A Recipe for Creating Recipes: An Ingredient Embedding Approach

Sibel Sozuer[1,2], Oded Netzer[3], Kriste Krstovski[3]

January 5, 2024

Abstract

An idea is a collection of existing concepts or words. What makes an idea original or appealing is how these concepts or words are combined in the context in which they appear. Similarly, a food recipe is a combination of ingredients, and it is often evaluated based on how these ingredients fit together to form the whole. In this research, we leverage representation learning methods, specifically word embeddings, to measure the fit among ingredients in the recipe and capture the possibly complex interactions between these ingredients. Using a large-scale online recipe dataset with over 57K recipes, we investigate how the fit between the ingredients relates to recipe popularity (trial) and favorability (ratings). Counter to prior research on creativity, which primarily suggests that creativity is mostly associated with positive outcomes, we find that recipes with unique ingredients have lower trial, but higher ratings given trial. We also find that high fit among ingredients promotes both trial and ratings. We use these findings to develop a generative recipe tool that suggests recipe improvements by adding, removing, or substituting ingredients (**http://recipecreativity.com/**). We validate our proposed recipe tool using online panel experiments.

Keywords: creativity, idea generation, food recipes, computational creativity, word embeddings, representation learning, online applications

[1] Sibel Sozuer (sibel_sozuer@kenan-flagler.unc.edu) is the corresponding author.
[2] Kenan-Flagler Business School, University of North Carolina at Chapel Hill
[3] Columbia Business School, Columbia University

"Who knew that the hardest part of being an adult is figuring out what to cook for dinner every single night for the rest of your life." – Anonymous

An average person makes about 200 food-related decisions every day (Wansink and Sobal 2007). One of the most frequent decisions a consumer faces daily is deciding what to cook for a meal. These decisions occur much more frequently than, for example, the decisions about which food products to purchase in the store, a topic commonly investigated in the marketing literature. Indeed, a recent survey reveals that most people (47%) cook at home three to four days a week on average, and unsurprisingly, a great majority (70%) are frequently bored cooking the same recipes (Buiano 2022; Jennie-O Turkey Store 2022). To add variety to their cooking, people often experiment with recipes (60%) and introduce unusual ingredients to the recipe (50%). In addition, a third of the respondents mention that they search the internet for new recipe inspirations. These patterns of behavior may explain the popularity of online recipe platforms such as allrecipes.com and foodnetwork.com and subscription-based meal kit delivery services that provide subscribers with recipes and ingredient kits (e.g., Blue Apron, Hello Fresh).

The prominence of making food recipe choices raises the question of what makes for a good recipe. Previous research on idea generation describes an idea as a collection of existing concepts or words. What makes an idea creative or successful is not only the concepts and words themselves but also how they are combined together in the context in which they appear. A comparable reasoning could apply to food recipes and recipe ingredients. A food recipe is a collection of ingredients and is likely to be perceived as novel or appealing based on the combination of its ingredients. To put it another way, the interactions between ingredients are often

just as important as the ingredients themselves because their gestalt arises from the fit of all ingredients together.

Accordingly, the objective of this paper is to examine the impact of the novelty and fit among recipe ingredients on recipe popularity and favorability evaluation and leverage these insights to develop an online app that helps consumers improve food recipes by adjusting the ingredients in the recipes. To this end, we propose an approach that builds on representation learning to combine recipe ingredients for generating promising new recipe ideas that are likely to be successful. We use a crowdsourced recipe dataset on Allrecipes.com to identify characteristics and ingredient combinations of successful recipes based on user engagement metrics. Then, we take these findings one step further to propose a generative recipe application that provides specific ingredient suggestions to improve recipe ideas.

To investigate the degree to which ingredients "fit together" in a food recipe, we build on research in text analysis and representation learning, specifically word embeddings. Word embeddings are used to represent the meaning of a word based on the surrounding words and help identify semantically similar and complementary words based on these representations. Similarly, we represent each ingredient based on the other ingredients in the same recipe and use these representations to quantify the fit of each ingredient in a specific recipe based on its context in that recipe (i.e., the other ingredients in that recipe) to determine the novelty in the combination of recipe ingredients. This can then help us find similar and complementary ingredients to add, or substitute in a recipe. We define two types of ingredient-based novelty measures: 1) the extent to which the ingredients in the recipe form a coherent combination derived by using embedding-based representations of ingredients; and 2) the prevalence or uniqueness of each ingredient across

recipes in the same category. A recipe can be novel by using novel ingredients (e.g., za'atar) or by combining unexpected ingredients togethers (e.g., chicken and chocolate). We then investigate how these measures of ingredient fit and uniqueness are linked to recipe performance in terms of popularity (recipe trial) and favorability given trial (ratings).

To device our fit measures and empirically test the relationship between food recipe ingredients and recipe performance, we collect an extensive dataset of more than 57K recipes from the popular recipe website Allrecipes.com. We extract information on each recipe's ingredients and cooking instructions, as well as recipe performances in terms of the number of people who tried it and recipe ratings. Investigating the relationship between our novelty measures of ingredient fit and uniqueness and recipe performance on the platform, we find that novelty may not always be preferable. Consumers are, on average, more likely to try recipes with common or familiar ingredients that fit well with one another. However, once consumers have tried the recipes, they tend to more favorably evaluate (higher star ratings) recipes that use unusual or less common ingredients—but these ingredients need to still fit well together. This result highlights the importance of capturing the gestalt fit among the recipe ingredients, which we achieve through a representation learning approach.

Importantly, we leverage our model of ingredient fit and uniqueness and the findings from the Allrecipes.com analysis to create an interactive generative recipe application (available online at http://recipecreativity.com) to allow average consumers to improve their recipe by adding, removing, or substituting ingredients. We also provide several validation exercises comparing the proposed recipe tool's recommendations for ingredient modifications to human preferences and recommendations. We find that the tool's recommendations are well-perceived by consumers, and

it produces better recommendations than most humans, including self-identified food domain experts.

The contribution of this research is threefold. First, this work relates to and augments the growing stream of literature that leverages machine learning and text analysis for idea generation (e.g., Kelly et al. 2021; Packard et al. 2016; Toubia and Netzer, 2017; Uzzi et al. 2013). We extend the research by employing a representation learning approach to capture the gestalt nature of ideas in an ideation domain where the interaction between idea ingredients is just as important as the ingredients themselves, food recipes. Second, whereas most existing research on ideation utilizes subjective measures of creativity or anticipated adoption, we include measures of product trial and favorability evaluation given trial. We find that novelty differentially affects product trial and product evaluation. Finally, we develop a useful tool for consumers, both novices and experts, to assist with one of the most common tasks in our daily lives: cooking.

In the next section, we discuss related work on idea generation, particularly in food recipes. Then, we describe the data from the popular online recipe website, Allrecipes.com, detail the embedding approach to recipes, and introduce the recipe novelty measures. Next, we investigate the relationship between recipes' novelty and recipes' popularity and favorability evaluations. We then leverage these results to introduce the generative recipe online app and conduct validation studies comparing the performance of the proposed app to human preferences and recommendations. Finally, we conclude with the implications of this work for idea generation and food recipe recommendations, and directions for future work.

Idea Generation and Creativity in Food Recipes

Generating New Recipe Ideas

The topic of recipe recommendations systems has received some attention primarily in the computer science literature (Anderson 2018; Trang Tran et al. 2018). This line of research has adopted typical recommender system tools such as collaborative filtering that focus on user preferences and/or domain knowledge (i.e., nutrient or flavor profiles, ontologies) to model the relationship between different ingredients and/or provide recipe suggestions. Collaborative filtering models recommend a recipe to a user based on the user's past behavior (e.g., previous adoptions or ratings) and the behavior of other users who are similar to the focal user. On the other hand, content-based filtering relies on external characteristics of the recipe and its ingredients and user preferences for these characteristics, to recommend other recipes to the user. Much of this work has focused on recommending recipes to individuals (e.g., Freyne and Berkovsky 2010). Our work differs from these studies in the use of an embedding approach to measure ingredient fit and in recommending replacements, additions, and subtractions of recipe ingredients for existing recipes to create new recipe inspirations rather than suggesting recipes to users.

Computational creativity, a relatively new subfield of artificial intelligence, uses automated systems to generate ideas that are considered creative. There have been some recent applications of computational creativity in evaluating and generating food recipes. Several papers have examined the association between ingredients using common flavor compounds or common food ontologies such as food groups or growing region of the world (e.g., Ahn et al. 2011; Amorim et al. 2017; Pinel, Varshney and Bhattacharjya 2015; Morris et al. 2012; Varshney, Wang, Varshney 2016; Varshney et al. 2019). To explore the relationship between ingredients and recipe characteristics and generate new recipes, these studies use a combination of network analysis,

machine learning, and genetic algorithms. A limitation of these approaches is that they require a priori domain knowledge in order to link ingredients to flavors or other ontologies. Additionally, the generation of recipes is limited to such a priori categorization groups (e.g., tofu as a soy product and chicken as a meat product). On the other hand, the representation learning-based approach allows us to capture these relationships by understanding the context of the ingredients. For example, if two ingredients (e.g., tofu vs. chicken) often appear with other ingredients that have similar characteristics (e.g., almond milk + peanuts vs. whole milk + cashews), they would have similar representations and appear closer in the embedding space.

More similar to our line of work, several papers have looked at co-occurrences among recipes and/or recipe reviews (e.g., Cromwell, Galeota-Sprung and Ramanujan 2015; Teng, Lin and Adamic 2012). However, by focusing on pairwise ingredient co-occurrence rather than the embedding space of the ingredients, such approaches fail to capture the gestalt of ingredient fit and the interactions among ingredients. Additionally, coming from a computer science domain, these papers do not focus on the relationship between the novelty of recipe ingredients and consumer preferences for these recipes.

The Process of Ideation Through Representation Learning

The Geneplore model of idea generation proposed by Finke, Ward and Smith (1996) presents idea generation as an iteratively performed two-stage process that consists of a generative phase and an exploratory phase. Related to our application, Moreau and Dahl (2005) illustrate this process in the context of creating a dinner recipe. In the generative phase, people start with a set of ingredients (e.g., peanut butter, spaghetti noodles, carrots, etc.) that form a pre-inventive structure for the dinner recipe. In the exploratory phase, this pre-inventive structure is interpreted

as satisfactory or unsatisfactory and modified, if necessary, by using only a subset of the ingredients and/or adding new ingredients to achieve a satisfactory dinner recipe. These pre-inventive structures may include representations or mental models of specific ingredients (e.g., spaghetti noodles) or conceptual ingredients (e.g., a type of pasta) and category exemplars (e.g., mac and cheese) (Burroughs, Moreau and Mick 2008; Ward 2001).

The idea of representation or mental models of specific ingredients is related to a very different stream of literature in machine learning and textual analysis, representation learning. In representation learning, the semantic similarity between words is determined based on their distributional properties in large samples of text data, considering the words that commonly co-occur with these words (Firth 1957; Harris 1954). Accordingly, words with similar meanings that often co-occur with similar other words are close to each other in the embedding space. For example, based on the embedding approach, the word "queen" may have a similar representation to the word "king" because both appear in similar contexts, i.e., they are surrounded by similar words (e.g., "crown", "palace", "royalty").

Geneplore's two-step creative process lends well to the embedding approach for creating a recipe. In this approach, each ingredient is represented with a vector that captures its unique features, learned through how it is used across different recipes. The probability of having an ingredient in a recipe depends on its vector representation and the vector representations of the other ingredients in the same recipe. As a result, ingredients that frequently appear in similar contexts (i.e., other ingredients in the recipe) will have more similar vector representations (e.g., whole milk and skim milk) and will appear closer in the embedding space whereas ingredients that fit well together and create satisfactory combinations will have a higher probability of appearing

together. Studying the relationship between the characteristics of recipe ingredients and recipe success further informs the development of an automated tool to assist human creators in the exploratory phase by identifying promising recipe ideas and guiding them to modify their recipe ingredients for improvement.

Measuring Ingredient Fit and Novelty

The literature on creativity consistently supports that creativity stems from a balance between novelty and conventionality (Giora 2003; Toubia and Netzer 2017; Ward 1995). To automatically assess the degree of novelty in the idea, researchers have developed natural language processing (NLP) and/or network analysis-based approaches to automatically evaluate idea features. For example, Uzzi et al. (2013) examine references of scientific papers, comparing the frequency of each co-cited journal pair across all papers with the frequency distribution created in a randomized citation network. Packard et al. (2016) represent movie "ingredients" (i.e., cast and crew) based on their positional and junctional roles in the film industry network. Toubia and Netzer (2017) evaluate the semantic similarity between pairs of words in ideas by the co-occurrence of these words in a typical text in the idea domain. Kelly et al. (2021) use cosine similarity among words in a patent document relative to the previous patents to measure the novelty of the patent. Wei, Hong and Tellis (2022) couple the Word Mover's Distance with Google's pre-trained Word2vec model to find the similarity between projects.

A limitation of these approaches is that they look at one pair of idea ingredients at a time and/or fail to evaluate how each idea ingredient fits in the context of the other ingredients in the idea. While the assumption of looking at how each pair of words in the idea fit together may be reasonable for creative ideas, it may not hold true for food recipes, in which the gestalt of how

well all ingredients in the recipe fit together is crucial. For example, lime goes well with both cilantro and pineapple; adding cilantro may work if you are making a savory lime dish (containing salt) whereas it may not work if you are making a sweet lime cake (containing sugar). The embedding approach we propose captures how an ingredient interacts with the entire set of ingredients in the recipe rather than relying on pairwise associations between ingredients. Moreover, for unique ingredients, the lack of co-occurrence in a dataset does not necessarily mean that the combination is novel. If we observe that walnuts and dates are often used together in many recipes and that a unique ingredient such as Brazil nuts has never been used with dates in a recipe corpus, the embedding approach allows us to infer the fit of dates and Brazil nuts together based on the context in which walnuts and Brazil nuts appear in other recipes. Additionally, when recommending substitutions, the embedding approach helps evaluating the similarity and substitution of ingredients, facilitating substitutions within a "family" of ingredients.

The notion that the success of food recipes depends on the fit among ingredients is aligned with how famous chefs and restauranteurs think about food recipes (Page 2017). For example, Wolfgang Puck of Spago draws similarities between cooking and art and states that even though the flavors are limited, how they are combined sets one apart; Daniel Patterson of Coi and LocoL emphasizes that "just because two components are amazing doesn't mean that combining them will work"; and the famous bartender Audrey Saunders of Pegu Club tells aspiring chefs to "make sure (your) ingredients dance well together" (Page 2017, p. 27-28). Therefore, when studying ideation in the context of food recipes, it is crucial to take a holistic perspective of the interaction among ingredients. Our embedding approach allows us to model this context-dependent relationship of ingredients in a way that has not been done in previous work on ideation.

The Impact of Novelty on Behavioral Outcomes

Researchers have looked at different measures to assess idea quality such as customer appeal (Dahl, Chattopadhyay and Gorn 1999), peer or expert evaluation (Kornish and Ulrich 2011, 2014), purchase or adoption intent (Girotra, Terwiesch and Ulrich 2010; Kornish and Ulrich 2014; Luo and Toubia 2015), predicted and actual sales (Kornish and Ulrich 2014), funding performance (Wei, Hong and Tellis 2022) and the number of citations (Kelly et al. 2021; Singh and Fleming 2010; Uzzi et al. 2013).

In this research, we consider two outcome measures: idea popularity, which is an idea's actual adoption by consumers (i.e., trial), and idea favorability, which is a subjective assessment of how much people like the idea after trial (i.e., ratings). We demonstrate that recipe novelty has differential effects on these two measures. For example, consumers may be reluctant to try food recipes with unique ingredients, but if these ingredients fit well within the recipe, they may be more likely to favorably evaluate the recipe once they try it.

Data and Information Extraction

We use recipe data from Allrecipes.com, the world's largest food community with 1.5 billion annual visits made by 60 million unique visitors.[4] In Allrecipes.com, users can find and review recipes that are created by other users and share their own recipes. We obtained more than 57K public and kitchen-approved recipes[5] that were posted on the website before November 2019. For each recipe, we have basic information (recipe title, recipe description, recipe creator, recipe category, and number of people who tried, rated, and reviewed the recipe), the list of ingredients,

[4] https://www.meredith.com/brand/allrecipes (03/09/2021).
[5] We use only recipes that are accessible through the main page, category pages and search function by all users. Recipes are reviewed and approved by the website based on completeness and redundancy etc.

cooking instructions, and average star ratings. We further obtain user profiles about the recipe creators (number of followers, number of personal recipes, number of reviews, number of favorite recipes). After we remove ingredients that appear in less than 10 recipes and the recipes that contain those ingredients (as described in Web Appendix A), we have more than 57K recipes in the dataset (see Table 1 for summary statistics of the recipe dataset).

Table 1: Summary Statistics for the Recipe Data (N = 57,709 Recipes)

	mean	stdev	min	median	max
Number of trials	127.78	601.23	0	20	34,233
Number of ratings	82.12	364.27	0	13	15,402
Average rating	4.31	0.54	1	4.4	5
Preparation time (min)*	18.62	165.38	1	15	30,240
Total time (min)*	167.35	1,914.64	1	50	211,700
Number of ingredients	8.85	3.56	2	8	42
Number of preparation steps	3.39	1.78	1	3	27

*The number of recipes reporting preparation time and total time is 45,317 and 45,510, respectively.

The website classifies recipes into different categories and subcategories, with a hierarchy of up to 6 levels of subcategorization (e.g., Side Dish > Sauces and Condiments > Sauces > Pasta Sauces > Creamy > Alfredo Sauce). We mainly leverage the 2nd level of the hierarchy for this analysis because it provides a good balance of differentiation between recipes and complexity of coding these subcategories (see Appendix A for main category summary statistics and a list of all subcategories). Overall, we have 138 2nd-level subcategories.

The creator of each recipe provides a list of ingredients to be used, and Allrecipes.com matches each ingredient description to an ingredient ID. Overall, Allrecipes.com includes 5,966 unique ingredient IDs. There are at least three issues with directly using these ingredient IDs and ingredient descriptions for this analysis. First, the text descriptions for the same ingredient ID may

be slightly different in different recipes (e.g., ingredient ID 1767 appears as "2 cups unbleached all-purpose flour" and "3 tablespoons all-purpose flour"). Second, ingredients with different ingredient IDs often have similar or even identical text descriptions (e.g., ingredient IDs 1684, 1767 and 19183 are all described as "all-purpose flour" in the text). Third, the list of ingredient IDs has a long tail, where only about 35% of ingredient IDs have been used in at least 10 recipes and excluding recipes with ingredients that appear in less than 10 recipes leads to removing 8.2K recipes from the dataset (i.e., over 14% of the data). Therefore, we identify the most representative label for each ingredient ID using text analysis and group them based on the similarity of their labels. We remove the ingredients that appear in less than 10 recipes after grouping. As a result, we remove from the dataset 279 ingredient IDs that can't be grouped with other ingredient IDs and appear in less than 10 recipes as well as 666 recipes that include them. After completing these steps, we reduce the number of unique ingredients to 1,249, each of which appears in at least 10 recipes, and have a total list of 57,709 recipes that are constructed from these ingredients (see Web Appendix A for details).

Another important recipe feature that can affect users' attitudes towards the recipe is the preparation technique (e.g., boil, grill, bake), which can be extracted from preparation instructions. We automatically text-mine the preparation instructions using the Python *spaCy* package (Honnibal and Montani 2017). We use part-of-speech tagging to identify the verbs used in the preparation instructions. We obtain a total of 1,193 verbs in the lemmatized form across recipes.

Empirical Approach

To identify the coherence of the ingredients in each recipe, we use embeddings to learn the vector representations of each ingredient. Using these representations, we derive a measure of fit

for each ingredient relative to the other ingredients in the recipe. We also construct measures that capture the prevalence or uniqueness of the recipe ingredients and preparation methods.

Embedding Approach to Recipe Ingredients

Word embeddings are a collection of self-supervised representation learning methods for analyzing language and capturing semantic similarity between words. Unlike typical bag-of-words approaches (e.g., tf-idf feature vectors), which assume that the order in which words appear is irrelevant, word embeddings represent each word in the context of the words that appear in its vicinity, thus taking into account the context of the word. Word embeddings are based on the distributional hypothesis which states that the words that occur in the same contexts tend to have similar meanings (Firth 1957; Harris 1954). Since the introduction of word embeddings (Bengio et al. 2003; Rumelhart, Hinton and Williams 1986), many variants and extensions have been developed (e.g., Levy and Goldberg 2014; Mikolov et al. 2013a, 2013b; Pennington et al. 2014; Ruiz, Athey and Blei 2020).

The main idea underlying word embeddings is to represent each word with two feature vectors: an embedding vector and a context vector. Both are multidimensional vectors with the same length which is much smaller than the total number of words in the vocabulary. The embedding vector represents the focal word and the context vector represents the words that surround the focal word in a similar embedding space. This approach attempts to maximize the conditional probabilities of the observed text in a given corpus by combining the embedding vector of the focal word with the context vectors of the surrounding words. Recently, Rudolph et al. (2016) develop the exponential family embeddings model which extends the idea of word

embeddings to other types of data (e.g., Gaussian distribution for real-valued data, Poisson distribution for count data, Bernoulli distribution for binary data).

We borrow from that text analysis literature to model the appearance of ingredient in recipes. To evaluate the context-dependent relationship among ingredients, we focus only on whether the ingredient is used in the recipe and ignore the exact quantity (e.g., 1 tablespoon, 2 cups, etc.). Our reasoning for ignoring amount of the ingredient is threefold. First, the existence of an ingredient in the recipe is a more important factor than the amount of the ingredient, especially for consumer to decide if they want to make or eat that recipe. Second, the contribution of the ingredient in the recipe may not always be attributed to the amount used in the recipe. For example, 1 tablespoon of ground cinnamon may be more potent than 1 cup of rice flour in a recipe. Third, the conversion across different units for quantities is not straightforward (e.g., 1 cup sugar = 200 grams, 1 cup oatmeal = 100 grams).

To model the occurrence of an ingredient in a recipe, we use Bernoulli embedding, which is a specific case of the exponential family embeddings. The Bernoulli embedding model can be defined as follows. Assume there are N ingredients across R recipes. Each recipe r is a collection of ingredients. The data consist of an $N \times R$ matrix with binary entries where $x_{nr} = 1$ indicates that ingredient n appears in recipe r. Each ingredient n is represented by an embedding and a context vector (ρ_n, α_n) of length t, where t defines the number of dimensions in the embedding space which is determined by the researcher, often based on model fit criteria.

We define the context of an ingredient n in recipe r as all the other ingredients in recipe r, $c_{nr} = \{m | x_{mr} = 1; m = 1, \dots, N;\ m \neq n\}$. We represent the context c_{nr} by averaging the context

vectors of all the other ingredients in recipe r: $g(\boldsymbol{\alpha}, c_{nr}) = \frac{1}{|c_{nr}|} \sum_{m \in c_{nr}} \alpha_m$. The conditional distribution of ingredient n appearing in recipe r given its context c_{nr} is given by:

$$x_{nr}|x_{c_{nr}} \sim \text{Bernoulli}\left(\sigma\left(\rho_n^T g(\boldsymbol{\alpha}, c_{nr})\right)\right),$$

where $\sigma(.)$ is the logistic function. Thus, intuitively an ingredient is more likely to appear in a recipe if its embedding vector is similar to the context vectors of the other ingredients in the recipe (i.e., the inner product of the two vectors is high). Summing across recipes and ingredients, the objective function contains conditional log-likelihoods of observed ingredients in recipes:

$$\mathcal{L}(\boldsymbol{\rho}, \boldsymbol{\alpha}) = \sum_{r=1}^{R} \sum_{n=1}^{N} \log p\left(x_{nr}|x_{c_{nr}}\right) + \log p(\boldsymbol{\rho}) + \log p(\boldsymbol{\alpha}).$$

Stochastic gradient descent with Adagrad (Duchi, Hazan and Singer 2011) is used to estimate the embedding and context vectors that maximize the objective function. We adopt the default settings of Rudolph et al. (2016) for inference to reduce the computational cost: computing the exact gradient for nonzero entities – ingredients that appear in a given recipe –, subsampling 10 zero entities instead of computing the exact gradient, and downweigh the contribution of zeros by 0.1 due to the sparsity of the data. We split the dataset of 57K recipes into 90% training, 5% validation (containing held-out ingredients for some recipes in the training sample), and 5% test samples (containing recipes that do not appear in the training sample) to determine the size of the latent space. We vary the size of the embedding space from 20 to 100 by increments of 10 at 2500 iterations and use validation and test performance to determine the final size of the embedding space. Because the log-likelihood performance on the validation sample and test sample does not significantly improve as we increase the size of the latent space above 80 dimensions, we choose $t = 80$ as the size of the latent space.

Figure 1 presents a 2-dimensional illustration of ingredients in the embedding space using t-Distributed Stochastic Neighbor Embedding (t-SNE), which is a tool to visualize high-dimensional data (Van der Maaten and Hinton 2008). The t-SNE plot shows that ingredients that are close substitutes appear closer together in the embedding space. For example, feta cheese is closest to goat cheese and blue cheese, which are also soft and fresh cheeses, but feta appears in close proximity to aged (e.g., parmesan, cheddar) and processed cheeses too (e.g., American cheese, processed cheese). Similarly, rice appears closest to white rice, brown rice, and quinoa but is also in close proximity to other carb side dishes such as egg noodles, spaghetti, macaroni, and pasta. As another example, we observe that shrimp appears most closely to another shellfish, crabmeat, but it is also near fish such as tuna and salmon and somewhat farther from (but still nearby) turkey, chicken, beef, and pork, which are all reasonable substitutes for seafood in many dishes. This analysis provides face validity to the embedding analysis.

To further examine whether the embedding model captures meaningful relationships among ingredients, we compare ingredient substitutions suggested by Allrecipes.com[6] to those suggested by the embedding model (see Table 2), where a suggested substitution by the Embedding Model is defined as an ingredient whose embedding vector has a cosine similarity of 0.45 or higher compared to the original ingredient. Indeed, the embedding model captures most of the substitution suggestions offered by Allrecipes.com and provides additional substitutes that seem reasonable.

[6] https://www.allrecipes.com/article/common-ingredient-substitutions (03/09/2021).

Figure 1: Two-Dimensional Projection of the Embedding Space (t = 80) for Ingredients that Appear in at least 250 Recipes Using t-SNE

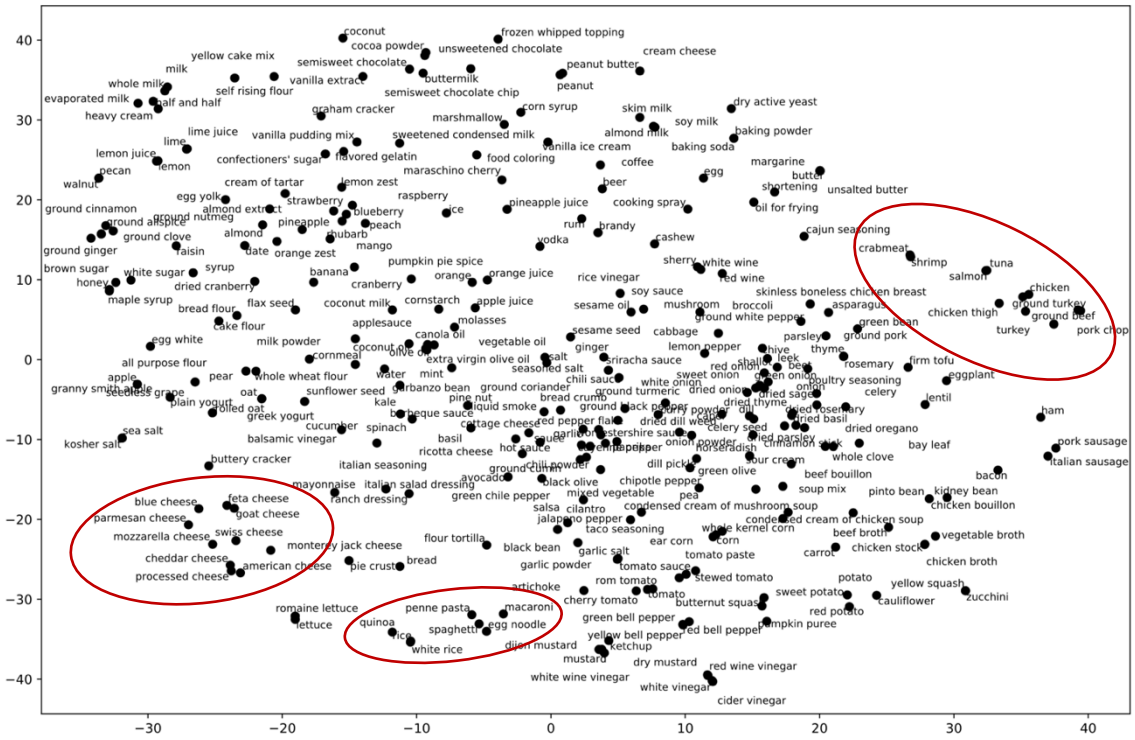


Table 2: Ingredient Substitutions Suggested by Allrecipes.com vs. Embedding Model

Source of Ingredient Substitution	**Green onion**	**Evaporated milk**	**Chicken/beef stock**	**Margarine**
Allrecipes.com	Onion Leek Shallot	Light cream	Chicken bouillon Beef bouillon	Shortening Butter Vegetable oil Lard
Embedding Model	Red onion Sweet onion **Onion** White onion **Shallot** Chive **Leek** Dried onion	**Light cream** Half and half Butterscotch chip Skim milk Milk Whole milk Almond milk	Chicken broth Beef broth Vegetable broth Vegetable stock Browning sauce **Chicken bouillon** Beef base Beef bone Turkey broth	**Butter** Unsalted butter **Shortening** Peanut oil Canola oil

Recipe Novelty Measures

Ingredient fit measure

We define the ingredient fit measure based on the similarity between the vector representation of an ingredient and that of other ingredients in the recipe. It is operationalized as the cosine similarity between an ingredient's embedding vector and its context representation (the average of context vectors of the other recipe ingredients). For each ingredient n in a given recipe r, we then define ingredient fit as $\theta_{nr} = \frac{\rho_n^T g(\alpha, c_{nr})}{\|\rho_n\| \|g(\alpha, c_{nr})\|}$.

Table 3: "Molasses Cookies" and "Cashew and Peanut Butter Cookies" in the "Desserts > Cookies" Subcategory

Molasses Cookies V		
Ingredients	Ingredient frequency	Ingredient fit
all-purpose flour	69.94%	0.1297
white sugar	63.94%	0.1340
baking soda	41.97%	0.1578
shortening	16.18%	0.0789
ground cinnamon	15.32%	0.1943
ground nutmeg	5.06%	0.1303
ground ginger	5.03%	0.1645
molasses	4.50%	0.1080
ground clove	4.35%	0.1905
buttermilk	1.74%	0.0450
MEAN	22.80%	0.1333
STDEV	26.10%	0.0471

Cashew and Peanut Butter Cookies		
Ingredients	Ingredient frequency	Ingredient fit
egg	65.03%	0.1168
white sugar	63.94%	0.0962
baking soda	41.97%	-0.0113
brown sugar	37.56%	0.0246
peanut butter	13.68%	0.0155
margarine	5.65%	0.0016
tapioca	0.32%	-0.1443
potato flour	0.09%	-0.3894
cashew butter	0.09%	-0.4373
corn flour	0.03%	-0.1608
MEAN	22.84%	-0.0888
STDEV	26.86%	0.1928

Looking across the ingredients in a recipe, we compute the mean and standard deviation of ingredient fits in each recipe. A recipe with high ingredient fit mean value can be described as conventional or coherent whereas a recipe with a lower ingredient fit mean value has a more novel or incoherent ingredient combination. Note that a recipe can have a novel ingredient combination even if the ingredients themselves are common (e.g., cilantro and sugar) and vice versa. For

example, the "Molasses Cookies" recipe in Table 3 contains some unusual cookie ingredients like molasses, clove or ground ginger, but the fit of these unusual ingredients is fairly high (an average fit score of 0.1333). On the other hand, the "Cashew and Peanut Butter Cookies" recipe includes some ingredients that are novel (e.g., tapioca and cashew butter), but also unique *combinations* of common ingredients (e.g., baking soda and peanut butter). While both recipes have similar overall uniqueness or prevalence of the recipe ingredients themselves, as we discuss next, they differ substantially in terms of the embedding-based ingredient fit measure (see Table 3).

Ingredient and verb frequency measures

In addition to assessing the recipe novelty via the gestalt fit of its ingredients, we also consider a more traditional measure of prevalence or uniqueness of recipe ingredients and preparation methods. In measuring the frequency of ingredients, we compute subcategory-adjusted ingredient frequencies because the usage of ingredients may be different across subcategories. For example, brown sugar (ground black pepper) is used in 37.6% (0.2%) of the recipes in the "Desserts > Cookies" subcategory, whereas it is used in 6.6% (29.8%) of the recipes in the "Meat and Poultry > Chicken" subcategory. Similarly, we compute subcategory-adjusted frequencies of extracted verbs from recipe preparation instructions (e.g., mix, bake, boil etc.).

Looking across the ingredients (verbs) in the recipe, we compute the mean and standard deviation of ingredient (verb) frequencies in each recipe. Recipes with common ingredients (preparation techniques) will have high ingredient (verb) frequencies whereas recipes with more unconventional ingredients (preparation techniques) will have lower ingredient (verb) frequencies. For example, the "Congo Bars" recipe contains a set of relatively conventional ingredients whereas "Coconut Date Bars" includes more uncommon ingredients (e.g., cashew) in the "Desserts >

Cookies" subcategory even though overall ingredient fits are very similar for both recipes as illustrated in Table 4.

Table 4: "Congo Bars" and "Coconut Date Bars" in the "Desserts > Cookies" Subcategory

Congo Bars		
Ingredients	Ingredient frequency	Ingredient fit
all-purpose flour	69.94%	0.1297
brown sugar	37.55%	0.1009
butter	56.91%	0.2116
egg	65.02%	0.0915
semisweet chocolate chip	22.65%	0.0346
MEAN	50.42%	**0.1137**
STDEV	19.83%	**0.0647**

Coconut Date Bars		
Ingredients	Ingredient frequency	Ingredient fit
date	3.52%	0.1534
coconut	11.47%	0.1258
almond	2.88%	0.1552
cashew	0.71%	0.0815
coconut oil	1.91%	0.0696
MEAN	4.10%	**0.1171**
STDEV	4.26%	**0.0399**

Recipe Success Measures: Popularity and Favorability

We measure recipe popularity by the number of consumers who have tried the recipe based on Allrecipes.com. Users self-report this measure by clicking "I made it" button on the recipe page and then receive a prompt to rate and review the recipe. Due to the long tail of the trial distribution, we use log trial as a measure of recipe popularity. The recipe favorability given trial is reflected in the distribution of the recipe's star ratings. Due to the large proportion of positive reviews (Schoenmueller, Netzer and Stahl 2020), we use the percentage of 5-star ratings as a measure of recipe favorability.

Results

Relationship Between Recipe Novelty and Recipe Popularity and Favorability

We investigate the relationship between recipe novelty and recipe popularity and favorability by regressing each of the outcome variables on the derived novelty measures related

to ingredient fits and ingredient and verb frequencies, controlling for a host of recipe and recipe creator characteristics. Each observation corresponds to a recipe in our dataset.

In this analysis, we reduce the sample of recipes to subcategories with more than 50 recipes (110 subcategories out of 138 subcategories) to allow inclusion of subcategory fixed effects, and recipes with at least one review and information on preparation and total time. After we remove recipes with missing information on the dependent and independent variables, the final dataset for analysis includes 29,821 recipes.

In addition to their main effect, we explore the non-linear effects of the mean ingredient fit and ingredient and verb frequency. Because users may consult the website more frequently for some subcategories or evaluate some subcategories more favorably, we include recipe subcategory fixed effects. Recipes posted earlier tend to have a higher number of trials. Since the data on when the recipe was posted are not directly available, we use the days since the first review as a proxy for the recipe age. Recipe complexity may also affect trials, ratings, and reviews, as users may be inclined to make simpler, quicker, or more elaborate recipes. Therefore, we control for the number of ingredients, preparation time, total time, and number of words in the preparation instructions. Kornish and Jones (2021) show that it is important to control for the number of concepts (ingredients) in the idea when assessing idea's novelty. Moreover, we control for the number of followers of the recipe creator, and the number of personal recipes the recipe creator has posted on Allrecipes.com, as more connected and more experienced creators may receive more trials and favorable ratings and reviews.

As can be seen in Table 5, we find that for both trial and evaluation (proportion of 5-star ratings), consumers prefer recipes in which the ingredients fit well together. This finding is

consistent with the typical view in the food industry that one of the most important factors in a successful recipe is how well the ingredients fit together (Page 2017).

Table 5: The Relationship between the Ingredient-Based Recipe Fit and Novelty Measures and Recipe Success Measures

		Recipe popularity		Recipe favorability	
		log number of trials		percentage of 5 star ratings	
		coef	std err	coef	std err
Ingredient fit measures	ing fit mean (mean centered)	**4.1864**	**0.213**	**0.0844**	**0.035**
	ing fit mean (mean centered) squared	-3.8790	2.085	0.3784	0.336
	ing fit std dev	**-5.4165**	**0.311**	-0.0183	0.050
Ingredient frequency measures	ing freq mean (mean centered)	**3.4502**	**0.193**	**-0.1033**	**0.031**
	ing freq mean (mean centered) squared	**-6.3496**	**0.720**	0.0932	0.116
	ing freq std dev	**-2.3499**	**0.243**	-0.0292	0.039
Verb frequency measures	verb freq mean (mean centered)	**0.5009**	**0.097**	**-0.1196**	**0.016**
	verb freq mean (mean centered) squared	-0.3416	0.483	**0.3837**	**0.078**
	verb freq std dev	**0.5580**	**0.186**	**0.0652**	**0.030**
Recipe complexity variables	number of ingredients (mean centered)	0.0049	0.003	**0.0032**	**0.001**
	number of ingredients (mean centered) squared	-0.0006	0.000	**0.0001**	**7.34E-05**
	prep time (in100 min)	-0.0016	0.004	0.0005	0.001
	total time (in100 min)	0.0003	0.001	0.0002	0.000
	word count in instruction (in 100 cases)	0.0197	0.018	**0.0189**	**0.003**
Recipe creator variables	creator's follower count (in 100,000 cases)	-0.0085	0.005	**0.0034**	**0.001**
	creator's personal recipe count (in 100 cases)	**0.0956**	**0.005**	**0.0046**	**0.001**
Other control variables	recipe age (in 100 days)	**0.0356**	**0.000**	**-0.0023**	**6.59E-05**
	percentage of 5 star ratings	**2.3815**	**0.033**		
	log number of trials			**0.0619**	**0.001**
Subcategory fixed effects		Yes		Yes	
Number of observations		29,821		29,821	
R^2		0.420		0.194	
Adjusted R^2		0.418		0.190	

Note: Bold font for p-value ≤ 0.05. For brevity, we do not report the estimates for subcategory fixed effects in this table.

However, when it comes to the ingredients themselves and the method of preparation, we find a differential effect between trial and favorability evaluation given trial. On the one hand, consumers are more likely to try recipes with common or typical ingredients and preparation

methods, though the marginal impact on trial is diminishing for extremely common ingredients. On the other hand, when it comes to favorability given trial, consumers prefer recipes with more novel ingredients and either extremely novel or extremely common preparation methods. Note that these effects hold after controlling for the complexity of the recipe as measured by the number of ingredients and preparation steps.

Thus, we conclude that when it comes to food recipes the effect of novelty may be more nuanced than commonly found in the ideation literature. First, novelty may not be preferable when it comes to recipe popularity. However, when we look at favorability given trial, which is a more common measure in the ideation literature, consumers prefer recipes that have unique ingredients. Furthermore, distinguishing between novelty of the ingredients themselves and the combination of ingredients, we find that conventional rather than novel ingredient *combinations* (high ingredient fit) are preferred for both trial and evaluation, and novel ingredients are preferred for favorability evaluation given trial, but not for trial. In other words, balanced creativity (i.e., novelty of the components and coherence of the whole) is positively related with favorability given trial. To the best of our knowledge, this is the first paper to separate these two important behavioral outcomes of novelty, namely, trial and evaluation given trial, which is likely to be the antecedent for repeat consumption. As we show, separating these two outcome measures is meaningful as one is positively, and the other is negatively related to novelty.

It can be argued that recipe popularity is driven by ingredient availability rather than ingredient novelty. That is, the reason for observing higher trial for frequently used ingredient is that consumers are more likely to have these ingredients readily available at home. To address this concern, we leverage the fact that our ingredient novelty measure captures the frequency of

appearance of an ingredient within a subcategory rather than across all recipes in Allrecipes.com. Thus, availability should be more related to frequency of appearance of the ingredient across subcategories rather than within a subcategory. For example, "salt" may be a common and widely available ingredient in general; however, it is a novel ingredient for Drink recipes. On the other hand, an uncommon and less accessible ingredient like "Jägermeister" may be used often in Shot recipes. Thus, we re-ran the analysis in Table 5 controlling for the overall ingredient frequency across categories as a proxy for ingredient availability in addition to ingredient frequency within each subcategory as a measure of novelty (See Appendix B). We find that while our proxy for availability—ingredient novelty across categories—is a significant factor, ingredient novelty within subcategory is still a significant driver of recipe popularity, suggesting that the negative effect of ingredient novelty on trial is not fully driven by availability.

Predictive Ability

To assess predictive ability, we explore how well recipe ingredients and their combination help to predict recipe success in terms of the recipe popularity and favorability. To do so, we conduct 10-fold cross-validation across 10 random shufflings of the recipes in the dataset. In each fold, we estimate the model coefficients using 90% of the data and leave the remaining 10% of the data to create a clean out-of-sample validation.

We compare four nested versions of our model: Model 1 – the full set of predictors including ingredient fit, ingredient frequency, preparation frequency, and controls; Model 2 – same as Model 1 removing the embedding based ingredient fit measures; Model 3 – same as Model 2 but without both ingredient novelty measure (fit and frequency); and Model 4, which remove all text-based variable (ingredients and preparation verbs) from Model 1 and only includes only

control variables which consist of recipe complexity variables, recipe creator variables, other control variables, and subcategory fixed effects (see Table 5 for details).

In Table 6, we report the mean squared errors of the cross-validation analysis. The table shows that adding verb frequency measures and ingredient frequency measures significantly improves predictions for all dependent variables (i.e., improvement from Models 4 to 3 and 3 to 2, respectively). Additionally, ingredient fit measures lead to significant improvements for recipe popularity predictions (i.e., improvement from Model 2 to 1). We conclude that novelty of recipe ingredients (as measured by the extent to which the ingredients in the recipe form a coherent combination and the prevalence of the ingredients across recipes in the same subcategory) can improve predictions of recipe popularity and favorability over and beyond all other readily available information, including the novelty of preparation techniques.

Table 6: Predictive Ability (Mean Square Error) of Nested Model Specifications

	Dependent variable	
Model specification	**Recipe popularity**	**Recipe favorability**
(1) Ingredient fit + Ingredient freq. + Verb freq. + Controls	1.734674	0.044765
(2) Ingredient freq. + Verb freq. + Controls	1.796658	0.044765
(3) Verb freq. + Controls	1.863769	0.044802
(4) Controls	1.874614	0.044955
Improvement		
(4) → (1)	**7.46%**	**0.42%**
(2) → (1)	**3.45%**	**0.00%**
(3) → (2)	**3.60%**	**0.08%**
(4) → (3)	**0.58%**	**0.34%**

Mean squared errors (MSE) are averaged across 10 replications of 10-folds mean.
Bold font for p-value ≤ 0.001.

Developing an Application for Improving Recipes

As we mentioned in the opening quote of this paper, one of the most frequent decisions faced by consumers is determining what food to make for a meal. This decision is made daily, and sometimes even multiple times a day. Now that we have identified the novelty drivers of successful recipes using an embedding-based model and demonstrated that these drivers have predictive power, we aim to use these findings to develop a practical application (hereafter app) to inspire recipe ideation for chefs and cooks (professional or amateur). We aim to assist users in their exploration of new recipe ideas by providing an evaluation of their recipe and a shortlist of suggestions for potential improvements for each recipe.

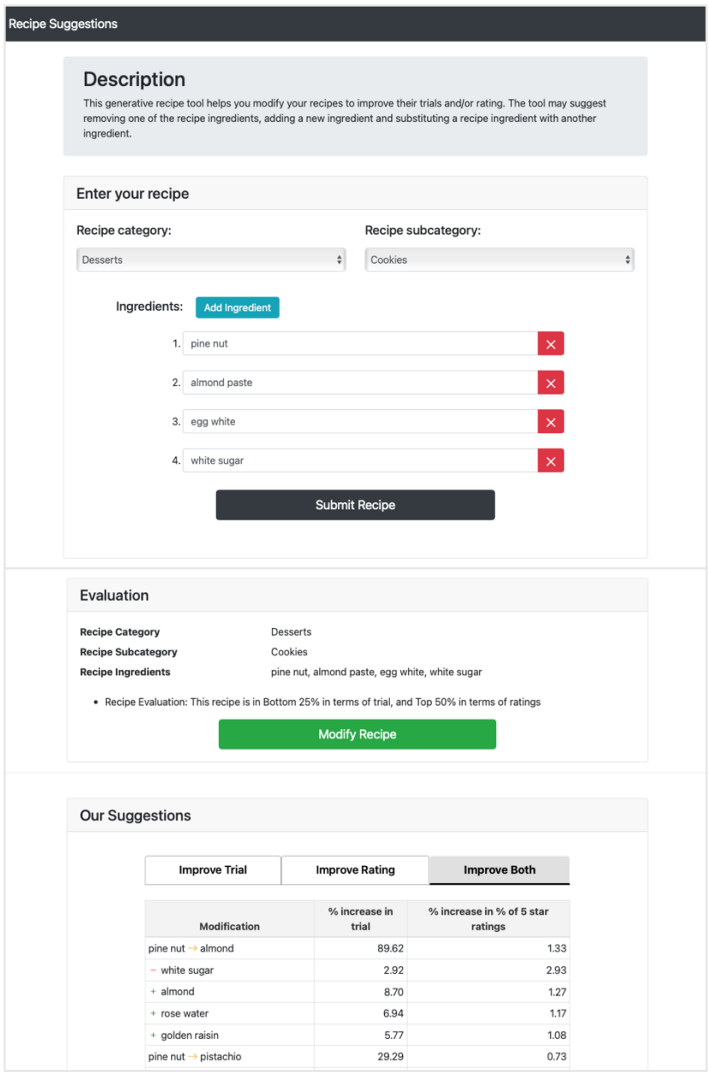
Specifically, we develop a generative recipe tool[7] that leverages our findings and provides specific ingredient suggestions to improve recipe ideas through elaboration (adding ingredients), simplification (removing ingredients), or editing (replacing ingredients). In making recipe ingredient recommendations, we aim to reconcile possibly contradicting objectives (increasing adoption rate vs. increasing liking given trial) to propose better recipe ideas. We will refer to this generative recipe tool as "Gentool" for the remainder of this paper. A screenshot of the app is shown in Figure 2.

When interacting with Gentool, the user first inputs the category and subcategory for the recipe they want to prepare and provides an initial list of ingredients included in this recipe. Then Gentool computes the measures of ingredient fit and ingredient frequency for the given recipe ingredients and subcategory and evaluates the initial recipe's performance based on predicted

[7] Available at recipecreativity.com

recipe popularity and favorability using the model in Table 5 after re-estimating the model coefficients only with the variables that were significantly different from zero. To account for the other non-ingredient related significant variables in the model (e.g., verb frequency, recipe age, recipe creator), we use the average values across the recipes in the same subcategory for each non-ingredient related variable.

Figure 2: A Screenshot of Gentool



Whereas additions of ingredients requires going once through the list ingredient in our corpus, substitutions pose a more difficult problem for the generative algorithm as one would want to substitute an ingredient with a similar ingredient that plays a similar role (e.g., substituting for flour in a cake recipe should use a similar substance that takes the flour role in the recipe). Akin to Wang et al. (2022), who combine word embeddings and hierarchical clustering approaches to generate meta-attributes from engineered attributes, we use a hierarchical clustering procedure to cluster ingredients based on the similarity of their embedding representations. The clustering algorithm procedure suggests very reasonable clusters with a high degree of substitutability at 200 clusters (see Web Appendix B for details). For example, Cluster 2 includes the ingredients walnut, pecan, almond, mixed nut, cashew, pistachio, macadamia nut, pine nut, and hazelnut; and Cluster 4 includes the ingredients parmesan cheese, Romano cheese, feta cheese, goat cheese, gorgonzola cheese, and brie cheese. Once we create the clusters, the substitution operation allows an ingredient to be replaced only with another ingredient in the same cluster.

After evaluating all possible one-at-a-time modifications to the given recipe, Gentool recommends the top 10 ingredient modifications that increase trial, ratings, or both; the modifications can be accessed moving across the app's tabs. A detailed guide on how to use the app is provided in Web Appendix C.

Assessing the App's Performance

In this section, we show the efficacy of Gentool using online panel studies.[8] First, we conduct an online panel study to evaluate the modification rankings of Gentool. Then, we conduct

[8] Human Subjects protocol IRB-********.

another study to compare the quality recipe modifications created by Gentool with those of humans in our panel including those who self-identify as food experts.

Validating the App's Recommendations

To validate the performance of Gentool's recommendations, we conduct an online panel study in which we compare respondents' evaluation of the recipe modifications with the predictions based on the Gentool algorithm. This online panel study provides external validity for our findings because in this study we hold constant possible confounders that may exist in the secondary observational data like the ranking of recipes on the platform or the effect of individual chefs (Blanchard et al. 2022).

Specifically, we recruited 59 participants from Amazon Mechanical Turk (MTurk). Each participant evaluated modification to 5 recipes resulting in 295 observations. The participants were screened for cooking experience (i.e., personally cook the daily meal at home at least 2 days a week) and lack of strict dietary restrictions (i.e., vegan, vegetarian, pescatarian, kosher, halal, raw food diets or gluten, lactose, or egg allergy). We selected five recipes from Allrecipes.com from different categories (breakfasts, main dishes, salads, side dishes, and desserts). For each recipe, we chose 10-12 candidate ingredients to add. The candidate ingredients were selected to generate variation in the recipe's ingredient fit, uniqueness, and predicted change in trial based on our model.

Each participant was asked to rank-order the candidate ingredient additions for each of the recipes in terms of how appealing the recipe would be with this ingredient added to the original recipe. At the end of the survey, participants were asked whether they know the ingredients (1 – yes, 0 – no), whether they had any allergic/dietary restrictions for the ingredients (1 – yes, 0 – no),

and how much they like the ingredients on a 5-point Likert scale (1 = strongly dislike, 5 = strongly like). This allows us to control for familiarity with ingredients, recipe-independent preference for ingredients, and food restrictions beyond those used in the screening criteria.

Because respondents do not get to cook, taste, and evaluate the recipes themselves, their rankings are akin to the trial predictions of Gentool. The responses in our study are rank-ordering of preferences. Hence, we analyze the relationship between the predicted change in trial likelihood due to the addition based on our model (or the predicted rank order of the ingredient addition) and the observed ranking by respondents using an exploded logit model (also known as rank-ordered logit; Beggs, Cardell and Hausman 1981), estimated using the *Apollo* package in R (Hess and Palma 2019). If Gentool's rankings are indeed aligned with respondent preferences, we expect to observe positive coefficients for Gentool's predicted improvement on trial and negative coefficients for Gentool's predicted rank (as the recommendations are ranked in descending order based on their predicted improvement upon trial).

Table 7 presents the exploded logit estimates. We observe that the candidate ingredients with higher predicted rank order based on our model were ranked higher by the participants (b = -0.0124, *p*-value = 0.035). Similarly, the ingredients with higher predicted improvement on trial are ranked higher by respondents (b = 0.0037, *p*-value = 0.001).

This online panel study provides external validity for the app suggestions. It demonstrates that Gentool's predictions are aligned with respondent preferences. It should be noted that Gentool's suggestions are informed by aggregate preferences as opposed to individual preferences. In other words, a specific user with strong like or dislike towards an ingredient may not agree with

Gentool's rankings. However, our analysis accounts for such respondent's idiosyncratic preferences for ingredients using the "Ingredient like" variable[9].

Table 7: Exploded Logit Model Estimates for Rank Ordered Candidate Ingredients

	coef.	std err	coef.	std err
Gen tool predicted rank	**-0.0124**	**0.0059**		
Gen tool predicted improvement on trial			**0.0037**	**0.0012**
Ingredient knowledge	0.0103	0.0599	0.0072	0.0599
Ingredient restriction	0.0255	0.0575	0.0250	0.0575
Ingredient like	**0.2031**	**0.0243**	**0.2029**	**0.0243**
Number of observations	295		295	
R^2	0.0072		0.0077	
Adjusted R^2	0.0064		0.0070	
AIC	10,853		10,847	
BIC	10,867		10,861	

Bold font for p-value ≤ 0.05

With the recipe held constant and only modifying the added ingredients, this online panel study also helps to alleviate possible confounds that may exist in the secondary-data analysis like the presentation of recipes on the website, the website ranking algorithm, or the recipe creator's decision to add certain ingredients to a recipe.

Comparing The App's Recommendations to Human Recommendations

In order to compare the recommendations generated by Gentool to those of human creators, we selected 5 recipes from different subcategories and obtain ingredient-modification ideas from humans. For each recipe, we considered two types of modification: addition and replacement of a selected ingredient in the recipe. We conducted an online panel study on MTurk and ask each

[9] We obtain similar results without controlling for ingredient knowledge, restriction and like.

participant to generate 7 ingredient additions and 3 replacements to make the recipe more appealing to most people. Respondents were instructed to make their suggestion in the order of their preference with most preferred recommendation first. Each participant completed this task for 3 randomly chosen recipes out of the set of 5 recipes. We recruited a total of 297 MTurkers. 97 respondents were discarded because they did not follow the instructions, either providing non-sensical answers for food recommendations (e.g., "five" or "anything") or recommended an ingredient that is already in the recipe. After we eliminated these respondents from our survey responses, we ended up with 200 valid participants. We further removed food recommendations that did match our set of known ingredients (e.g., garnish, dairy). Overall, we obtained 253-323 unique recommendations (additions and replacements) from our participants for each of the five recipes[10]. We generated addition and replacement recommendations for the same recipes using the Gentool for trial improvement.

To evaluate the quality of the recommendations from both the humans and Gentool, we conducted another panel study with a different set of MTurk participants. Before we began the survey, we screened for bots using simple English fluency questionnaire[11]. We asked 208 MTurkers to rate unique additions or replacements proposed by both Gentool and MTurkers on a scale of 0 to 100. Across the two methods and 5 the recipes, there were 1,044 unique ingredient additions and 457 unique ingredient replacements. Each participant evaluated 3 recipes and 30 modifications per recipe. We obtained on average 8.92 ratings for each modification (with a

[10] AsPredicted Pre-Registration #******
[11] Sample question: I saw the horse jump __ the fence. A) over, B) until, C) with.

minimum of 5 ratings for each modification). We averaged the individual ratings to get the quality score for each modification.

In our analysis, we compare the ingredient modification quality of each participant with the modification quality of Gentool's recommendations for each of the recipes observed by participants. We first compare Gentool to each individual human recommender to assess whether Gentool provides better recommendations than most humans. Next, we examine how Gentool performs relative to the wisdom of the crowd of humans by averaging responses across human recommenders.

Table 8: Recommendations of Each MTurk Participant vs. Gentool, Evaluated by Other MTurkers

		Addition	Replacement	All Modifications
Number of MTurk participants		200	200	200
Number of evaluations		594	596	590
Top ideas	**Number of ideas for each recipe**	1	1	2
	Gentool win rate against MTurk	**57.41%**	**55.20%**	**59.66%**
	p*-value	**<0.001**	**<0.005**	**<0.001**
All ideas	**Average number of ideas for each recipe**	**6.69**	**2.86**	**9.54**
	Gentool win rate against MTurk	50.51%	**59.23%**	**54.41%**
	p-value*	0.403	**<0.001**	**0.016**

* One-tailed test with the alternative hypothesis that the win rate is greater than 0.5 (chance).

To compare Gentool to individual human recommenders, first we consider the top ingredient addition and replacement modifications proposed by each MTurk participant and by Gentool. Table 8 shows that Gentool's top ideas for recipe modifications have higher quality than 59.66% of the MTurkers with a success rate significantly surpassing the chance level of 50% (*p*-value < 0.001). Next, we consider all ingredient additions and replacements proposed by each Mturk participants and by Gentool. Table 8 shows that Gentool produces higher-quality

modifications than 54.41% of MTurkers and this success rate is significantly greater than chance (p-value = 0.016).

One concern with MTurk respondents is that they may not be sufficiently familiar with cooking and food recipes. To address this concern, next we focus on modifications generated by self-identified experts. In this survey, we adapt a measurement scale developed by Mitchell and Dacin (1996) and ask a 7-item question to assess domain-specific expertise in the context of cooking, food, and recipes (see Web Appendix D for details). We compute an expertise score by averaging the responses to the 7 items and categorize those above the median as experts[12]. 77 participants out of 200 MTurkers are identified as experts with this method. Table 9 repeats the analysis in Table 8, this time comparing the app performance only with food experts. The results are similar to the previous analysis: Gentool has a higher recommendation quality than most experts.

Table 9: Recommendations of Each *Expert* MTurk Participant vs. Gentool, Evaluated by Other MTurkers

		Addition	Replacement	All Modifications
Number of MTurk participants		77	77	77
Number of evaluations		229	228	228
Top ideas	**Number of ideas for each recipe**	1	1	2
	Gentool win rate against MTurk	**61.57%**	54.39%	**62.28%**
	p*-value	**<0.001**	0.093	**<0.001**
All ideas	**Average number of ideas for each recipe**	**6.74**	**2.87**	**9.61**
	Gentool win rate against MTurk	48.91%	**58.77%**	**56.14%**
	p-value*	0.629	**0.004**	**0.032**

* One-tailed test with alternative hypothesis that the win rate is greater than 0.5 (chance).

[12] There are 77 MTurkers above the median score and 25 MTurkers with exactly the median score.

Finally, we compare the app recommendations with the collective MTurk recommendations. This analysis can be thought of as a wisdom of the crowd analysis (Surowiecki, 2004) because we investigate how the app recommendations compare to the most popular ideas recommended by the collective of MTurkers. We find that the app recommendations have some degree of agreement with the wisdom of the crowd of the MTurkers' recommendation. Of the 50 possible modifications (5 recipes times 7 addition and 3 replacements), 45 are also recommended by at least one MTurker. However, only 17 of the app modifications are among the most frequently proposed ideas *across* MTurk respondents (most popular 7 additions and 3 replacements for each of the 5 recipes). We further compare the idea quality of most common 50 MTurk ideas and the app's 50 recommendations. We find that the mean idea quality of most common MTurk ideas is 50.47 with a standard deviation of 8.22 while the mean idea quality of the app recommendations is 49.64 with a standard deviation of 8.62. We conclude that the overall idea quality is fairy similar between the app and the aggregation of humans. This analysis demonstrates that when leveraging the wisdom of the crowds—collecting the most common modifications across a large number of respondents—human respondents reach the same, or even better, level of quality of recommendations relative to app. However, it should be noted that such crowdsourcing is impractical for most applications, such as Allrecipes.com with tens of thousands of recipes and many possible modifications for each one. Additionally, consumers expect to receive such recommendations in real-time and are not likely to wait for crowdsourcing of such modifications in real-time.

Overall, this analysis suggests that app proposes significantly better ideas than most human creators, including those with domain-specific expertise. Moreover, the app provides great efficiency and scalability while performing these tasks.

Conclusion

In this paper, we investigate the role of novelty in food recipe ingredients and ingredients' combinations on recipe success. Borrowing from the NLP literature, we propose an embedding approach to capture the novelty of the ingredients in a recipe by examining the fit of each ingredient to the context in which it appears (the other ingredients). By doing so, we extend the view of ingredient novelty in ideation products beyond merely the frequency of the ingredients or the pairwise relationships among ingredients to a more holistic view. Considering the holistic fit of all ingredients is particularly important in creatives such as recipes.

Using a multi-pronged approach combining NLP and machine learning tools, econometric models, secondary data from tens of thousands of food recipes, and online panel studies, our work sheds a new light on the impact of creativity on consumer attitudes and preferences. Using our measures of embedding-based ingredient novelty we find that, in the context of food recipes, the effect of creativity may be more nuanced than has been discussed before in the literature. Contrary to existing research that primarily emphasized the positive impact of creativity, we find that in some cases creativity may have a negative effect; recipes with novel ingredients (as measured by the prevalence of these ingredients across recipes in the same category) are likely to have lower trial, but higher favorability evaluation given trial. Additionally, we find that whereas the traditional measure of ingredient prevalence has a differential effect on trial and favorability

evaluation given trial, having a high fit among ingredients in the recipe (more conventional ingredient combinations) is preferable for both trial and favorability evaluations.

Our work in the context of food recipes may guide how chefs and restauranteurs design their menus in order to increase customer adoption (trial) and customer retention, and/or improve their ratings and reviews on online platforms such as Yelp. The food and beverage companies may use these findings to develop new products. This work may also help the rapidly growing meal kit delivery services (e.g., Blue Apron, HelloFresh) which provide their subscribers with weekly recipes and ingredient kits. While they may want to offer recipes with a familiar combination of ingredients for the purpose of customer acquisition (trial), to increase retention they may wish to come up with creative recipes that keep their customers engaged.

We further leverage these insights and findings to construct an app that improves recipes by proposing concrete ingredient modifications. To make the work more accessible to chefs and cooks (professional or amateur) who wish to improve their recipes, as well as to academics interested in creativity in the food domain, we created a publicly available web app (http://recipecreativity.com) of Gentool. In multiple validation exercises, we show the efficacy of this tool in capturing human preference and providing better recommendations than most human creators. Our application can help consumers in one of the most common decisions on a daily basis: what food to prepare. The app can also help consumers who wish to modify a recipe due to dietary restriction, weight lost preferences or simply dislike for a particular ingredient by recommending likely substitutions for specific ingredients.

The rise of generative AI tools in recent years has brought to the public's attention the use of such tools for food and recipe recommendations. Our work contributes to computational

creativity and automated idea-generation literature in general and recipe recommender systems in particular. It provides evidence for the efficacy of automated tools for food recipes by using big data and machine learning methods. Future work can explore emerging deep learning models approaches for food recipes recommendations.

Future research can also further explore the potential of using representation learning in other creativity and idea-generation domains. We believe that several creative domains such as art, fashion, or advertising creatives may particularly benefit from the gestalt view of these ingredient fit measures.

References

Ahn, Yong-Yeol, Sebastian E. Ahnert, James P. Bagrow, and Albert-László Barabási (2011), "Flavor network and the principles of food pairing," *Scientific Reports*, 1(1), 1-7.

Amorim, Álvaro, Luís Fabrício W. Góes, Alysson Ribeiro da Silva, and Celso França (2017), "Creative flavor pairing: using RDC metric to generate and assess ingredients combination," *Proceedings of the International Conference on Computational Creativity, June 2017*, 33-40.

Anderson, Carl (2018), "A survey of food recommenders," *arXiv preprint arXiv:1809.02862*.

Beggs, S., S. Cardell, and Jerry Hausman (1981), "Assessing the potential demand for electric cars," *Journal of Econometrics*, 17(1), 1-19.

Bengio, Y., Réjean Ducharme, Pascal Vincent, and Christian Jauvin (2003), "A neural probabilistic language model," *Journal of Machine Learning Research*, 3 (Feb), 1137–1155.

Blanchard, Simon J., Jacob Goldenberg, Koen Pauwels, and David A Schweidel (2022), "Promoting data richness in consumer research: How to develop and evaluate articles with multiple data sources," *Journal of Consumer Research*, 49(2), 359-372.

Buiano, Madeline (2022), "Where Do You Find Recipe Inspiration? A New Survey Says Most Home Cooks Get Ideas from Loved Ones or the Internet," *Martha Stewart* (March 21), https://www.marthastewart.com/8242003/where-find-recipe-inspiration-survey-2022/

Burroughs, James E., C. Page Moreau, and David Glen Mick (2008), "Toward a psychology of consumer creativity," in *Handbook of Consumer Psychology,* Curtis P. Haugtvedt, Paul M. Herr, Frank R. Kardes, eds. Lawrence Erlbaum Associates, 1011–1038.

Cromwell, Erol, Jonah Galeota-Sprung, and Raghuram Ramanujan (2015), "Computational creativity in the culinary arts," *Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference, April 2015*, 38–42.

Dahl, Darren W., Amitava Chattopadhyay and Gerald J. Gorn (1999), "The use of visual mental imagery in new product design," *Journal of Marketing Research*, 36(1), 18-28.

Duchi, John, Elad Hazan, and Yoram Singer (2011), "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, 12, 2121–2159.

Finke, Ronald A., Thomas B. Ward, and Steven M. Smith (1996), *Creative cognition: Theory, research, and applications*. MIT Press.

Firth, John Rupert (1957), "A synopsis of linguistic theory 1930-1955," *Studies in Linguistic Analysis*, 1–32.

Freyne, Jill and Shlomo Berkovsky (2010), "Recommending food: reasoning on recipes and ingredients," *International Conference on User Modeling, Adaptation, and Personalization UMAP* 2010, 381–386.

Giora, Rachel (2003), *On our mind: Salience, context and figurative language.* Oxford University Press.

Girotra, Karan, Christian Terwiesch, and Karl T. Ulrich (2010), "Idea generation and the quality of the best idea," *Management Science* 56(4), 591–605.

Harris, Zellig S. (1954), "Distributional structure," *Word*, 10(23), 146-162.

Hess, Stephane, and David Palma (2019), "Apollo: a flexible, powerful and customisable freeware package for choice model estimation and application," *Journal of Choice Modelling,* 32, 100170.

Honnibal, Matthew, and Ines Montani (2017), "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing."

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy (2021), "Measuring technological innovation over the long run," *American Economic Review: Insights* 3(3), 303-320.

Kornish, Laura J., and Sharaya M. Jones (2021), "Raw ideas in the fuzzy front end: Verbosity increases perceived creativity," *Marketing Science*, 40(6), 1106-1122.

Kornish, Laura J., and Karl T. Ulrich (2011), "Opportunity spaces in innovation: Empirical analysis of large samples of ideas," *Management Science*, 57 (1), 107–128.

Kornish, Laura J., and Karl T. Ulrich (2014), "The importance of the raw idea in innovation: Testing the sow's ear hypothesis," *Journal of Marketing Research*, 51(1), 14–26.

Jennie-O Turkey Store (2022). "Jennie-O Shares National Survey Results Revealing How Americans Are Combatting Recipe Boredom & Cooking Fatigue," *PR Newswire: Press Release Distribution, Targeting, Monitoring and Marketing* (March 15), www.prnewswire.com/news-releases/jennie-o-shares-national-survey-results-revealing-how-americans-are-combatting-recipe-boredom--cooking-fatigue-301502934.html.

Levy, Omer, and Yoav Goldberg (2014), "Neural word embedding as implicit matrix factorization," *Neural Information Processing Systems*, 2177–2185.

Luo, Lan, and Olivier Toubia (2015), "Improving online idea generation platforms and customizing the task structure based on consumers' domain specific knowledge," *Journal of Marketing*, 79(5), 100-114.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013a), "Efficient estimation of word representations in vector space," *ICLR Workshop Proceedings*. arXiv:1301.3781.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean (2013b), "Distributed representations of words and phrases and their compositionality," *Advances In Neural Information Processing Systems*, 26.

Mitchell, Andrew A., and Peter A. Dacin (1996), "The assessment of alternative measures of consumer expertise," *Journal of Consumer Research*, 23(3), 219-39.

Moreau, C. Page, and Darren W. Dahl (2005), "Designing the solution: The impact of constraints on consumers' creativity," *Journal of Consumer Research*, 32 (1), 13–22.

Morris, Richard G., Scott H. Burton, Paul M. Bodily, and Dan Ventura (2012), "Soup over bean of pure joy: Culinary ruminations of an artificial chef," *Proceedings of the International Conference on Computational Creativity*, *May 2012*, 119–125.

Packard, Grant, Anocha Aribarg, Jehoshua Eliashberg, and Natasha Z. Foutz (2016), "The role of network embeddedness in film success," *International Journal of Research in Marketing*, 33(2), 328-342.

Page, Karen (2017), *Kitchen creativity*. Little, Brown and Company.

Pennington, Jeffrey, Richard Socher, Christopher D. Manning (2014), "Glove: Global vectors for word representation," *Proceedings of the 2014 Conference on Empirical Methods on Natural Language Processing, October 2014*, 1532-1543.

Pinel, Florian, Lav R. Varshney and Debarun Bhattacharjya (2015), "A culinary computational creativity system," in *Computational creativity research: Towards creative machines,* Tarek R. Besold, Marco Schorlemmer, Alan Smaill, eds. Springer, 327–346.

Rudolph, Maja, Francisco J. R. Ruiz, Stephan Mandt, and David M. Blei (2016), "Exponential family embeddings," *Advances in Neural Information Processing Systems*, 29, 478–486.

Ruiz, Francisco J., Susan Athey, and David M. Blei (2020), "Shopper: A probabilistic model of consumer choice with substitutes and complements," *Annals of Applied Statistics*, 14(1), 1–27.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (1986), "Learning representations by back-propagating errors," *Nature*, 323(6088), 533–536.

Schoenmueller, Verena, Oded Netzer, and Florian Stahl (2020), "The polarity of online reviews: Prevalence, drivers and implications," *Journal of Marketing Research*, 57(5), 853-877.

Singh, Jasjit, and Lee Fleming (2010), "Lone inventors as sources of breakthroughs: Myth or reality?," *Management Science*, 56(1), 41-56.

Surowiecki, James (2004), *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. Doubleday & Co.

Teng, Chun-Yuen, Yu-Ru Lin, and Lada A. Adamic (2012), "Recipe recommendation using ingredient networks," In *Proceedings of the 4th annual ACM web science conference, June 2012*, 298-307.

Toubia, Olivier, and Oded Netzer (2017), "Idea generation, creativity, and prototypicality," *Marketing Science*, 36(1), 1–20.

Trang Tran, Thi Ngoc, Müslüm Atas, Alexander Felfernig and Martin Stettinger (2018), "An overview of recommender systems in the healthy food domain," *Journal of Intelligent Information Systems*, 50, 501–526.

Uzzi, Brian, Satyam Mukherjee, Michael Stringer, and Ben Jones (2013), "Atypical combinations and scientific impact," *Science*, 342(6157), 468-472.

Van der Maaten, Laurens, and Geoffrey Hinton (2008), "Visualizing data using t-SNE," *Journal of Machine Learning Research*, 9(11), 2579-2605.

Varshney, Lav R., Florian Pinel, Kush R. Varshney, Debarun Bhattacharjya, Angela Schoergendorfer, and Yi-Min Chee (2019). A big data approach to computational creativity: The curious case of Chef Watson. *IBM Journal of Research and Development*, 63(1), 7:1-7:18.

Varshney, Lav R., Jun Wang, and Kush R. Varshney (2016), "Associative algorithms for computational creativity,". *The Journal of Creative Behavior*, 50(3), 211-223.

Wang, Xin (Shane), Jiaxiu He, David J. Curry, and Jun Hyun (Joseph) Ryoo (2022), "Attribute embedding: Learning hierarchical representations of product attributes from consumer reviews," *Journal of Marketing*, 86(6), 155-175.

Wansink, Brian, and Jeffery Sobal (2007), "Mindless eating," *Environment and Behavior,* 39(1), 3-142.

Ward, Thomas B. (1995), "What's old about new ideas?," in *The Creative Cognition Approach,* Steven M. Smith, Thomas B. Ward and Ronald A. Finke, eds. MIT Press, 157-178.

Ward, Thomas B. (2001), “Creative cognition, conceptual combination, and the creative writing of Stephen R. Donaldson,” *American Psychologist*, 56(4), 350–54.

Wei, Yanhao 'Max', Jihoon Hong, and Gerard J. Tellis (2022), “Machine learning for creativity: Using similarity networks to design better crowdfunding projects,” *Journal of Marketing*, 86(2), 87-104.

Appendix A: Recipe Category Summary Statistics

Categories	Subcategories	Recipe count	Number of ingredients				
			mean	stdev	min	median	max
Appetizers and Snacks	Antipasto, Beans and Peas, Bruschetta, Canapes and Crostini, Cheese, Deviled Eggs, Dips and Spreads, Fruit, Garlic Bread, Meat and Poultry, Nuts and Seeds, Pasta Appetizers, Pastries, Pickled Eggs, Seafood, Snacks, Spicy, Tapas, Vegetable, Wraps and Rolls	5363	7.78	3.32	2	7	42
Bread	Bread Machine, Holiday Bread, Pastries, Pizza Dough and Crusts, Quick Bread, Yeast Bread	3168	9.78	3.19	2	10	24
Breakfast and Brunch	Breakfast Bread, Cereals, Crepes, Drinks, Eggs, French Toast, Meat and Seafood, Pancakes, Potatoes, Waffles	3365	8.69	3.04	2	8	23
Desserts	Cakes, Candy, Chocolate, Cobbler, Cookies, Crisps and Crumbles, Custards and Puddings, Fillings, Frostings and Icings, Frozen Desserts, Fruit Desserts, Mousse, Nut Desserts, Pies, Specialty Desserts	12523	8.47	3.13	2	8	29
Drinks	Beer, Cider, Cocktails, Coffee, Eggnog, Hot Chocolate, Juice, Lemonade, Liqueurs, Mocktails, Mulled Wine, Punch, Sangria, Shakes and Floats, Shots, Slushies, Smoothies, Tea	2519	5.33	2.06	2	5	17
Everyday Cooking	More Meal Ideas, Vegan, Vegetarian	1159	8.71	3.94	2	8	27
Fruits and Vegetables	Vegetables	1054	10.41	3.28	2	10	22
Main Dish	Beef, Bowls, Burgers, Casseroles, Chicken, Curries, Dumplings, Meatballs, Meatloaf, Pasta, Pizza, Pork, Ribs, Rice, Roasts, Sandwiches, Savory Pies, Seafood, Steaks and Chops, Stir-Fry, Stuffed Main Dishes, Tacos, Vegetable Main Dishes	5426	9.55	3.59	2	9	29
Meat and Poultry	Beef, Chicken, Game Meats, Lamb, Pork, Turkey	4790	9.28	3.62	2	9	32
Pasta and Noodles	Noodles, Pasta by Shape	281	9.49	3.57	2	9	22
Salad	Beans, Beef and Pork Salads, Coleslaw, Curry Salad, Egg Salad, Fruit Salads, Grains, Green Salads, Pasta Salad, Potato Salad, Seafood Salad, Taco Salad, Vegetable Salads, Waldorf Salad	3769	9.55	3.19	3	9	27
Seafood	Fish, Shellfish	1204	9.27	3.22	2	9	24
Side Dish	Applesauce, Beans and Peas, Casseroles, Curry Side Dishes, Fries, Grains, Hushpuppies, Potatoes, Rice, Sauces and Condiments, Stuffing and Dressing, Vegetables	8240	7.83	3.16	2	7	26
Soups, Stews and Chili	Bisque, Broth and Stocks, Chili, Chowders, Soup, Stews	4848	11.91	3.84	2	12	28

Appendix B: Ingredient Availability and Recipe Popularity

We define ingredient availability as the overall ingredient frequency across categories and use the mean of the overall frequency of recipe ingredients for each recipe. Table B1 shows the relationship between the ingredient-based recipe fit and novelty measures and recipe popularity when we include the ingredient availability measure. While ingredient availability has a statistically significant relationship with trial (b = 0.7528), ingredient novelty measured by the mean frequency of ingredients *within the recipe subcategory* is still statistically significant and positively related to recipe popularity (b = 3.0730).

Table B1: The Relationship between the Ingredient-Based Recipe Fit and Novelty Measures, Ingredient Availability, and Recipe Popularity

		Recipe popularity	
		log number of trials	
		coef	std err
Ingredient fit measures	ing fit mean (mean centered)	**4.1369**	**0.214**
	ing fit mean (mean centered) squared	-4.0363	2.085
	ing fit std dev	**-5.6070**	**0.319**
Ingredient frequency measures	ing freq mean (mean centered)	**3.0730**	**0.240**
	ing freq mean (mean centered) squared	**-5.9454**	**0.736**
	ing freq std dev	**-2.2765**	**0.245**
Verb frequency measures	verb freq mean (mean centered)	**0.5255**	**0.098**
	verb freq mean (mean centered) squared	-0.3569	0.483
	verb freq std dev	**0.5547**	**0.186**
Recipe complexity variables	number of ingredients (mean centered)	0.0041	0.003
	number of ingredients (mean centered) squared	-0.0005	0.000
	prep time (in 100 min)	-0.0016	0.004
	total time (in 100 min)	0.0003	0.001
	number of words in instruction (in 100 cases)	0.0172	0.018
Recipe creator variables	creator's follower count (in 100,000 cases)	-0.0082	0.005
	creator's personal recipe count (in 100 cases)	**0.0955**	**0.005**
Other control variables	recipe age (in 100 days)	**0.0356**	**0.000**
	percentage of 5-star ratings	**2.3830**	**0.033**
	overall ing freq mean (mean centered)	**0.7528**	**0.284**
	Subcategory fixed effects	Yes	
	Number of observations	29,821	
	R2	0.420	
	Adjusted R2	0.418	

Bold font for p-value ≤ 0.05. For brevity, we do not report the estimates for subcategory fixed effects in this table.

Web Appendix A: Extracting Recipe Ingredients

Allrecipes.com includes 5,966 unique ingredient IDs. There are at least three issues with directly using these ingredient IDs and descriptions for this analysis. First, the text descriptions for the same ingredient ID may be slightly different in different recipes. Second, ingredients with different ingredient IDs often have similar text descriptions. Third, the list of ingredient IDs has a long tail, where only about 35% of ingredient IDs have been used in at least 10 recipes and excluding recipes with ingredients that appear in less than 10 recipes leads to removing 8.2K recipes from the dataset (i.e., over 14% of the data). Therefore, we identify the most representative label for each ingredient ID using text analysis and group them based on the similarity of their labels.

Step 1: Clean text descriptions.

First, we use text analysis to extract candidate labels for each ingredient ID. We start by removing quantities and measurements (e.g., 2 cups, 3 tablespoons, 16 pounds) as well as brand names (e.g., Heinz, Barilla) from the ingredient text descriptions (Table W1).

Table W1: Text Data Cleaning for Ingredients

Ingredient ID	Recipe ID	Ingredient text description	Cleaned and lemmatized text
10498	16678	2 (14.5 ounce) cans peeled and diced tomatoes	peeled and diced tomato
	222002	1 (10 ounce) can diced tomatoes	diced tomato
	165190	1 (14.5 ounce) can diced tomatoes	diced tomato
	78052	1 (14.5 ounce) can diced tomatoes with juice	diced tomato with juice
	87624	1 (14.5 ounce) can diced tomatoes, drained	diced tomato drained
	…	…	…

Step 2: Extract candidate labels.

We create a list of unique n-grams from cleaned and lemmatized ingredient text descriptions as candidate labels. For example, the candidate labels extracted from "peeled and diced tomato" include "peeled", "and", "diced", "tomato", "peeled and", "and diced", "diced tomato", "peeled and diced", "and diced tomato" and "peeled and diced tomato."

Step 3: Extract the most representative label.

To choose the most representative label for each ingredient ID, we score and rank candidate labels based on their representativeness. We define coverage and informativeness scores to measure the representativeness of a candidate label.

The coverage score is defined as the percentage of recipes with ingredient text description that contain the candidate label among recipes that include the ingredient ID. For example, in the case of ingredient ID 10498 which is used in 1,075 recipes, the candidate label "tomato" appears in 100% of these recipes and "diced tomato" in 97.3% of these recipes.

$$\text{coverage} = \frac{\text{number of recipes with candidate label for ingredient ID}}{\text{number of recipes for ingredient ID}}$$

The informativeness score is defined as the discrepancy between the candidate label length and the median text description length; it is the ratio (inverse ratio) of the candidate label length and the median text description length if the candidate label length is smaller (larger) than the median text description length. For the ingredient ID 10498, the median text description length is 2 and the informativeness scores are 1/2 for "tomato" and 2/2 = 1 for "diced tomato."

$$\text{informativeness} = \begin{cases} \dfrac{\text{candidate label length}}{\text{median text length}} & \text{, if median text length} \geq \text{candidate label length} \\ \dfrac{\text{median text length}}{\text{candidate label length}} & \text{, if median text length} < \text{candidate label length} \end{cases}$$

The most representative label for each ingredient ID is chosen based on the highest coverage × informativeness score. "Diced tomato" is the most representative label for ingredient ID 10498 (Table W2).

Table W2: Calculating Representativeness Score for Ingredient ID 10498 that Appears in 1,075 Recipes with Median Text Description Length of 2

Candidate label	Number of recipes with candidate label	Coverage	Text length	Informativeness	Representativeness score
diced tomato	**1046**	**0.973**	**2**	**1.000**	**0.973**
tomato	1075	1.000	1	0.500	0.500
tomato drained	131	0.122	2	1.000	0.122
diced tomato drained	130	0.121	3	0.667	0.089
…	…	…	…	…	…

Steps 4&5&6: Group ingredient IDs with similar labels.

Different ingredient IDs are grouped together first based on exact match of their label, then based on similarity of their representative labels after removing words such as "diced" and "crushed" that refer to mechanical form, and finally, rare ingredients are manually added to groups that are the most similar (Table W3).

In step 4, we identify ingredient IDs with identical most representative labels and group them together as one ingredient. Thus, 1,095 ingredient IDs match with at least one other ingredient ID and are classified under 495 unique labels.

In step 5, we relax the matching criteria and group the ingredient IDs based on the similarity of their most representative label. For this step, we ignore differentiating words that are related to the mechanical form of the ingredient (e.g., crushed, diced, chopped, drained, peeled, seeded, large, small, whole, half etc.) in the most representative labels and group the ingredient IDs if the remaining substrings in the most representative labels match. Note that "diced tomato" and "crushed tomato" refer to different mechanical forms of the same ingredient ("tomato"), whereas "cherry tomato", "green tomato", "sun dried tomato", "tomato juice" and "tomato paste" refer to different types of ingredients, which are not grouped together. We use the common substring in all most representative labels of the ingredient IDs in the same group as the group label.

In step 6, we manually go through the remaining ingredient IDs and group ingredients that appear in less than 10 recipes. At this step, we try to match them to one of the existing ingredient IDs or ingredient groups based on the similarity of the candidate labels (e.g., "Italian style tomato sauce", "Mexican style tomato sauce", "no salt added tomato sauce", "spicy tomato sauce", etc., are added to the "tomato sauce" group). Or we group them with other similar and rare ingredient IDs if they share a common candidate label with other rare ingredient IDs, but they are not similar enough to one of the existing ingredient IDs or ingredient groups (e.g., "goat's milk", "hemp milk", "oat milk" and "flax milk" are grouped together as "other milk").

The remaining unmatched ingredients that appear in less than 10 recipes are removed from the data. As a result, we remove from the dataset 279 ingredient IDs that can't be matched with

other ingredient IDs and appear in less than 10 recipes as well as 666 recipes that include them. After completing these steps, we reduce the number of unique ingredients to 1,249, each of which appears in at least 10 recipes, and have a total list of 57,709 recipes that are constructed from these ingredients.

Table W3: Grouping Multiple Ingredient IDs as a Single Ingredient

Ingredient ID	The most representative label	Group label (Step 4: identical match)	Group label (Step 5: similar match)	Final group label (Step 6: manual matching)
4572	tomato	tomato	tomato	tomato
21082	tomato	tomato	tomato	tomato
4574	whole peeled tomato		tomato	tomato
10494	canned crushed tomato		tomato	tomato
12327	canned whole tomato chopped		tomato	tomato
10214	diced tomato	diced tomato	tomato	tomato
10498	**diced tomato**	**diced tomato**	**tomato**	**tomato**
29960	diced tomato	diced tomato	tomato	tomato
23073	diced tomato	diced tomato	tomato	tomato
4664	crushed tomato	crushed tomato	tomato	tomato
27259	crushed tomato	crushed tomato	tomato	tomato
10502	italian style diced tomato	italian style diced tomato	italian style tomato	tomato
28084	italian style diced tomato	italian style diced tomato	italian style tomato	tomato
12333	italian style tomato		italian style tomato	tomato
3638	no salt added tomato	no salt added tomato	no salt added tomato	tomato
10499	no salt added tomato	no salt added tomato	no salt added tomato	tomato
27954	no salt added diced tomato		no salt added tomato	tomato
21589	pear tomato			tomato
4666	yellow tomato			tomato
25936	diced chili style tomato			tomato

Web Appendix B: Ingredients Clustering

We use Hierarchical Clustering to cluster ingredients, where each ingredient is represented by its context vector. Specifically, we use Agglomerative Clustering, which recursively merges pairs of ingredients or clusters that minimally increase the linkage distance, using the *scikit-learn* package in Python. For the linkage criterion, we use Ward's method as it minimizes the variance of the clusters being merged and provides coherent clusters.

Figure W1: The Linkage Distance Threshold and Number of Clusters

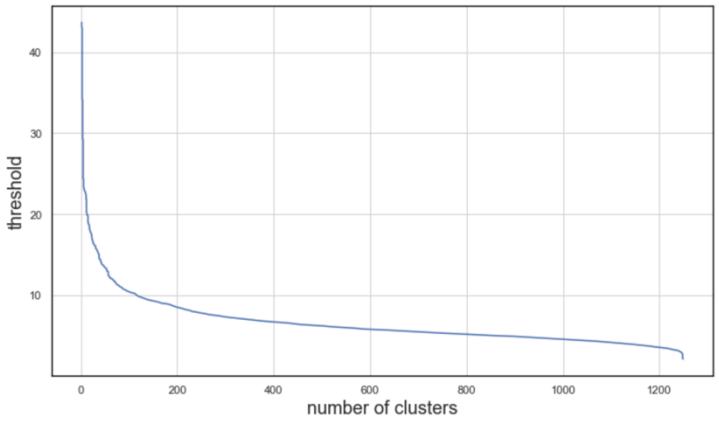


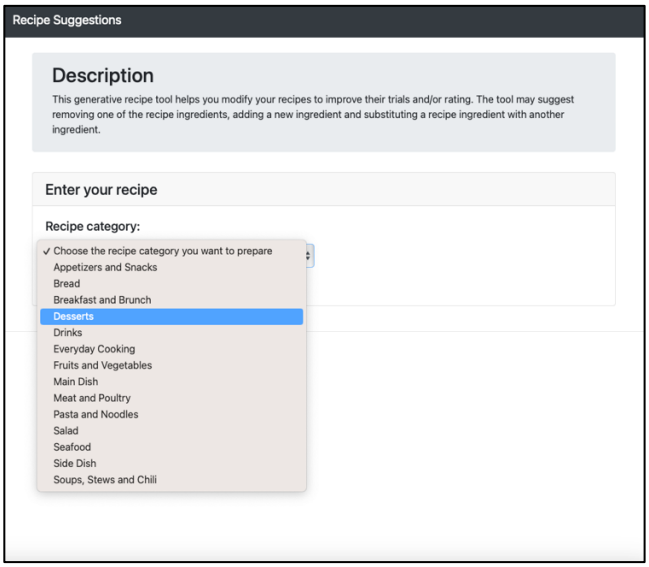
Figure W1 shows how the linkage distance threshold changes at different numbers of clusters. We decided to use 200 clusters, as after this point the decrease in the linkage distance with additional clusters is less pronounced. Table W4 shows the first 20 clusters and the ingredients that belong to these clusters.

Table W4: The First 20 Clusters and Ingredients

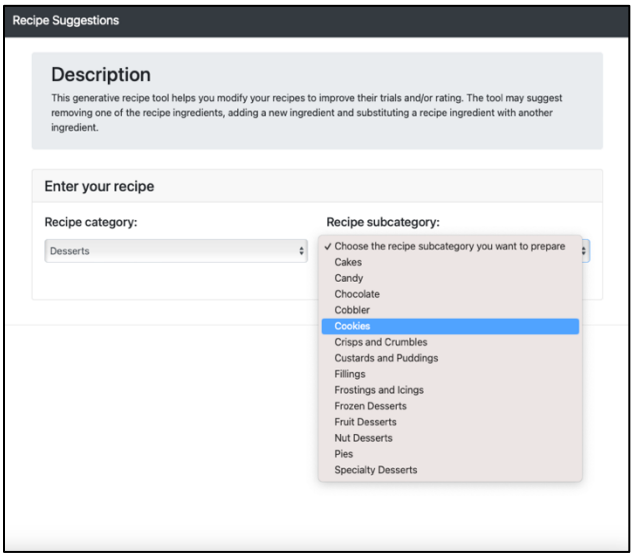
Cluster no	Ingredients
1	flavored gelatin, food coloring, limeade, lemonade, drink mix, cinnamon candy, sherbet, orange drink mix, unflavored gelatin
2	walnut, pecan, almond, mixed nut, cashew, pistachio, macadamia nut, pine nut, hazelnut
3	banana, pineapple, coconut
4	parmesan cheese, romano cheese, feta cheese, goat cheese, gorgonzola cheese, brie cheese
5	poultry seasoning, mixed vegetable, turkey stock, turkey gravy, turkey dripping, turkey carcass, turkey giblet, turkey broth
6	shortening, oil for frying, lard, bacon dripping
7	provolone cheese, mozzarella cheese, muenster cheese, monterey jack cheese, asiago cheese, gruyere cheese, swiss cheese, gouda cheese, italian cheese blend, hollandaise sauce, fontina cheese, havarti cheese, jarlsberg cheese, string cheese
8	ham, smoked sausage, corned beef, bone in ham, smoked ham, hot dog, kielbasa sausage, bratwurst, smoked turkey, ham steak
9	lamb chop, rack of lamb, leg of lamb, lamb shoulder, ground lamb, beef shank, cranberry bean, lamb stew meat, rabbit, beef bone, oxtail, scotch bonnet chile pepper, lamb shank, chicken carcass
10	bean, molasses
11	dried thyme, dried savory, dried sage, dried rosemary, dried basil, dried oregano, dried parsley, dried marjoram, dried tarragon, dried chive, dried cilantro
12	hominy, tomatillo, chile pepper, serrano pepper, adobo seasoning, anaheim chile pepper, ancho chile pepper, green pea, cotija cheese, chayote squash, sazon seasoning, herb, cassava, chile de arbol pepper, pasilla chile pepper, sofrito, mexican crema, guajillo chile pepper, hatch chile pepper, queso fresco, manchego cheese, cactus, spanish onion
13	coffee, coffee liqueur, irish cream liqueur, espresso, creme de menthe liqueur, liqueur, hazelnut liqueur, schnapps, creme de cacao liqueur
14	beef dripping, port wine, orecchiette pasta, pearl onion, beef liver, emmentaler cheese, animal fat, marjoram, veal, porcini mushroom, merlot wine, ground veal, pecorino cheese, walleye, cheese curd, beef gravy, morel mushroom, pheasant, fiddlehead fern, duck fat, dandelion green, chanterelle mushroom, zucchini blossom, chestnut
15	kiwi, citrus soda, caramel topping, vanilla ice cream, strawberry topping, ice cream, hot chocolate mix, irish whiskey, chocolate flavored syrup, strawberry flavored yogurt, custard powder, chocolate drink mix, snow, strawberry ice cream, lemon curd, chocolate, chocolate ice cream, flavored syrup, vanilla powder, raspberry flavored syrup, pudding mix, chocolate milk, topping, vanilla flavored syrup, jagermeister liqueur, blueberry jam
16	sweetener, superfine sugar, agave nectar, sugar, raw sugar, sucralose sweetener, stevia, cane sugar, coconut sugar
17	milk, half and half, evaporated milk, whole milk, light cream
18	turkey, skinless boneless chicken breast, chicken, chicken chunk, turkey breast, chicken breast, chicken tender, meatball, ground chicken
19	golden raisin, raisin, date, dried currant, dried mixed fruit, dried cherry, dried apricot, dried fig, prune, dried blueberry
20	ground cinnamon, ground nutmeg, ground ginger, ground clove, ground allspice

Web Appendix C: User Manual for Web App Interface[13]

Step 1: The user chooses the category of the recipe they want to prepare from the dropdown menu.

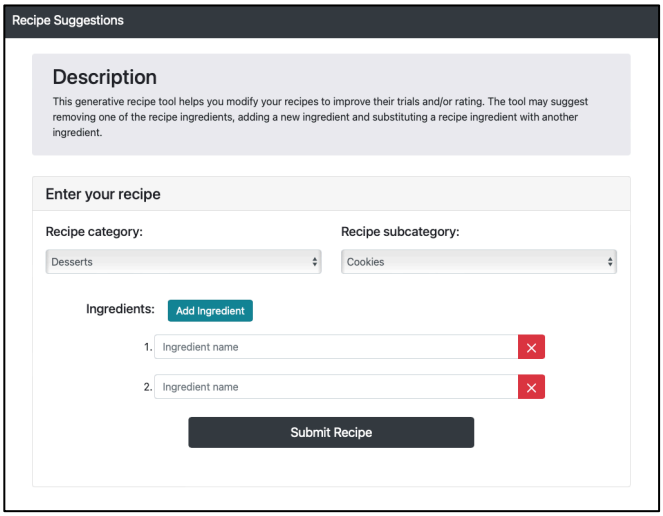


Step 2: The user chooses the subcategory of the recipe they want to prepare from the dropdown menu. The list of recipe subcategories depends on the category selection in Step 1.[14]



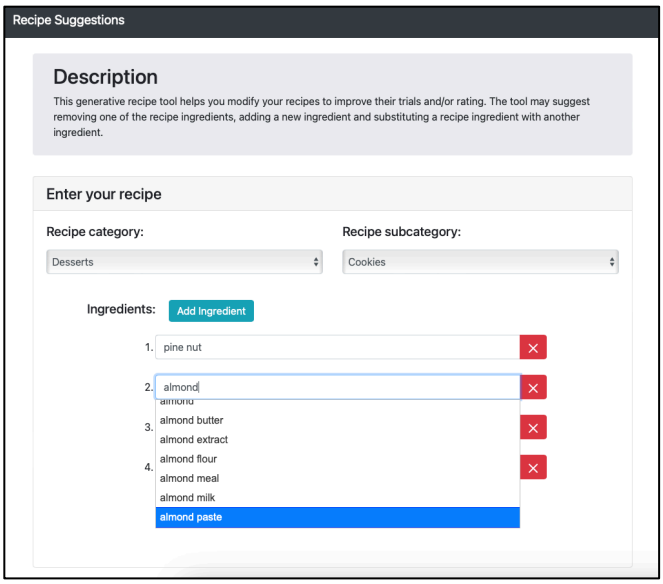
[13] The app is available at http://recipecreativity.com
[14] The list of categories and subcategories matches the list of categories in Allrecipes.com.

Step 3: The ingredient prompt appears on the screen. The user can add additional lines by clicking “Add Ingredient” or delete any line by clicking the “X” next to it.

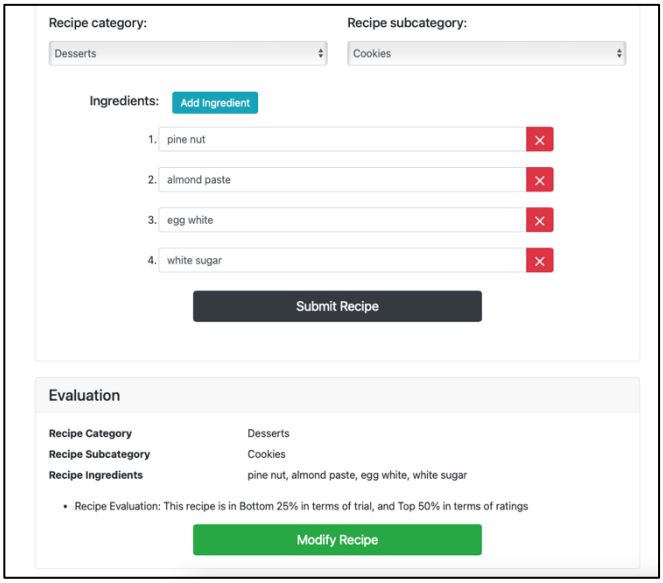


Step 4: As the user enters an ingredient name, alternative ingredients that have partial matches to the entry appear for selection. The user is forced to select from the list to make sure we identify the ingredient.[15] A valid recipe must have at least two ingredients.

[15] The list of ingredients matches the list of ingredients in our analysis for a total of 1,249 possible ingredients.

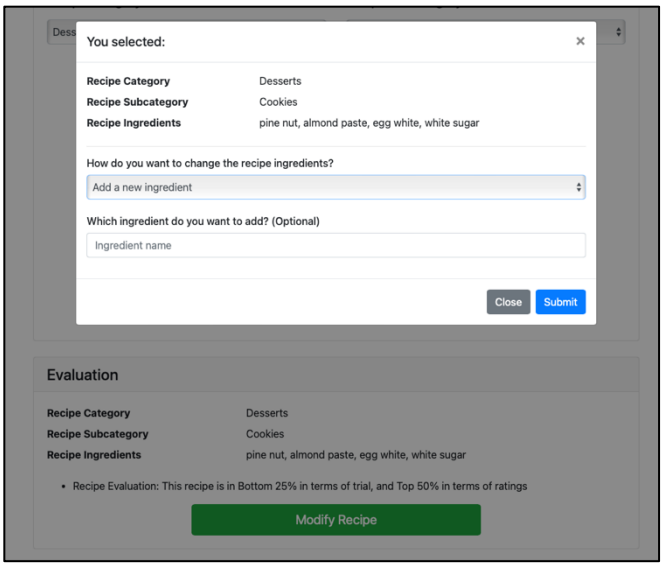


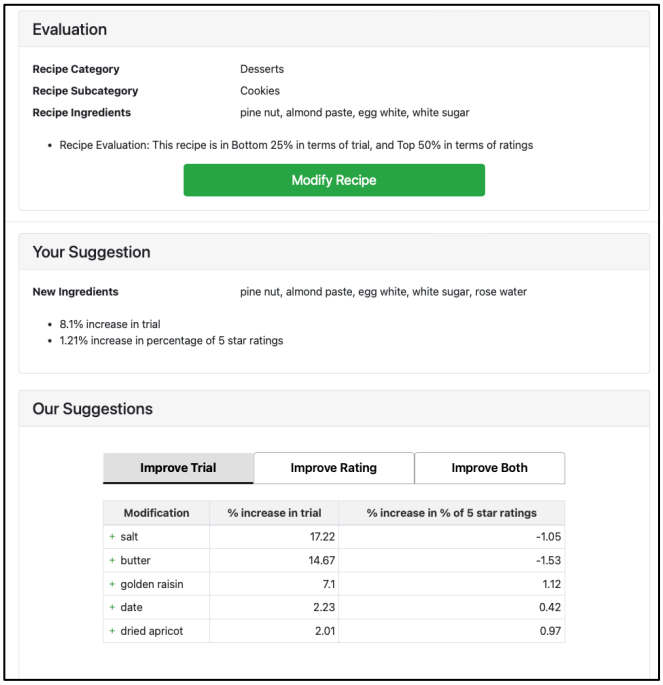
Step 5: After the user enters all the recipe ingredients and click the "Submit Recipe" button, the tool provides an evaluation of the recipe in terms of its trial and ratings percentiles relative to recipes in the subcategory on Allrecipes.com. At this point the user can change the recipe ingredients and submit the recipe again to get a new evaluation.



Step 6: After the user clicks “Modify Recipe” button, the tool prompts the user to select how they want to modify the recipe. The user can choose “Add a new ingredient”, “Change an existing ingredient” or “Show all suggestions.” After choosing the required ingredient modification type, the user clicks the “Submit” button to receive the suggestions for ingredient modification with the predicted increases in trial and ratings. The suggestions are presented in three tabs: “Improve Trial,” “Improve Ratings,” and “Improve Both.” Each tab presents the top 10 ingredient modifications that increase trail, ratings, or both, respectively. The user can navigate across the tabs for the ranked ingredient suggestion under each objective. “Improve Both” rankings are based on summation of trial and ratings rankings.

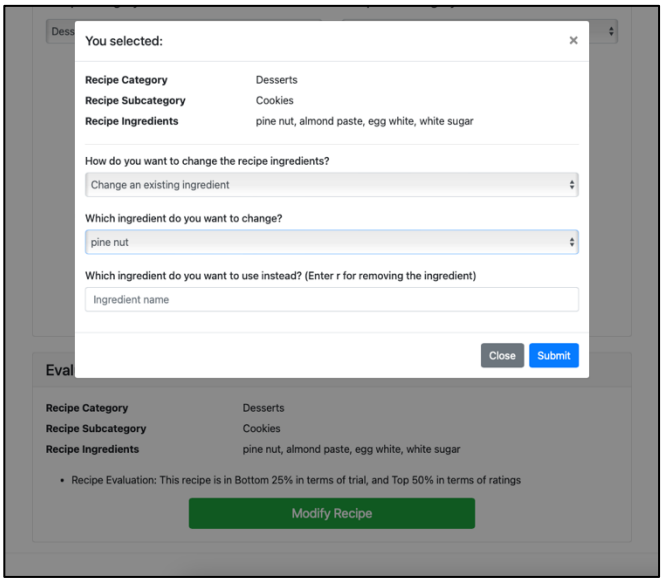
Step 6a: If the user chooses “Add a new ingredient”, they can also provide an ingredient to add to the recipe. If they have provided an ingredient suggestion, the tool also shows the recipe evaluation with their ingredient suggestion. In the example below, the user chooses to add rose water.

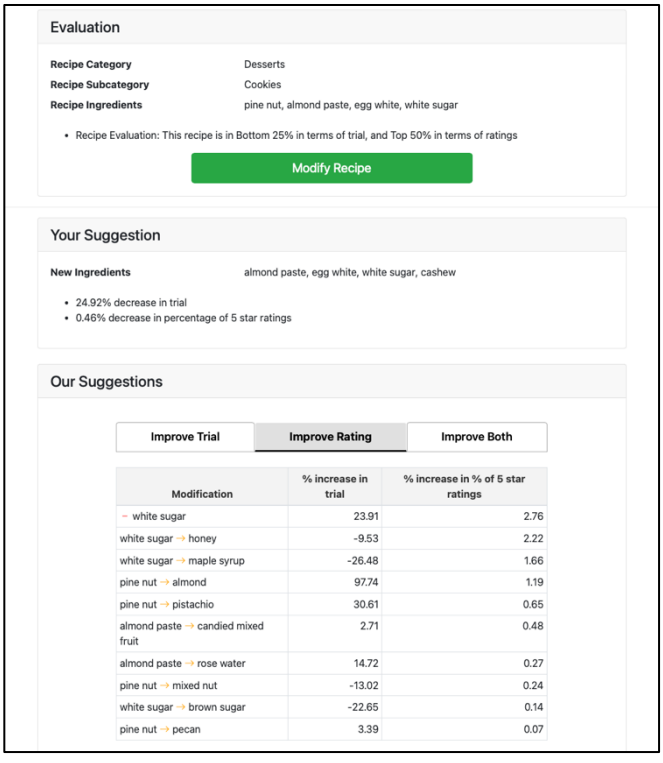




Step 6b: If the user chooses “Change an existing ingredient”, they can choose one of the recipe ingredients to change and state how they want to change it (i.e., remove or substitute with another ingredient). If they have provided an ingredient suggestion, the tool also shows the recipe evaluation with the modification. This feature can allow users to test the quality of their

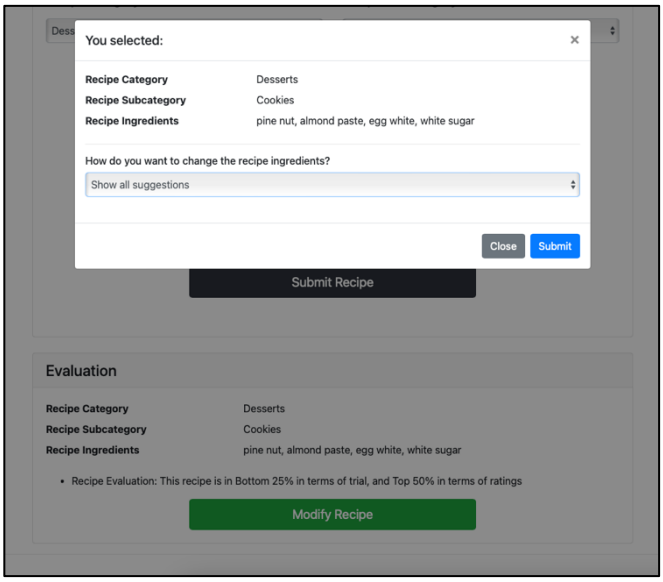
own modifications, for example due to dietary restrictions. In the example below, the user chooses to replace pine nut with cashew.

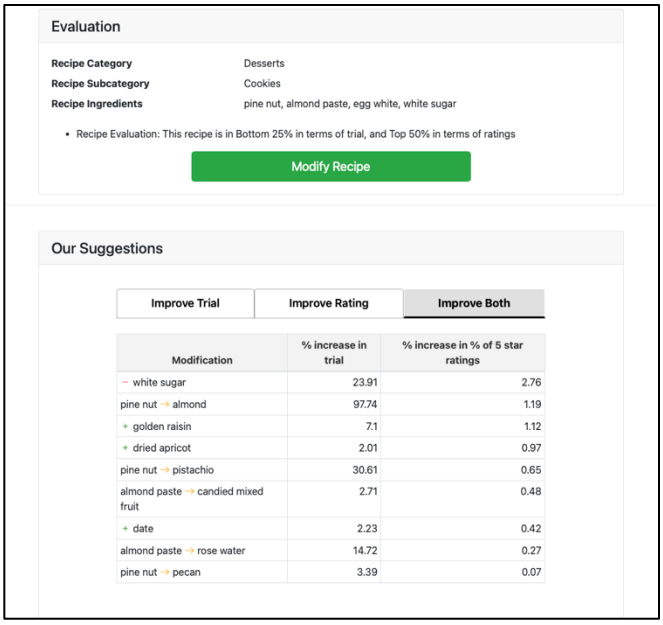




Modification	% increase in trial	% increase in % of 5 star ratings
− white sugar	23.91	2.76
white sugar → honey	-9.53	2.22
white sugar → maple syrup	-26.48	1.66
pine nut → almond	97.74	1.19
pine nut → pistachio	30.61	0.65
almond paste → candied mixed fruit	2.71	0.48
almond paste → rose water	14.72	0.27
pine nut → mixed nut	-13.02	0.24
white sugar → brown sugar	-22.65	0.14
pine nut → pecan	3.39	0.07

Step 6c: If the user chooses “Show all suggestions”, the tool provides the top 10 suggestions based on our algorithm.





Web Appendix D: Measurement Scales Used to Categorize Experts

Domain-Specific Consumer Expertise Scale

(Adapted from Mitchell and Dacin 1996; Cronbach's alpha = 0.735)

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
Compared to the average person, I know a lot about cooking.	○	○	○	○	○
I am very interested in cooking.	○	○	○	○	○
I often read about cooking and food recipes (e.g., recipe books, blogs, articles, websites).	○	○	○	○	○
I do most of the cooking in my household.	○	○	○	○	○
I enjoy experimenting with recipes.	○	○	○	○	○
I feel confident about assessing a recipe by reading the ingredient list and instructions.	○	○	○	○	○
I would describe myself as a foodie, gourmet or food connoisseur.	○	○	○	○	○