J Biosci (2022)47:12
DOI: 10.1007/s12038-021-00248-1

Review

Computational gastronomy: A data science approach to food

Mansi Goel and Ganesh Bagler*

Center for Computational Biology, Indraprastha Institute of Information Technology Delhi (IIIT-Delhi), New Delhi 110 020, India

**Corresponding author (Email, bagler@iiitd.ac.in)*

MS received 20 September 2021; accepted 2 December 2021

Cooking forms the core of our cultural identity other than being the basis of nutrition and health. The increasing availability of culinary data and the advent of computational methods for their scrutiny are dramatically changing the artistic outlook towards gastronomy. Starting with a seemingly simple question, 'Why do we eat what we eat?', data-driven research conducted in our lab has led to interesting explorations of traditional recipes, their flavor composition, and health associations. Our investigations have revealed 'culinary fingerprints' of regional cuisines across the world. Application of data-driven strategies for investigating the gastronomic data has opened up exciting avenues, giving rise to an all-new field of 'computational gastronomy'. This emerging interdisciplinary science asks questions of culinary origin to seek their answers via the compilation of culinary data and their analysis using methods of complex systems, statistics, computer science, and artificial intelligence. Along with complementary experimental studies, these endeavors have the potential to transform the food landscape by effectively leveraging data-driven food innovations for better health and nutrition.

Keywords. Computational gastronomy; food pairing; cuisine evolution; novel recipe generation

1. Introduction

Gastronomy has largely been considered an artistic endeavor despite efforts made for understanding its scientific basis (Savarin 2009). Cooking is the art of transforming raw ingredients into delicious dishes that are central to nutrition and health. Cultures across the world have imbibed idiosyncratic practices for processing ingredients into their respective cuisines. Understanding the nuances of cuisines, food, and cooking enables asking many interesting questions. Why do we eat what we eat? What is the molecular basis of flavor, of ingredients, and recipes? Can we quantify the taste of a recipe? How do we measure the nutritional profile of a recipe? How does one make sense of contradictory assertions about the health consequences of food ingredients? How have world cuisines evolved? Can we create a tasty and healthy recipe?

It is increasingly becoming apparent that investigating such questions of gastronomic origin requires a strong foundation in data and computation. Clearly, a data science approach to food is something novel and a path hitherto not taken. Similar to many other domains (weather prediction and face recognition, among others), which have been dramatically transformed with the application of data science and computation, there is abundant scope to investigate food from a data-driven perspective (Zeevi *et al.* 2015; Sonnenburg and Sonnenburg 2015). This is the vision portrayed by computational gastronomy, a data science that blends food, data, and the power of computation for achieving data-driven food innovations. This review investigates various dimensions of computational gastronomy – availability of structured gastronomical data, the molecular basis of flavor, quantification of taste, evolution of world cuisines, algorithms for novel recipe generation, among others (Figure 1).

This article is part of the Topical Collection: Emergent dynamics of biological networks.

2. Recipe data

Probing food from a data-driven perspective requires the availability of well-curated, structured data resources. Traditional recipes encode the cultural wisdom that has gone into making a delicious dish. While there is a plethora of websites (such as allrecipes.com, geniuskitchen.com, epicurious.com, foodnetwork.com, and tarladalal.com) that provide compilation of human-readable recipes, there has been a dearth of structured compilation of these recipes which enables complex queries. Addressing this problem needs implementation of natural language processing (NLP) algorithms that capture relevant information (quantity, unit, temperature, processing state, etc.) from the recipe text. Diwan *et al.* (2020) designed named entity recognition (NER) models for extracting such details using knowledge-mining techniques. These models have applications when translating recipes between languages, determining similarity between recipes, generating novel recipes, and estimating the nutritional profile of recipes.

RecipeDB (*https://cosylab.iiitd.edu.in/recipedb/*) has been created to investigate the culinary correlates of dietary elements for probing their association with sensory responses as well as consequences for nutrition and health (Batra *et al.* 2020). It is a structured compilation of recipes, ingredients, and nutrition profiles interlinked with flavor profiles and health associations. This repertoire comprises more than 118,000 recipes from cuisines across the globe (6 continents, 26 geo-cultural regions, and 74 countries), cooked using various processes (heat, cook, boil, simmer, bake, etc.), by blending thousands of ingredients. Ingredients are further linked to their flavor molecules, nutritional profiles, and empirical records of disease associations obtained from Medline.

3. Flavor data

Ingredients are chosen to be used in recipes, primarily, by virtue of their taste and odor (together referred to as flavor). Experimental assays such as gas chromatography and mass spectroscopy probe the constituent flavor molecules that are present in ingredients. A structured compilation of flavor compounds from natural ingredients is a prerequisite for quantification of taste of ingredient combinations and any subsequent investigations. FooDB (*http://foodb.ca*), one of the resources that addressed this need, compiled molecules from food ingredients, albeit its focus was not on the chemical basis of flavor. Another resource called Flavornet provides a list of flavor molecules and their odor profiles, but does not furnish information of their natural sources (Arn and Acree 1998). Some other attempts in this direction have focused on the flavor-specific compilation of data, such as bitter (BitterDB) and sweet (SuperSweet), and volatile compounds of scents (SuperScent) (Dunkel *et al.* 2009; Ahmed *et al.* 2011; Wiener *et al.* 2012). Some other efforts have targeted nutritional factors (NutriChem), polyphenols (Phenol-Explorer), and the medicinal value of food (Neveu *et al.* 2010; Scalbert *et al.* 2011; Rothwell *et al.* 2013; Jensen *et al.* 2015).

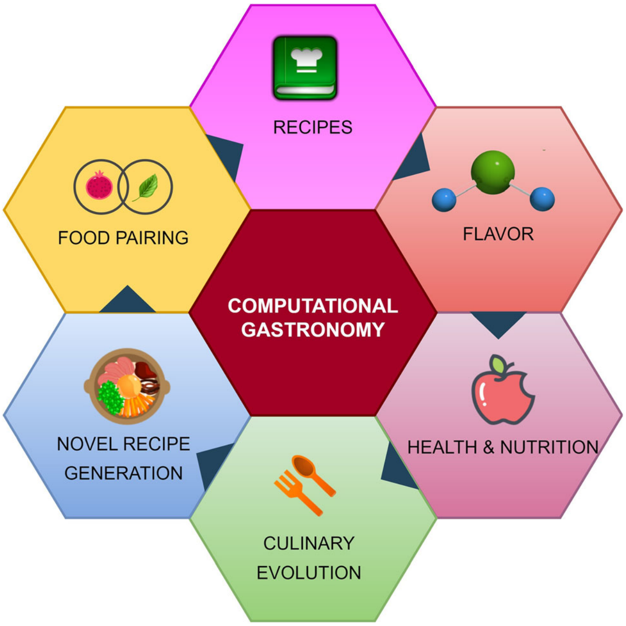

Figure 1. Computational gastronomy is a science that blends food with data and the power of computation for achieving data-driven food innovations.

FlavorDB (*http://cosylab.iiitd.edu.in/flavordb/*) is a comprehensive repository of flavor compounds, physicochemical structures, their natural sources, flavor percepts, and functional groups (Garg *et al.* 2018). It helps to find molecules of desired flavor percepts, explore ingredient molecules, discover relevant food blendings, and know the chemical properties of flavor compounds. FlavorDB contains 25,595 flavor molecules, 34 ingredient categories covering 936 ingredients linked to 527 unique natural resources. The resource is conceptually divided into two parts – entity space, representing natural ingredients utilized in food, and flavor space, representing flavor molecules and their chemical properties. Each molecule is described with a range of descriptors such as PubChem ID, CAS number, IUPAC, SMILES, and 2D/3D properties. In connection with the flavor percepts, it lists 33 taste and

1,068 odor receptors with unique UniProt ID and name. Users can search FlavorDB via the entity (ingredient) name, their category, flavor molecules, and their percepts.

4. Food pairing

In a reductionist manner, a recipe can be perceived as nothing but a combination of ingredients. With this notion, the theoretical number of recipes that can be generated is astronomically large. Even with a conservative estimate, one gets 10^{30} recipes with an average recipe size of 10 and 1,000 available ingredients. Traditional recipes, obviously, do not represent all possible combinations of ingredients due to bias in preferences, climate, geography, cultural lock-ins, religious taboos, and, to an extent, even genetics. One of the interesting questions in this regard is, 'Are there guiding principles for the choice of ingredient combinations in traditional recipes?'

One of the possible answers to this question was provided by Chef Heston Blumenthal (2008) through the food pairing hypothesis – 'ingredients with similar taste tend to blend well together'. Availability of data on ingredients used in the recipes and flavor profiles of natural ingredients facilitates investigation of food pairing hypothesis in a quantifiable manner. Food pairing index is defined as the average number of shared flavor molecules across all ingredient pairs in a recipe.

In one of the earliest computational gastronomy investigations, Ahn *et al.* probed the data of 56,498 recipes as a bipartite network of 381 ingredients used in the recipes from a range of world cuisines and their 1,021 flavor molecules (Ahn *et al.* 2011; Ahnert 2013). Here, ingredients and flavor molecules are nodes, and an edge represents association between the two. The authors created a monopartite projection of this network representing the flavor graph and investigated the food pairing phenomenon in cuisines under consideration. For this purpose, they compared the average food pairing of recipes in a cuisine with that of a random cuisine (a random control created by generating recipes by an arbitrary combination of ingredients while maintaining the recipe size distribution). In the absence of statistically significant difference between the food pairing index of a real cuisine vis-à-vis its random counterpart, one may infer that the prior does not differ from the latter in terms of preferential use of ingredient combinations. A positive deviation in food pairing would indicate uniform blend of ingredients confirming the food pairing hypothesis. On the contrary, a negative deviation would be a signature of contrasting blend of ingredient combinations.

Ahn *et al.* (2011) observed that cuisines from North America, Western Europe, Southern Europe, and Latin American show uniform (positive) food pairing, confirming Chef Blumenthal's food pairing hypothesis. East Asian cuisine (recipes from South Korea), on the other hand, showed the signature of a contrasting blend of flavor profiles.

In another study that focused on recipes from the Indian subcontinent, authors investigated food pairing phenomenon in 2,543 recipes from across eight regional cuisines of India (Jain *et al.* 2015a, b). These recipes (compiled from TarlaDalal.com) were composed of 194 ingredients each of which was exclusively categorized into one of the 15 ingredient categories such as vegetables, herbs, lentils, fruits, herbs, etc. Building on the study of Ahn *et al.* (2011), this study compiled 1,170 unique flavor compounds (Burdock 2010) and reported a signature of contrasting food pairing in Indian recipes.

Importantly, this study probed for the role of placement of specific ingredient in the recipes by performing intra-category shuffling of ingredients. For example, to find the role of placement of vegetables, all vegetable category ingredients were shuffled with a randomly selected vegetable while maintaining the position of ingredients of all other categories. No or little (statistically insignificant) change in the food pairing index of the cuisine indicates that the placement of an ingredient does not play a key role in specifying the food pairing. It was observed that, barring the 'spice' category of ingredients, the shuffling of ingredients does not affect the food pairing in Indian cuisine. However, shuffling of spice ingredients had a dramatic (statistically significant) impact on the food pairing, suggesting that spices are the molecular fulcrum of Indian recipes. This study also built statistical models to identify factors that contribute to the food pairing to conclude that popularity of ingredients (frequency of use) plays a key role in specifying the contrasting food pairing observed in Indian cuisine. In studies conducted with a similar spirit, data of Arabian (Tallab and Alrazgan 2016) and Medieval European (Varshney *et al.* 2013) cuisines were investigated and it was observed that they were characterized with positive food pairing.

Further to food pairing, Simas *et al.* (2017) proposed the food bridging hypothesis which identifies molecular association between two ingredients even in the absence of shared flavor profiles. It generalizes the idea of food pairing by creating a bridge between any two ingredients via intermediate ingredients. Starting with the data of 1,530 ingredients and 1,106 flavor compounds, same as the one used by Ahn *et al.* (2011), this study identified semi-metric paths between ingredient pairs (paths with non-zero intermediate ingredients). The food bridging score for a recipe is defined as the mean of all semi-metric links present in the recipe's sub-graph. It was concluded that while Eastern cuisine follows the food bridging hypothesis, Western cuisines do not.

These early studies paved the way for food pairing analysis of worldwide cuisines. Issa *et al.* (2018) investigated Eastern Mediterranean cuisine by probing recipes from Jordan (136), Palestine (193), Lebanon (169), and Syria (214). The analysis revealed positive food pairing in these cuisines with a strong signature of uniform blending of ingredients in Jordanian recipes, followed by those from Palestine, Lebanon, and Syria. It was observed that the food pairing could be either explained by ingredient popularity (Syrian and Lebanese cuisines) or by category composition (Jordanian and Palestinian) of recipes. In another study, Chinese cuisine comprising 8,498 recipes and 2,911 ingredients was investigated by Zhu *et al.* (2013) for the food pairing phenomenon to conclude a uniform blend of ingredients.

In one of the comprehensive food pairing studies that probed 45,772 recipes from 22 regions worldwide, Singh and Bagler (2018) showed the ubiquitous nature of positive food pairing. Consistent with earlier studies, it showed that recipe size follows a Gaussian distribution with an average recipe containing 9 ingredients. Cuisines from 16 regions (Italy, Africa, USA, France, China, Thailand, Canada, East Asia, Caribbean, Spain, Greece, Mexico, India, Middle East, Australia & New Zealand, and South America) showed positive food pairing while negative food pairing in the remaining six regions (Japan, Korea, DACH countries, Scandinavia, Eastern Europe, and Britain). By investigating the global culinary landscape in thorough detail along with an improved dataset of flavor profiles (FlavorDB), this article highlighted the importance of comprehensive data compilation.

One of the relevant topics in the context of food pairing is food-beverage pairing. Charles Spence proposed two pairing principles, viz. perceptual (sense-based) pairing and cognitive/intellectual (mind-based) pairing (Spence 2020). While the former involves food-beverage pairing based on similar aroma, taste, color, and flavor, the latter involves pairing based on cultural and geographical identity. Perceptual pairing has become quite popular lately vis-à-vis cognitive pairing.

Another question around the notion of food pairing is whether one can predict the new pairings. Park *et al.* (2019) proposed a model called KitcheNette to determine the food-ingredient pairings scores and to predict the unknown pairings. Towards this objective, they used MIT's Recipe1M database. KitcheNette utilized Siamese neural network and in-depth network to obtain maximum performance in predicting unknown pairings. Pointwise mutual information score, the ratio of probability of two ingredients occurring in a recipe to probability of ingredients occurring separately, was used to find good and bad pairings. A positive score indicated good pairing and a negative score, a bad pairing. The pairings obtained by the KitcheNette model were compared with those from FlavorDB as a reference. Such data-driven models can propel the identification of compatible ingredients in future.

As an extension of food pairing analysis, one can consider whether a recipe from one cuisine could be 'transformed' into a recipe from an altogether different cuisine (Kazama *et al.* 2018). This goal was achieved starting with the Yummly dataset of 39,774 recipes from 20 countries. The cuisine of given recipe was identified, and ingredient replacements were suggested using a modified word2vec algorithm. Further, a neural network containing 2 hidden layers and 20 output layers matching the number of cuisines was implemented to classify the cuisine. The input to the model is a recipe (binary) vector in which each dimension represents an ingredient in the dataset. As a demonstration, the authors converted a Japanese recipe 'Sukiyaki' into its French avatar by substituting ingredients based on the cosine similarity.

5. Health and nutrition

Beyond the flavor, nutrition and health impacts of food ingredients are among the major aspects. Food interacts with the human body in a complex manner, making it a challenging problem to predict their health consequences. Not so surprisingly, the scientific literature is replete with many contradictory assertions. Towards the objective of laying the data-driven foundation of computational gastronomy, it is critical to create an interactive data repository of food-disease associations.

Towards this goal, as a first step, Rakhi *et al.* (2018) mined the MEDLINE database for spice-disease associations. A list of 188 culinary herbs and spices were identified from foodb, Wikipedia, Food Plant International, Plants For A Future and FlavorDB. A total of

8,957 spice-disease (8,172 positive and 783 negative) associations were extracted from the abstracts of research articles using named entity recognition and machine-learning-based candidate sentence extraction. The Convolution Neural Network (CNN) classifier presented with an accuracy of 86.7% and precision of 90.7%. TaggerOne was used to map disease entities with their MeSH (Medical Subject Headings) IDs at three levels: category, sub-category, and specific disease.

The analysis of spice-disease associations suggests strong positive association with diabetes mellitus, carcinogenesis and inflammation, while hypertension, dermatitis, hypersensitivity, and allergy show strong negative (adverse) associations. For the quantitative analysis of spice-disease associations, a 'spectrum score' was defined which captures adverse and therapeutic effects across the MeSH disease spectrum and its integral sub-categories. The benevolence score highlights the relative benevolence vis-à-vis the adverse impact. In general, culinary herbs and spices were observed to have a high benevolence score compared to their adverse score. Spices like ginger, turmeric, and garlic showed a high benevolence spectrum score, whereas liquorice, fenugreek, ginger, sunflower, celery, and ginkgo showed an adverse spectrum score. The article also probed the spice-phytochemicals association using Phenol Explorer (Neveu *et al.* 2010) and KNapSAcK (Afendi *et al.* 2012) and association with diseases using CTD (Comparative Toxicogenomic Database) (Davis *et al.* 2017). The tri-partite data of spice-phytochemical-disease associations is available as an online repository, SpiceRx (*https://cosylab.iiitd.edu.in/spicerx*).

SpiceRx (Nk *et al.* 2018) is a repository of evidence-based information on the health effects of culinary herbs and spices, along with their phytochemicals and disease associations. Users can search associations based on the spices, disease or phytochemical names. Spices can be queried based on their common name, scientific name, and NCBI ID. Diseases can be searched based on the name, MeSH ID, disease category or sub-category. Apart from spice and disease, SpiceRx can also be searched using chemical constituents of herbs and spices through IUPAC name, PubChem ID, SMILES, molecular weight, hydrogen donors/acceptors or molecular hydrophobicity. SpiceRx thus provides a platform to investigate tripartite associations between spices, their chemicals and associated health effects.

Extending the scope of SpiceRx, Tuwani *et al.* have built a repository DietRx (*https://cosylab.iiitd.edu.in/dietrx*) which integrates the associations among food, diseases, chemicals, and genes. DietRx represents 1,781 food entities obtained by text-mining around 38,000 Medline articles using named entity recognition techniques. The data is further enriched with 6,992 food-chemicals and 20,550 food-genes associations, thereby providing a platform for the investigation of dietary ingredients and their health consequences.

Data from social networks is another dimension that can be explored to associate food with diseases. Abbar *et al.* (2015) analyzed the food consumption of US citizens through 502 million tweets of 210,000 users. The authors compared the Twitter data obtained across multiple dimensions such as gender, education and income levels with the actual obesity data provided by the Center for Disease Control (CDC) of around 50 US districts. The authors identified the tweets by checking specific keywords, e.g., food, eating, cooking, calorie count and famous food chains. The results were correlated with the CDC's state-wise obesity and diabetes data. The highest Pearson correlation with obesity was 77.2%, and with diabetes was 65.8%. Further, diets were analyzed based on users' interests across different dimensions. Interests such as television shows, education and business were positively related to obesity, whereas baseball, football, travel and such were negatively correlated. The chances of becoming obese increase by 57% if one had an obese friend in the network.

In another study that probed food-nutrition axis, Sajadmanesh *et al.* (2017) classified dishes based on ingredients, their flavor molecules and nutritional advantages. Various health statistics were used to investigate the relation of these factors with distinct health indicators such as obesity, diabetes and health expenditure. By probing 157,000 recipes the authors conclude that while sugar and carbohydrates show negative association with diabetes and obesity, proteins have positive association.

To determine the link between nutrition, sustainability and recipes, van Erp *et al.* (2021) explored nutrition and sustainability with recipe and nutrition databases. For analyzing the food's nutritional information, these authors used a nutriscore scale that helps users to make a healthier purchase. A tool resulting out of the analysis provides the health, social, economic and environmental footprints of recipes submitted by users. The authors proposed a recommendation system to minimize food waste, to recommend recipes based on ingredients, and to link recipes with supply chain and modes of production.

In another study that looked into the nutritional and dietary preferences, a knowledge-graph-based semantic approach was developed that takes ingredients as input and provides healthier ingredient substitutions (Shirai *et al.* 2021). An ingredient substitution heuristic rooted in the semantics of the ingredients was developed for the identification of substitutions. This approach was evaluated using web-scraped databases and was shown to outperform the state-of-the-art methods by suggesting healthy substitutions.

6. Taste prediction

Flavor molecules are characterized by taste and odor percepts. While there are five broad taste attributes, namely, sweet, bitter, salty, sour and umami, the number of odor percepts are much larger (Garg *et al.* 2018). Prediction of taste and odor based on molecular attributes is one of the outstanding problems in computational gastronomy (Keller *et al.* 2017). The latter is far more challenging than the former due to the complexity of odor space. Various attempts have been made towards the prediction of taste, specifically that of bitter and sweet taste (Dagan-Wiener *et al.*; Wiener *et al.* 2012; Rojas *et al.* 2017; Zheng *et al.* 2018; Banerjee and Preissner 2018).

Significantly, Tuwani *et al.* (2019b) built on the existing work to contribute state-of-the-art, machine-learning algorithms for the prediction of bitter-sweet taste. Towards this they curated 918 bitter, 1510 non-bitter, 1205 sweet, and 1171 non-sweet molecules. Molecular descriptors were generated using ChemoPy, Dragon 2D, Dragon 2D/3D, Canvas and ECFPs, which served as features. The dataset was trained using 5-fold stratified cross-validation and feature selection was done using Boruta algorithm. All the models were evaluated using various performance metrics such as precision, recall, AuPR, AuROC, F1 and sensitivity score. To determine the redundant components, preprocessing was implemented using PCA (principal component analysis) and Boruta algorithm. Random Forest, Ridge Logistic Regression and Adaboost algorithms were used to predict the taste of molecules. Adaboost and Random Forest models in coordination with Boruta algorithm with Dragon 2D/3D molecular descriptors gave more accurate results for sweet/non-sweet, whereas Random forest with PCA gave good results for bitter/non-bitter prediction. It was inferred that the average precision is a better metric as opposed to other metrics, and yielded a performance of 90%.

The dataset used in this study as well as an user-friendly implementation has been made publicly available, BitterSweet (*https://cosylab.iiitd.edu.in/bittersweet*). The webserver can be used for predicting the taste of small molecules starting with their IUPAC, SMILES, or common name. The database comprises of 394,152 molecules with predicted BitterSweet taste profiles and 3,086 with verified taste profiles.

7. Culinary evolution

Cuisines have been modified over a long time to acquire their present shape. ‘What are the factors that have gone into the evolution of cuisines?’ is one of the most interesting questions in culinary sciences. One of the earliest studies that examined this question probed the recipes from British, Medieval, Brazilian and French cuisines and proposed a copy mutate model to explain their ingredient composition (Kinouchi *et al.* 2008). According to this model, starting with an initial cuisine containing randomly concocted recipes, a recipe is chosen randomly and mutated (modified) into a daughter recipe based on the ingredient’s fitness. The daughter recipe is treated as a novel recipe if the same is not present in the existing cuisine. With such a simple ‘copy-mutate model’, the authors showed that fitness-based ingredient modification can solely explain the observed ingredient popularity in present-day cuisine.

Building on this work, Tuwani *et al.* (2019a) added a layer of flavor profile information and investigated 25 regional cuisines across the world by copy-mutate models. The authors created a variation of this model to account for random ingredient replacement, ingredient category-based replacement and a mixed strategy. They concluded that different models matched with reality depending on the cuisine under consideration.

Going beyond investigating evolution of the cuisines, efforts have been made to probe the interrelatedness of cuisines by virtue of their similarity. In one of the studies (Sajadmanesh *et al.* 2017), using the data of 157,000 recipes from World Bank, BBC and Yummly website, the authors obtained Jensen-Shannon divergence to find the ingredient-based similarity between the cuisine. To find a similar cuisine based on flavor molecules, a flavor-based correlation matrix was formed between recipes. The results showed that Welsh cuisine was dominated by Asian culture, whereas Indian cuisine was found to be having similarities with Ethiopian and African cuisines with the dominance of spices.

In another study, Sharma *et al.* (2020) curated 118,071 recipes and identified the most significant patterns in 26 geo-cultural cuisines from across the world using frequent itemset mining and ingredient authenticity. To find the similarity between cuisines, these authors implemented hierarchical agglomerative clustering using three distance metrics (Jaccard, Cosine, and Euclidean). Euclidean distance-based clustering gave good results when validated using geographical distance among cuisines. Some clusters (India and Africa, French and Canada) showed similarities in their cuisines despite being geographically separated.

8. Novel recipe generation

For ages chefs and culinary enthusiasts have attempted to create novel recipes by tweaking ingredients and processing. This process is inherently laborious and time consuming due to trail-and-error methods involved. Recent advances in computational gastronomy have led to the question, 'Can computers generate novel recipes?' One of the earliest approaches used in this direction is text generation. Text generation is a subfield of natural language processing which captures the stylistic nuances of a corpus to generate novel text similar in style to that of the corpus. It involves training of the models using various word representations for imbibing the notion of phrases, grammar and sentence construction. Recipes are structured texts with a variety of named entities such as ingredient names, quantity and unit. Novel recipe generation algorithms attempt to capture the structure and style of recipes from a large corpus of recipes in an effort to generate healthy, tasty, vegan or non-vegan, personalized, disease or diet-specific recipes.

In one of the earliest studies in novel recipe generation, Kiddon *et al.* (2016) proposed a globally coherent text generation model, the 'neural checklist model' which takes the list of ingredients as an input. This model used a traditional encoder-decoder architecture that implemented Recurrent Neural Network for generating text. It further implemented an attention mechanism that keeps track of ingredients used and those unused from a user-provided list. The recipes generated were evaluated for the grammatical correctness and use of suggested ingredients. The outcome of the checklist model was compared with four baseline models and subsequent human evaluations were performed. While the model had an average human score of 4.2 on a scale of 1 to 5, it was observed to be heavily reliant on fine-tuning of the hyper-parameters.

In another study conducted with a similar spirit, Lee *et al.* (2020) introduced an online pre-trained transformer-based application to generate recipes – RecipeGPT. RecipeGPT generates recipe instructions starting from a title and list of ingredients or produces a list of ingredients given a title and set of instructions. The GPT-2 transformer, tuned for MIT's Recipe1M recipe dataset, served as the backbone of RecipeGPT. The model implemented an evaluation system to identify overlapping ingredients and to compare recipes generated with reference recipes.

Another dimension of this domain is to generate recipes from food images. However, since food images are often not accompanied by ingredients or recipes, Salvador *et al.* (2019) proposed an 'inverse cooking system' that generates recipes from images using ingredients of a recipe. Both images and ingredients serve as input to the model, leading to the cooking instructions as the output. The proposed model demonstrated that image-to-ingredients prediction was more effective as compared to image-to-recipe retrieval. By comparing two retrieval models, the authors showed that their model outperforms both human and retrieval baselines. Further, on comparing the quality of recipes generated from the dataset and retrieval baselines, it was observed that human-written recipes were shorter than the computer generated recipes. Failure in recognizing ingredients, lack of correlation between ingredient and instruction, and repeated ingredient enumeration were among the key shortcomings of this study.

Recently, Wang *et al.* (2020) proposed Decomposed Generation Network model, which generates cooking instructions by predicting structure of food images. The instructions were split into phases based on cooking process and sub-generator was assigned, which was concatenated to generate the recipe. The paper achieved a new state-of-the-art score to generate a recipe from food images using MIT's Recipe1M dataset.

Depending on the user's culinary preference, Majumder *et al.* (2020) generated personalized recipes by taking ingredients, recipe name, and calorie level as inputs. The process comprised two sub-computational tasks, viz. text generation and a recommendation system and used dataset of 180,000 recipes and 700,000 user reviews. The base architecture used bidirectional Grated Recurrent Unit focusing on user's previously liked recipes. The results were promising, but the architecture is heavily dependent on the user's history.

In view of previous research, Yu *et al.* (2020) proposed Routing Enforced Generative Model (RGM) to generate recipes considering ingredients and user preference. The authors grouped ingredients into dish categories (low sugar, high fiber, low fat, grilling, and frying) to generate category vectors and probability of a category-specific recipe was computed using a non-linear squash function. Then, the authors performed human evaluation with the help of nine judges from Amazon Mechanical Turk and three native English speakers on the following factors: readability, accuracy, feasibility, creativity and overall quality. The recipes generated by RGM were adequate, while those from baselines achieved low scores due to repetition of ingredients or instructions and larger recipe length.

Knowing the growing need of adapting a recipe based on user preferences, Morales-Garzón *et al.* (2021) suggested an unsupervised method for adapting recipes. They used three datasets, one which they scraped to train word embedding model, second from food.com for validation of adaptation, and third was a nutritional database from COFID, containing 2,913 unique ingredients. These authors adapted a recipe in three ways: similarity-based (substitute ingredients based on similar recipes), preference-based (substitute ingredients based on user preferences), and restriction-based (substitute ingredients based on user's food constraints). While they implemented three models (word2vec, fast-text, and GloVe), they chose word2vec due to its simplicity and speed of execution. Finally, recipes were evaluated through online surveys. This model did not implement any advanced neural language models such as BERT, Elmo or GPT.

Agarwal *et al.* (2020) presented a Hierarchically Disentangled Model for text generation with named entities while curating 158,473 recipes using named entity recognition and unsupervised methods. The authors used two models, the type prediction model, and the entity transformation model, which captured typical patterns among the same type of entities and then identified the most relevant entity. The output of the type prediction model was fed into the entity transformation model. This architecture can be used for ingredient-to-recipe generations, image-to-recipe generation, action-graphs-to-recipe generation, and translating a recipe from one language to another. This study used transfer learning and action graph techniques to generate relevant recipe text. This approach heavily relies on named entities which are troublesome to annotate.

9. Discussion

Computational gastronomy is a new niche that blends food with data and computation. As a data and computation-intensive field it is rooted in well-curated, structured data and computational algorithms for their investigation. While, as described in this review, the computational gastronomy data resources on various aspects of food have been growing for the past few years, there is much to be desired on this front.

While RecipeDB provides a structured repository of recipes from cuisines across the world, the granularity with which the recipe data is compiled has much scope for improvisation. Similarly, while FlavorDB provides a comprehensive picture of flavor compounds in natural ingredients it lacks the information of their concentrations. Despite such a fine-grained compilation of data, the notion of taste of an ingredient and that of a recipe are way too complex and nuanced due to emergent properties. BitterSweet provides an extensive compilation of sweet and taste molecules to predict the taste of compounds; however, its utility is limited to single molecules and it cannot be used to predict taste of compound mixtures. Both SpiceRx and DietRx, while providing one of the most comprehensive data resources of food-disease associations, are limited in having inferred associations from the title and abstract.

The food pairing analysis provides a useful strategy for quantifying a recipe and cuisine. Going beyond pairwise ingredient combinations, one may probe the patterns in recipes via higher-order tuples. When generating recipes using language modeling, improvements in training strategy are needed to be done. The novel recipes generated achieve low scores due to repetition of ingredients or instructions and large recipe length. As the recipes dataset is crawled from online repositories which tend to be noisy, rigorous preprocessing is a prerequisite to generate novel recipes. When generating recipes from food images, one needs high-resolution food images to generate more accurate recipes. Going further, the novel recipes can be improvised to include the information of servings and nutritional profile.

Acknowledgements

MG and GB thank IIIT-Delhi for the computational resources. MG thanks IIIT-Delhi for the fellowship.

References

Abbar S, Mejova Y and Weber I 2015 You tweet what you eat: Studying food consumption through twitter. *Conf.*

Hum. Factors Comput. Syst. Proc. **2015-April** 3197–3206

Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, *et al.* 2012 KNApSAcK family databases: Integrated metabolite-plant species databases for multi-faceted plant research. *Plant Cell Physiol.* **53** https://doi.org/10.1093/pcp/pcr165

Agarwal Y, Batra D and Bagler G 2020 Building Hierarchically Disentangled Language Models for Text Generation with Named Entities Building Hierarchically Disentangled Language Models for Text Generation with Named Entities (Online)

Ahmed J, Preissner S, Dunkel M, Worth CL, Eckert A and Preissner R 2011 SuperSweet-A resource on natural and artificial sweetening agents. *Nucleic Acids Res.* **39** https://doi.org/10.1093/nar/gkq917

Ahn YY, Ahnert SE, Bagrow JP and Barabási AL 2011 Flavor network and the principles of food pairing. *Sci. Rep.* **1** 1–7

Ahnert SE 2013 Network analysis and data mining in food science: the emergence of computational gastronomy. *Flavour* **2** 2–4

Arn H and Acree TE 1998 Flavornet: A database of aroma compounds based on odor potency in natural products. *Dev. Food Sci.* **40** 27

Banerjee P and Preissner R 2018 Bitter sweet forest: A Random Forest based binary classifier to predict bitterness and sweetness of chemical compounds. *Front. Chem.*. https://doi.org/10.3389/fchem.2018.00093

Batra D, Diwan N, Upadhyay U, Kalra JS, Sharma T, *et al.* 2020 RecipeDB: A resource for exploring recipes. *Database* **2020** 77

Blumenthal H 2008 *The big fat duck cookbook* (Bloomsbury)

Burdock GA 2010 *Fenaroli's handbook of flavor ingredients*, (CRC Press)

Dagan-Wiener A, Nissim I, Abu NB, Borgonovo G, Bassoli A, Niv MY 2017 Bitter or not? BitterPredict, a tool for predicting taste from chemical structure. *nature.com* **7** 1–13

Davis AP, Grondin CJ, Johnson RJ, Sciaky D, King BL, *et al.* 2017 2017 The comparative toxicogenomics database: Update. *Nucleic Acids Res.* **45** D972-978

Diwan N, Batra D and Bagler G 2020 A named entity based approach to model recipes. *Proc. 2020 IEEE 36th Int. Conf. Data Eng. Work. ICDEW 2020* 88–93

Dunkel M, Schmidt U, Struck S, Berger L, Gruening B, *et al.* 2009 SuperScent - A database of flavors and scents. *Nucleic Acids Res.* **37** D291-294

Garg N, Sethupathy A, Tuwani R, Nk R, Dokania S, *et al.* 2018 FlavorDB: A database of flavor molecules. *Nucleic Acids Res.* **46** D1210–216

Issa L, Alghanim F and Obeid N 2018 Analysis of food pairing in some Eastern Mediterranean countries. *2018 8th Int. Conf. Comput. Sci. Inf. Technol. CSIT 2018* 167–172

Jain A, K RN and Bagler G 2015a Spices form the basis of food pairing in Indian cuisine. *arXiv1502.03815*

Jain A, Rakhi NK and Bagler G 2015b Analysis of food pairing in regional cuisines of India. *PLoS One* **10** 1–17

Jensen K, Panagiotou G and Kouskoumvekaki I 2015 NutriChem: A systems chemical biology resource to explore the medicinal value of plant-based foods. *Nucleic Acids Res.* **43** D940–945

Kazama M, Sugimoto M, Hosokawa C, Matsushima K, Varshney LR and Ishikawa Y 2018 A neural network system for transformation of regional cuisine style. *Front. ICT* **5** 1–8

Keller A, Gerkin RC, Guan Y, Dhurandhar A, Turu G, *et al.* 2017 Predicting human olfactory perception from chemical features of odor molecules. *Science* **355** 820–826

Kiddon C, Zettlemoyer L and Choi Y 2016 Globally coherent text generation with neural checklist models. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* 329-339

Kinouchi O, Diez-Garcia RW, Holanda AJ, Zambianchi P and Roque AC 2008 The non-equilibrium nature of culinary evolution. *New J. Phys.* https://doi.org/10.1088/1367-2630/10/7/073020

Lee H, Shu K, Achananuparp P, Prasetyo PK, Liu Y, Lim EP and Varshney LR 2020 RecipeGPT: Generative pre-training based cooking recipe generation and evaluation system. *Web Conf. WWW 2020* 181–184

Majumder BP, Li S, Ni J and McAuley J 2020 *Generating personalized recipes from historical user preferences*

Morales-Garzón A, Gomez-Romero J and Martin-Bautista MJ 2021 A word embedding-based method for unsupervised adaptation of cooking recipes. *IEEE Access* **9** 27389–27404

Neveu V, Perez-Jiménez J, Vos F, Crespy V, du Chaffaut L, *et al.* 2010 Phenol-Explorer: an online comprehensive database on polyphenol contents in foods. *Database*. https://doi.org/10.1093/database/bap024

Nk R, Tuwani R, Garg N, Mukherjee J and Bagler G 2018 SpiceRx: an integrated resource for the health impacts of culinary spices and herbs. *bioRxiv* 273599

Park D, Kim K, Park Y, Shin J and Kang J 2019 Kitchenette: Predicting and ranking food ingredient pairings using siamese neural networks. *IJCAI Int. Jt. Conf. Artif. Intell.* **2019** 5930–5936

Rakhi NK, Tuwani R, Mukherjee J and Bagler G 2018 Data-driven analysis of biomedical literature suggests broad-spectrum benevolence of culinary herbs and spices. *PLoS One* **13** 276105

Rojas C, Todeschini R, Ballabio D, Mauri A, Consonni V, Tripaldi P and Grisoni F 2017 A QSTR-based expert system to predict sweetness of molecules. *Front. Chem.* https://doi.org/10.3389/fchem.2017.00053

Rothwell JA, Perez-Jimenez J, Neveu V, Medina-Remón A, M'Hiri N, *et al.* 2013 Phenol-Explorer 3.0: A major update of the Phenol-Explorer database to incorporate data on the effects of food processing on polyphenol content. *Database*. *https://doi.org/10.1093/database/bat070*

Sajadmanesh S, Jafarzadeh S, Ossia SA, Rabiee HR, Haddadi H, *et al.* 2017 Kissing cuisines: Exploring worldwide culinary habits on the web. *26th Int. World Wide Web Conf. 2017, WWW 2017 Companion* 1013–1021

Salvador A, Drozdzal M, Giro-I-Nieto X and Romero A 2019 Inverse cooking: Recipe generation from food images. *arXiv:1812.06164*

Savarin B 2009 *The Physiology of Taste: Or Transcendental Gastronomy* (The Floating Press)

Scalbert A, Andres-Lacueva C, Arita M, Kroon P, Manach C, Urpi-Sarda M and Wishart D 2011 Databases on food phytochemicals and their health-promoting effects. *J. Agric. Food Chem.* **59** 4331–4348

Sharma T, Kalra J, Ahmad S, Aggarwal B and Bagler G 2020 Hierarchical clustering of world cuisines. *https://doi.org/10.1109/ICDEW49219.2020.00007*

Shirai SS, Seneviratne O, Gordon ME, Chen C-H and McGuinness DL 2021 Identifying ingredient substitutions using a knowledge graph of food. *Front. Artif. Intell.* *https://doi.org/10.3389/frai.2020.621766*

Simas T, Ficek M, Diaz-Guilera A, Obrador P and Rodriguez PR 2017 Food-bridging: A new network construction to Unveil the principles of cooking. *Front. ICT https://doi.org/10.3389/fict.2017.00014*

Singh N and Bagler G 2018 Data-driven investigations of culinary patterns in traditional recipes across the world. *Proc. IEEE 34th Int. Conf. Data Eng. Work. ICDEW 2018* 157–162

Sonnenburg ED and Sonnenburg JL 2015 A personal forecast. *Nature* **528** 484–486

Spence C 2020 Food and beverage flavour pairing: A critical review of the literature. *Food Res. Int.* **133** 109124

Tallab ST and Alrazgan MS 2016 Exploring the food pairing hypothesis in Arab cuisine: A study in computational gastronomy. *Procedia Comput. Sci.* **82** 135–137

Tuwani R, Sahoo N, Singh N and Bagler G 2019a Computational models for the evolution of world cuisines *Proc. 2019 IEEE 35th Int. Conf. Data Eng. Work. ICDEW 2019* 85–90

Tuwani R, Wadhwa S and Bagler G 2019b BitterSweet: Building machine learning models for predicting the bitter and sweet taste of small molecules. *Sci. Rep.* **9** 1–13

van Erp M, Reynolds C, Maynard D, Starke A, Ibáñez Martín R, *et al.* 2021 Using natural language processing and artificial intelligence to explore the nutrition and sustainability of recipes and food. *Front. Artif. Intell.* *https://doi.org/10.3389/frai.2020.621577*

Varshney KR, Varshney LR, Wang J and Myers D 2013 Flavor pairing in medieval European Cuisine: A study in cooking with dirty data. *arXiv1307.7982*

Wang H, Lin G, Hoi SCH and Miao C 2020 Decomposed generation networks with structure prediction for recipe generation from food images. *arXiv:2007.13374*

Wiener A, Shudler M, Levit A and Niv MY 2012 BitterDB: A database of bitter compounds. *Nucleic Acids Res.* **40** D413–419

Yu Z, Zang H and Wan X 2020 Routing enforced generative model for recipe generation. *https://doi.org/10.18653/v1/2020.emnlp-main.311*

Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, *et al.* 2015 Personalized nutrition by prediction of glycemic responses. *Cell* **163** 1079–1094

Zheng S, Jiang M, Zhao C, Zhu R, Hu Z, Xu Y and Lin F 2018 e-Bitter: Bitterant prediction by the consensus voting from the machine-learning methods. *Front. Chem.* *https://doi.org/10.3389/fchem.2018.00082*

Zhu YX, Huang J, Zhang ZK, Zhang QM, Zhou T and Ahn YY 2013 Geography and similarity of regional cuisines in China. *PLoS One* **8** *https://doi.org/10.1371/journal.pone.0079161*

Corresponding editor: Susmita Roy