company behind Gastrograph AI, is a market leader in the application of machine learning and artificial intelligence to model human sensory perception of flavor, aroma, texture, and its resulting consumer preference. The core predictive frameworks of Gastrograph AI are the translation of perception and the

prediction of the distribution of preference for the target population.

The predictions of Gastrograph AI yield similar perception and preference results from a CLT panel. Gastrograph AI models topological subspace of a 24-dimensional flavor space that includes common sensory attributes that are tested in CLT. Gastrograph AI also predicts a distribution of perceived quality (PQ) scores based on the perception inputs. The PQ score is comparable to the overall liking scores which are commonly used to measure consumer liking in CLT.

In order to prove Gastrograph AI's robustness as a predictive sensory platform, an independent validation is needed to test its predictability against CLT panels. In this research, Gastrograph AI partnered with Ajinomoto, a trusted supplier in the food and beverage industry to conduct a double blinded validation study. The objectives of this study are to validate Gastrograph AI's capability to 1) translate perception across demographics and 2) to predict distributions of preference for the target population.

2. Materials and Methods

2.1 Experimental approach

The study directly compared perception and preference results generated between a traditional CLT and Gastrograph AI's predictions. A traditional CLT panel is organized by Ajinomoto and a major global market research firm to collect consumer reviews of 9 commercial products. Gastrograph AI analyzed the products from a non-representative group of tasters in Japan and made predictions for 10 demographics in China. The recruitment and screener of the CLT conducted afterward in Shanghai matched the parameters of the demographic groups predicted for by the AI. A series statistical tests were then used to compare the results obtained from CLT and Gastrograph AI's predictions. This study was carefully designed to be double blinded.

In order to prove Gastrograph AI's robustness as a predictive sensory platform, an independent validation is needed to test its predictability against CLT panels. In this research, Gastrograph AI partnered with Ajinomoto, a trusted supplier in the food and beverage industry to conduct a double blinded validation study. The objectives of this study are to validate Gastrograph AI's capability to 1) translate perception

across demographics and 2) to predict distributions of preference for the target population.

2.2 Gastrograph Reviews and Predictions

Ajinomoto recruited 12 panelists in Japan for this study. Each panelist attended a 2 hour training session about the "Gastrograph Review" mobile software before tasting the tested products in a fully randomized order. The reviews collected on the 9 commercial products were analyzed based on Gastrograph AI's demographic translation (section 5.2.1) and preference predictions (section 5.2.2). Briefly, the Japanese reviews of the products were translated into 10 different demographics of Chinese consumers. For each demographic, flavor perception scores and a distribution of perceived quality (PQ) score were predicted. The general Chinese demographic is used to compare to the CLT consumer results and directly evaluate Gastrograph AI's prediction accuracy.

In order to prove Gastrograph AI's robustness as a predictive sensory platform, an independent validation is needed to test its predictability against CLT panels. In this research, Gastrograph AI partnered with Ajinomoto, a trusted supplier in the food and beverage industry to conduct a double blinded validation study. The objectives of this study are to validate Gastrograph AI's capability to 1) translate perception across demographics and 2) to predict distributions of preference for the target population.

2.3 Central Location Test Panel

A consumer CLT panel was organized by a major global market research firm and Ajinomoto in Shanghai, China. In brief, 242 Chinese consumers with balanced demographics evaluated 9 commercial food products with overall liking scores (1 to 7, 7 being most liked) and perception scores of 6 common sensory attributes (roasted, dairy, bitter, rich, sweet, and aftertaste).The detailed panel recruitment process and panel questionnaire can be found in section 5.1.

2.4 Statistical Analysis and Result Comparisons

The perception scores of the 6 attributes were compared between the CLT panel and Gastrograph AI's predictions. A student's t-test was used to conduct pair-comparisons 3 of the intensity scores. The distributions of overall liking and PQ scores were compared using the two-sample Kolmogorov–Smirnov (K-S) test. The KS test quantifies the distance between 2 distributions and tests if the distributions are drawn from the same population. The mean overall liking and PQ scores of the 9 products were tested in a Pearson's correlation test. A significance level of 0.05 was used for all comparisons.

3. Results and Discussions

3.1 Perception Analysis

The distributions of flavor attribute intensity measured by CLT and Gastrograph AI are shown in Figure 1. CLT observed intensity scores across the flavor attributes resemble pseudo-normal distributions. This scale usage pattern is likely the result of forced ratings on attributes in the survey. Forced response questions on specific attributes increases the minimum response value and mean response values as individuals are primed to perceive it. There are no forced-response attributes on the Gastrograph AI System. Respondents can leave unperceived attributes blank resulting in different scale usage patterns and a tail towards zero intensity. The majority of the differences observed between Gastrograph's predictions and CLT perception data result from fundamental differences in survey methodology.

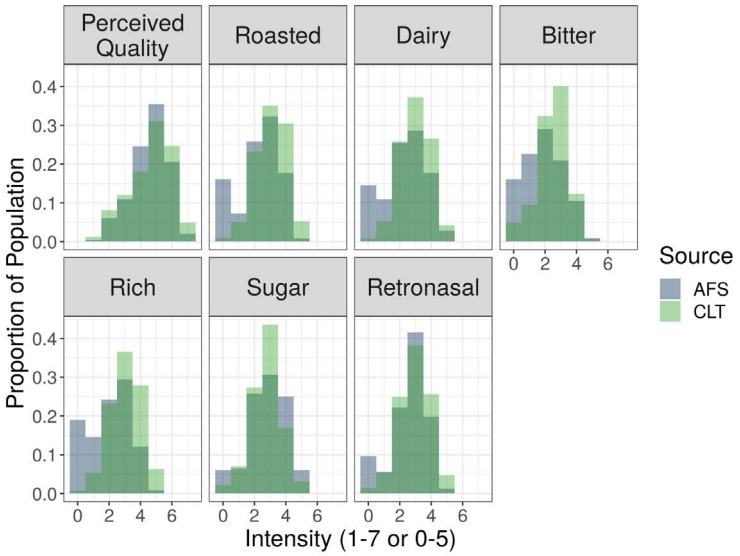

Figure 1. Comparison of the distributions of the Gastrograph AI general Chinese demographic predictions and the CLT data per product.

The mean and standard deviation of the attribute intensities are shown in per product (left) and aggregate (right) between the AFS predictions and the CLT data in Figure 2. The mean values are consistent on average and in 85 % of the product/attributes combinations. Even though the CLT data are systematically higher than the AFS predictions; the higher observed CLT values are caused by the use of forced-response ballots.

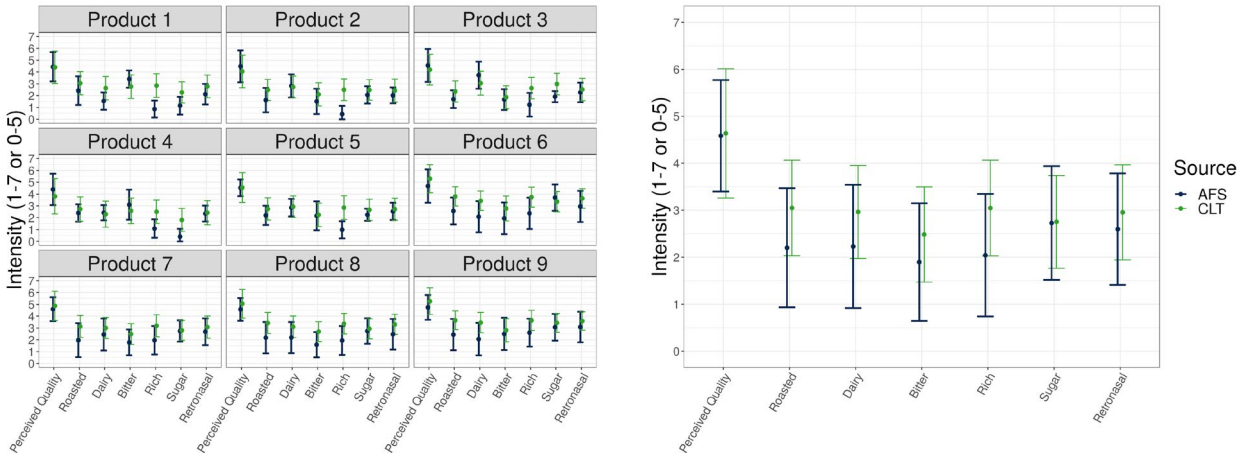

Figure 2. Comparison of intensity of flavor attributes per product (left panel) and aggregated for all products (right panel) between the Gastrograph AI general Chinese demographic predictions and the CLT results. The intensity of Perceived Quality is on a scale of 1 to 7 and the intensity of all other attributes is on a scale of 0 to 5.

In order to compare the perception between the two sets of data, a student t-test is used to determine if there is a statistical difference in predicted mean scores of the attributes between the CLT and Gastrograph AI's data. In Figure 3, no significance difference in attribute intensity was found in roasted, dairy, bitter, and aftertaste among the tested products. There were some minor disagreements in evaluating sweetness among 3 out of 9 products. The largest areas of disagreement are specifically around the term "rich". There is strong evidence to support the offloading of other flavors such as mouthfeel, earthy, and spices not included in the CLT questionnaire onto the category of rich. However, in the Gastragraph review system, this effect is not observed

because of the availability of other terms and the lack of forced response.

As evidence of this, we re-run the test removing 0's from all attributes on data collected on the Gastrograph system. In Figure 3 (right), a marked improvement in the p-value scores was observed. The 0-removed t-test shows the scale usage and sample-bias from forced choice tests where sensory attributes are primed to panelists. The lack of priming on the Gastrograph System is advantageous as it reduces offloading and false positive observations.

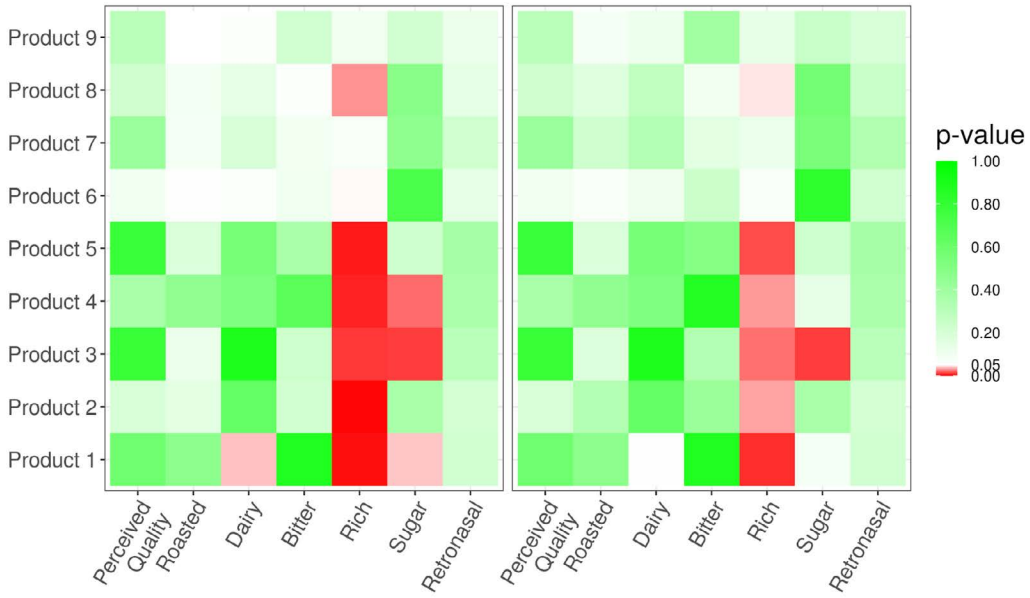

Figure 3. Results of the t-test per product and flavor attributes showing disagreements between the AFS Gastrograph AI general Chinese demographic predictions and the CLT distributions when the p-value is below 5% using the raw data (left panel) and ignoring when no data was entered on the 6 flavor attributes in the Gastrograph AI software (right).

3.2 Preference Comparisons

3.2.1 Distribution Comparisons

Figure 4 shows the distributions of overall liking and PQ scores of the products. The Area Under the Curve (AUC) was calculated as the overlap between the overall liking and PQ score distributions. An average AUC of 0.84 was found for 9 products, indicating that the predictions of Gastrograph AI preference distribution can reproduce the CLT observed overall liking scores.

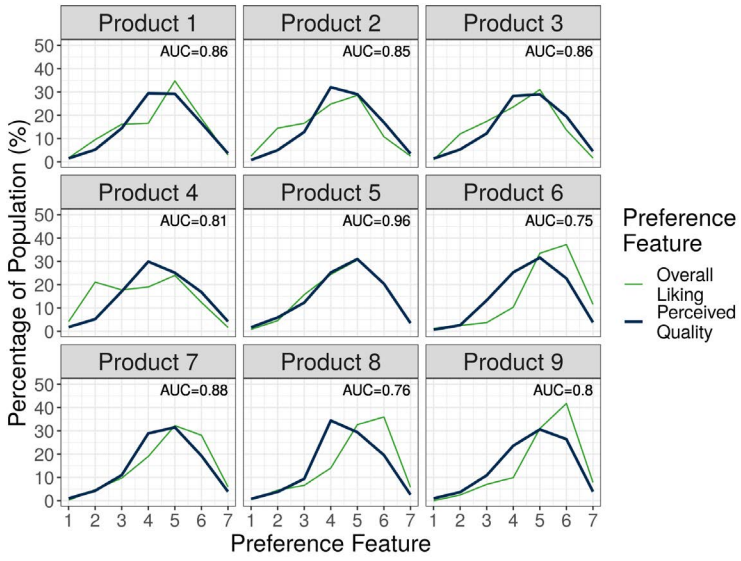

Figure 4. Distributions of flavor attributes and AUC value per product. The mean AUC for the 9 products is 0.84.

K-S test is used to examine if there is a difference between CLT's overall liking individuals and Gastrograph AI's PQ predictions. The cumulative distributions are shown in Figure 5. The p-values are all significantly higher than 0.05, meaning that the PQ and overall liking scores were sampled from the same population.The strong agreement between the distributions of preference for both the AFS predictions and the CLT data were observed for all 9 products.

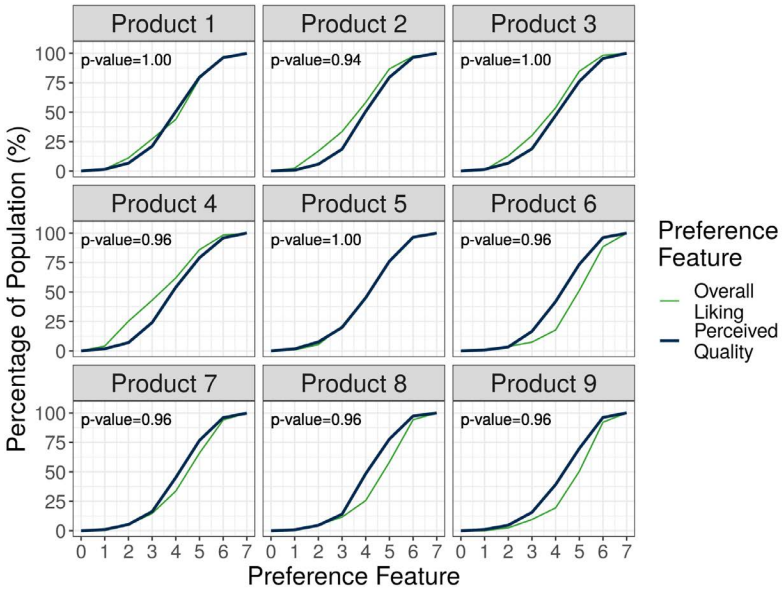

Figure 5. Cumulative distributions of flavor attributes per product and related K-S test p-values.

3.2.2 Mean Score Comparisons

A correlation test was conducted to compare the mean scores of overall liking and PQ. Based on demographic

translation, Gastrograph AI predicted PQ scores for 10 different Chinese demographics. A representative scatter plot is shown in Figure 6. There were strong correlations between Gastrograph AI predicted mean PQ scores and CLT observed overall liking scores. We observed that the highest mean preference score correlation (0.95) was found between the Chinese upper class demographic and the CLT results. This agreement is also reflected in the panel recruitment process because the screener requires a minimal monthly household income of 10,000 RMB per month. Gastrograph AI can segment prediction targets based on the demographic parameters, granting more flexibility when gaining consumer insights.

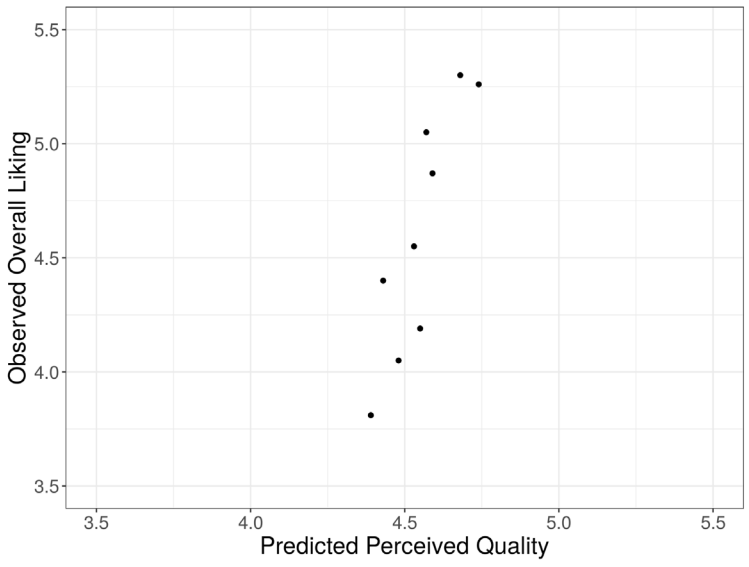

Figure 6. The overall liking data from the CLT as a function of the predicted perceived quality by Gastrograph AI general Chinese demographic predictions are highly correlated with a R2 of 0.77.

3.3 General Discussions

Sensory description is very complex and misinterpretation of attributes may easily occur, meaning that small variations in attributes understanding may lead to big differences in a panel result. The design of the two compared methods implied several variations in terms of data collection and interpretation. The number of attributes varied from 24 with the Gastrograph sensory software to 6 with the CLT design of sensory evaluation. The data collection method used for the CLT, paired with a limited number of attributes, incited panelists to avoid “0” as intensity values. Those minor changes in methodology lead to imprecision in the outcomes of sensory studies.

The Gastrograph System possesses several advantages, including the lack of priming as there is no incitation to rank an attribute intensity if not perceived by the panelist. This allows to reduce offloading and false positive observations. The flavor graph used in the Gastrograph system also covers a broader spectrum of attributes.

The K-S test and the correlation analysis showed that the predicted distribution by the Gastrograph System is drawn for the same population as for the CLT. The majority of the differences seen between Gastrograph AI predictions and observed CLT perception data result from fundamental differences in survey methodology, and Gastrograph AI predictions have been proven at least equivalent to the CLT

with fewer exogenous factors capable of adversely affecting the results.

3.4 Applications of Gastrograph AI

The ability to translate the perception of products across demographics with a limited number of panelists (10 panelists are sufficient to translate perception into a registered demographic) allows for the prediction of multiple new demographics at once. The accurate predictions can save a considerable amount of time compared to conducting CLT panels. Each CLT preparation is made for a one time use corresponding to a specific objective. Gastrograph AI's ability to translate perceptions across demographics can also provide rapid estimation of consumer preferences, allowing systematic data-driven market comparisons.

4. Conclusions and Outlooks

The ability to translate the perception of products across demographics with a limited number of panelists (10 panelists are sufficient to translate perception into a registered demographic) allows for the prediction of multiple new demographics at once. The accurate predictions can save a considerable amount of time compared to conducting CLT panels. Each CLT preparation is made for a one time use corresponding to a specific objective. Gastrograph AI's ability to translate perceptions across demographics can also provide rapid estimation of consumer preferences, allowing systematic data-driven market comparisons.

5. Appendix

5.1 Central Location Test Procedures

5.1.1 Panelists recruitment

A major global market research firm independently recruited 242 panelists with equally distributed across gender and age as shown in Table 1. The panelists also had to consume coffee at least once a week (but preferably at least three times a week), to be high wage earners with a monthly household income superior or equal to 10,000 RMB, and to have experience purchasing Japanese products over the Internet.

5.1.2 Questionnaire

Ajinomoto together created a questionnaire for product reviews structured in 2 parts: a screener interview for demographic information and a product evaluation questionnaire. The detailed example questionnaire can be found in the supplemental materials.

5.2 Mechanics of Gastrograph AI

5.2.1 Demographic translation

One of the goals of this research is to prove the accuracy of the demographic translation. If such can be proven, then the predictive model can be shown to have more than sufficient statistical power (as demonstrated through its predictive capabilities), low data requirements (requiring just 10 reviews to make a prediction on a new product), no requirement or dependency on experimental design (data is always collected in a sequential monadic form on the Gastrograph Sensory System), and no biases from non-representative panelists (as the panel has no requirements from the demographic it is drawn from). The ability to translate perception across demographics allows food and beverage companies to collect data from anywhere and predict how a product would be perceived by a new demographic without ever needing to ship a product or recruit a consumer panelist from the target group – a huge advantage over traditional methods.

Demographic translation is a complex series of models used to project the perception of any given demographic or consumer cohort into the probabilistic perception of another one across differences such as age, sex, race, socio-economic status, and past tasting experience. In a given project, demographic translation can be used to predict the perception of unobserved demographics at a high degree of accuracy. Gastrograph AI can predict the perception of flavors among the AFS database of reviews by translating the aggregated panelist reviews into the perception of the target demographic.

While the demographic translation model itself is proprietary, a conceptual overview of the method and theory is warranted and discussed below. The Gastrograph AI platform models each sensory review by predicting its location in infinite dimensional Hilbert Space. To do so, flavors present but unidentified within the products are predicted before translating the data to the target demographics perception and decomposing the product flavor profiles into their relative constitution parts called "signatures'', defined as any flavor, aroma, or texture. Translation is performed via semi-supervised Fisher nonlinear discriminant analysis for locality-and-covariance preserving projections. Thus, Gastrograph AI can take in data from any random sampling of panelists, drawn from any population, and translate their perception to the target population. This capability eliminates the need for new large-scale consumer testing and for testing products with the target population. Each product flavor profile is represented as a topological subspace of 24-dimensional flavor space.

5.2.1 Demographic translation

A major global market research firm independently recruited 242 panelists with equally distributed across gender and age as shown in Table 1. The panelists also had to consume coffee at least once a week (but preferably at least three times a week), to be high wage earners with a monthly household income superior or equal to 10,000 RMB, and to have experience purchasing Japanese products over the Internet.

5.2.2 Preference prediction

To model reviews of consumers in the target population, Gastrogaph AI builds an empirical preference model for each demographic target and infers its preferences from translated data. The preference prediction model outputs a distribution of perceived quality score (ranged 1 to 7, as 7 being the highest perceived quality) probabilities based on demographically translated reviews. The preference probabilities are calculated from the proportion of decision trees that voted for a given product and are interpreted as the predicted market preference. This is a valid interpretation because of the correct distribution of experience scores in the input dataset, and the understanding that the decision trees model various segments of the population who do not appreciate each flavor attribute equally. This results in a more accurate distribution than direct empirical data collection from standard sized consumer panels can achieve, as it includes the preferences of individuals who would not willingly participate in the tasting of certain products or categories due to known preferences or aversions. Because of this inclusion, Gastrograph AI more accurately predicts the long tale of product rejectors, commonly leading to a marginally lower mean score.

Random forest is a machine learning algorithm that uses hundreds of decision trees, each with a subset both of the variables and of the observations in the input data, to both

predict the output perceived quality and learn the variables of most importance. Decision trees are a set of rules used to classify the data into categories. In this case, the categories are the different possible perceived quality scores on a scale of 1 to 7.

5.2.3 Gastrograph Sensory System

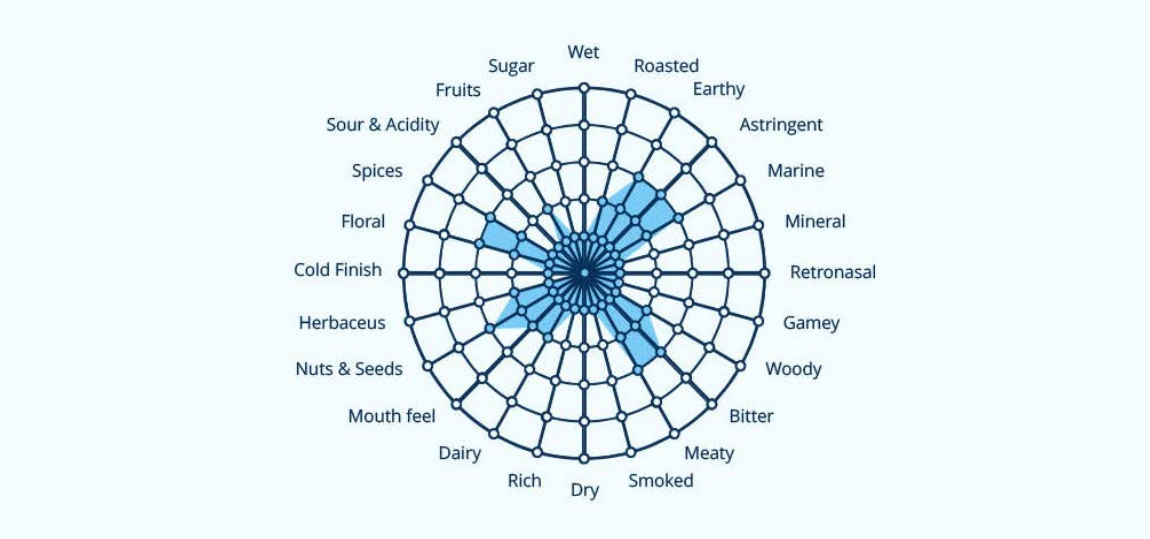

Figure 7. Gastrograph

The Gastrograph is a spider graph consisting of 24 broad-spectrum flavor attributes and somatosensations used to quantify a flavor profile of any homogeneous food or beverage product. Gastrograph variables are a measure

of intensity on a scale of 0 – 5: 0 is not present, 1 is edge of perception ('hint of'), 2 is least intense in product class, and 5 is most intense in product class (Figure 7 and 8). Product class refers to the category of food or beverage the product being sampled belongs to such as beer, coffee, wine, chocolate, etc.

Panelists review products on the Gastrograph Review application, which allows for the selection of different reference favors. Reference flavor selection is done within a Gastrograph variable, so that their semantic meaning is properly categorized. Some reference flavors may be marked in more than one Gastrograph variable. For example, the flavor of pine could both be categorized as 'herbaceous', 'woody', or 'earthy' depending on the expression of pine such as pine needles (herbaceous), resinous pine (woody), or decaying pine wood (earthy).

Additionally reference flavor sentiment can be categorized as 'neutral,' 'positive,' and 'negative.' Upon review submission, panelists are asked to give a Perceived Quality (PQ) score of the product being reviewed. PQ is rated from 1 – 7: 1 is lowest perceived quality and 7 is highest perceived quality of the specific sample being tasted. PQ is not necessarily a hedonic score (enjoyment score), but their 'best assessment of a product's quality' based on past tasting experiences. For low experience panelists in a given product class, PQ is approximately a hedonic score because they don't have the

formative experiences by which to separate their preferences from their ability to assess the product's quality. High experience panelists within product class are better able to assess the quality of a given product, regardless of personal preference due to their past tasting experiences in that product class.

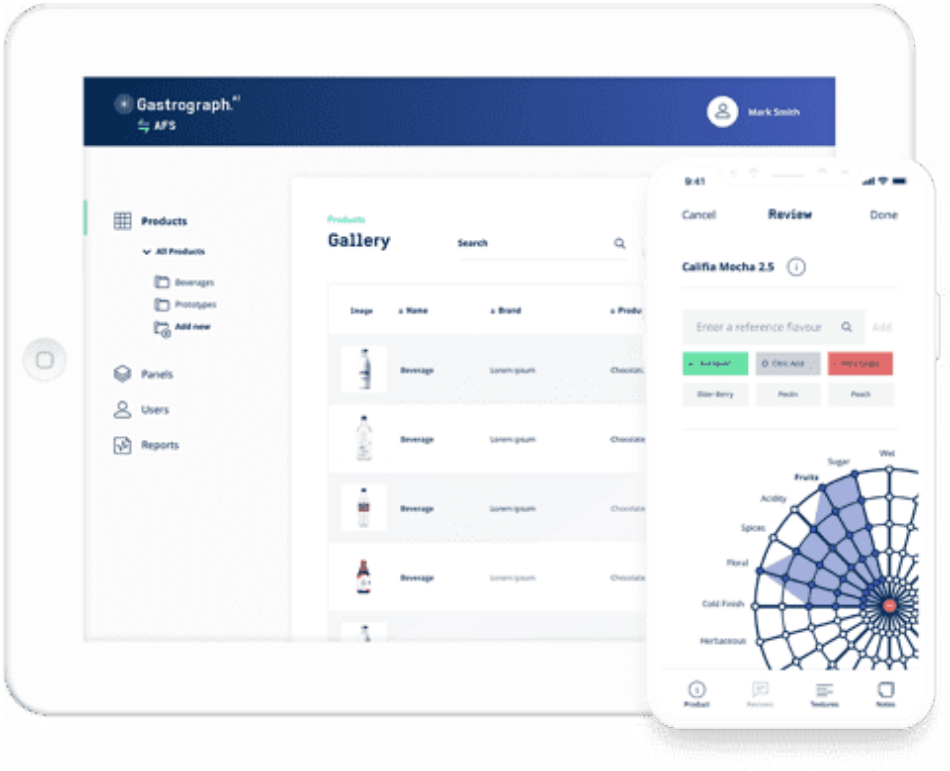

Figure 8. The Gastrograph Review Interface