CULINARY PUZZLE PIECES: HIGHER-ORDER INGREDIENT PAIRINGS IN WORLD CUISINES

by
James Marcogliese
BTech, McMaster University, 2017
AdvDip.Tech, Sheridan College, 2014

A Major Research Project
presented to Toronto Metropolitan University

In partial fulfillment of the
requirements for the degree of
Master of Science
in the program of
Data Science and Analytics

Toronto, Ontario, Canada, 2023
©James Marcogliese, 2023

AUTHOR'S DECLARATION FOR ELECTRONIC

SUBMISSION OF A MAJOR RESEARCH PROJECT (MRP)

I hereby declare that I am the sole author of this Major Research Paper. This is a true copy of the MRP, including any required final revisions.

I authorize Toronto Metropolitan University to lend this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize Toronto Metropolitan University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.

James Marcogliese

CULINARY PUZZLE PIECES: HIGHER-ORDER INGREDIENT PAIRINGS IN WORLD CUISINES

James Marcogliese
Master of Science 2023
Data Science and Analytics
Toronto Metropolitan University

ABSTRACT

This study applies data-driven techniques to explore the culinary structure of world cuisines, identifying correlations between shared and unique flavour compounds and frequent ingredient combinations. Analyzing food pairing and bridging, the research categorizes cuisines into four classes based on their flavour compound pairing and bridging characteristics. Frequent pattern mining is employed to identify common ingredient tuples across cuisines. It further utilizes decision tree classifiers to predict whether a cuisine's tendency towards flavour pairing or bridging can be inferred from its ingredient combinations. The findings can inform recipe creation, enhance culinary tradition understanding, and advance computational gastronomy by aiding AI recipe generators to generate palatable and tradition-respecting recipes.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Dr. Ravi Vatrapu, for his guidance and valuable suggestions during the course of this research.

I am also thankful to the faculty of the Masters Program at Toronto Metropolitan University for providing a stimulating environment, resources, and support which have been indispensable.

Perseverance has been key in this journey, and I would be remiss not to acknowledge my family and friends for being the very embodiment of this virtue. Each tough time was made lighter by their constant understanding, for which I hold heartfelt appreciation. Encouragement from them formed the backbone of my resilience. Never failing to rally around me, their steadfast friendship and support merit my sincere gratitude.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Problem Definition

Whether or not there are guiding principles for the choice of ingredient combinations in traditional recipes is at the heart of the computational gastronomy field. Chef Heston Blumenthal (2008) proposed that ingredients with similar tastes tended to blend well together, and data on recipe data has been used to investigate this food pairing hypothesis. Ahn et al. (2011) investigated the food pairing phenomenon in cuisines and found uniform and contrasting examples between Western and Eastern cuisines. Further to food pairing, Simas et al. (2017) proposed the food bridging hypothesis, which identified a molecular association between ingredients without shared flavours; a bridge can be created between two ingredients via intermediate components. While subsequent research since these foundational papers have uncovered much in the structures and variances within geographic and cultural sub-regions of a cuisine, findings have been limited to pairwise ingredient combinations and macro-level cuisine analysis; analysis of higher-order sets of elements within cuisines has not yet been probed.

Research Question

As mentioned in the problem definition, prior research has uncovered patterns of food pairing and food-bridging within world cuisines. However, findings have been limited to pairwise ingredient combinations and macro-level analysis of regional cuisines. Going beyond pairwise ingredient combinations, what are the patterns at high order n-tuples (i.e. triples and quadruples of ingredients) present within global cuisines? Do these higher-order ingredient combinations adhere to their parent cuisines' food pairing/briding patterns, or do they diverge? Are there commonalities in ingredient combinations between world cuisines? Such questions have been posted prior in Bagler et al. (2018), so this research is poised to contribute.

Datasets

CulinaryDB

CulinaryDB is a repository of structured data of recipes and ingredients across 22 world regions, intended to enable data-driven explorations of recipes. In conjunction with data on flavour molecules from FlavorDB, CulinaryDB facilitates multi-level analysis of traditional recipes (recipes, ingredient composition and flavour pairing).

Bagler, Ganesh & Singh, Navjot. (2018). Data-Driven Investigations of Culinary Patterns in Traditional Recipes Across the World. 157-162. 10.1109/ICDEW.2018.00033.

FlavourDB

FlavorDB is a resource with extensive coverage of 25,595 flavour molecules. Among molecules listed in the database, 2,254 have been reported to be found in 936 natural entities/ingredients. These natural ingredients have further been classified into 34 categories and mapped to 527 distinct natural sources. The features provided as part of these compounds' detailed molecular and flavour profiles impact their taste and odour through gustatory and olfactory sensory mechanisms.

Neelansh Garg†, Apuroop Sethupathy†, Rudraksh Tuwani†, Rakhi NK†, Shubham Dokania†, Arvind Iyer†, Ayushi Gupta†, Shubhra Agrawal†, Navjot Singh†, Shubham Shukla†, Kriti Kathuria†, Rahul Badhwar, Rakesh Kanji, Anupam Jain, Avneet Kaur, Rashmi Nagpal, and Ganesh Bagler*, FlavorDB: A database of flavour molecules, Nucleic Acids Research, gkx957, (2017). †Equal contribution *Corresponding Author

Flavor Network

The Flavor Network dataset provides recipe data and details on molecular compounds of ingredients. The recipe component comprises 56,498 recipes gathered from two American and one Korean repositories, grouped into 5 geographically distinct cuisines. The molecular data contains 1,106 compound names and their associated CAS number.

Ahn Y.-Y., Ahnert, S.E., Bagrow, J.P. & Barbasi, A.-L. Flavor network and the principles of food pairing. *Sci. Rep*. 1, 196; DOI:10.1038/srep00196 (2011)

LITERATURE REVIEW

A recipe for a dish, taken at its most basic form, can be understood as an amalgamation of ingredients. Even a modest number of ingredients can yield an enormous number of combinations (a set of 500 ingredients with a recipe size of 10 yields 3.05 x 10^23 combinations!), and we know traditional recipes do not represent all ingredient combinations due to geography, cultural inheritance, and in some cases, genetics. One crucial question, then, is if there are principles for ingredient combinations in traditional recipes. This question rests at the core of computational gastronomy, which applies data science principles to understand culinary practices. Research in this domain ranges from databases of flavour molecules to analysis of culinary patterns in world cuisines, giving insights into the science behind food preparation. Inspired by a hypothesis published by Chef Heston Blumenthal (2008) which suggested that ingredients that taste similar tend to pair well together, the academic community has undertaken a myriad of analyses to prove such a statement. This hypothesis has spawned research into topics such as food pairing and so-called food bridging through analytical means by disseminating large recipe datasets from the world wide web.

Food Pairing

Food pairing is the practice of combining ingredients based on their flavour profiles to create enjoyable dishes. This approach has been used in traditional cooking practices for centuries or more, but computational gastronomy has allowed a more systematic approach to identifying ingredient combinations. As mentioned previously, Chef Heston Blumenthal (2008) proposed that “ingredients with similar taste tend to blend well together”. One of the earliest studies in this field to prove or disprove this hypothesis was conducted by Ahn et al. (2011), who used a network-based approach to identify flavour compounds shared by different ingredients, with 381 ingredients used as the nodes and 1021 flavour molecules used as the edges. The created flavour network based on recipe data was used to explore the impact of flavour compounds on different ingredient combinations and identify which ingredients were interconnected to one another based on the frequency of their appearance together. Perhaps unsurprisingly,

natural categories of ingredients appeared clustered, but more valuably, network edges between categories linked different classes of food; as reported by the authors, fruits and dairy products are close to alcoholic drinks, and mushrooms appear isolated, as they share a statistically significant number of flavour compounds only with other mushrooms. The other significant findings of this influential paper were the identifications of both uniform and contrasting examples of ingredient combinations between Western and Eastern cuisines by comparing the average food pairing of recipes in a given cuisine against a random control cuisine made up of arbitrary ingredient combinations. The difference between the control and the actual cuisine compound sharing scores highlighted underlying patterns, finding that complementary food pairings are more likely to be found in Western cuisines and contrasting pairings in Eastern ones. The work put forth by this paper was highly influential and spawned a number of future research papers across the field to replicate the results and take a closer look into regions within cuisines to see if the patterns differed. To further prove the contrasting pairing primarily observed in Eastern cuisines, Jain et al. (2015) in "Spices form the basis of food pairing in Indian Cuisine," found that spices, due to their rich molecular composition, often serve as a foundation for food pairing in Indian cuisine, contrasting the Western tendency to pair based on shared flavour compounds. This is further expounded upon in their 2017 study "Analysis of Food pairing in regional cuisines of India," where they explored regional variations in Indian cuisine, showcasing how regional variations might impact food pairing principles.

In addition to the analysis of the pairing of ingredients by looking at the number of overlapping flavours within recipes in cuisine, the importance of individual ingredients was tested as well. The study mentioned earlier, published by Ahn et al. (2011) and then replicated by Jain et al. (2015) and Singh et al. (2018), proposed multiple control models to compare against the actual cuisine's shared compound score: random, preserved ingredient frequency, preserved category frequency, and a mixture of both ingredient and category frequency. By sampling ingredients accordingly and then computing a z-score to measure the statistical significance of the difference in food pairing scores between the actual and test, they found that preserving the observed frequency of the ingredients reproduced the food pairing patterns of the cuisine while

the preserved category models did not. One could infer that barring spices, as long as the general ratio of ingredient categories within a cuisine does not change, then the pairing index would not either. Further exploration into identifying the ‘fingerprints’ of particular cuisines using ingredient category ratios could yield interesting data.

Food Bridging

Related to the concept of food pairing is the concept of food bridging, explored in "Food-bridging: a new network construction to unveil the principles of cooking" (Simas et al., 2017), which refers to a practice in gastronomy that involves creating ingredient combinations based on a chain of molecular associations rather than relying solely on shared flavours. The concept suggests that certain ingredients can be connected or "bridged" through the presence of intermediate components, even if they do not have similar taste profiles. While food pairing intensifies flavour through ingredients with shared flavours, food bridging smoothes the contrasts.

While a single edge can exist between two ingredients that share flavours, which is considered “metric”, alternate extended paths can exist, considered “semi-metric”. The shorter the number of hops, the stronger the relationship is, but this flexibility in allowing hops allows for surprising combinations. Such an example is explained by the authors, which shows the relationship between garlic and strawberry, which do not share flavours, but can be bridged with intermediate hops between roasted onion and bantu beer, or alternatively roast beef. These two semi-metric paths may inspire a garlic-strawberry sauce served with roast beef.

Simas et al. (2017) confirmed the findings of Ahn et al. (2011), Jain et al. (2015) and Ahn and Ahnert (2013) while further redefining the placement of global cuisines within high and low food pairing and food bridging categories. Southern, Eastern, and Western European cuisines, along with North American cuisines, were characterized by higher food pairing and lower food bridging. Latin American cuisines have high pairing and bridging, East Asian low pairing and bridging, and lastly, Southeast Asian has high bridging and low pairing. From these findings, a myriad of analyses on worldwide cuisines followed, including publications by Issa et al. (2018) on Eastern Mediterranean and Zhu et al. (2013) on Chinese cuisine, respectively. In "Analysis of Food Pairing in

Some Eastern Mediterranean Countries," (Issa et al., 2018), it was noted that food pairing could be explained by category composition or ingredient popularity, which touches upon earlier notes by Ahn et al. (2011). Can cuisines be uniquely identified with a set of unique ingredients and food pairing/bridging indexes?

Intra-Cuisine Investigations

In a broader perspective, "Hierarchical clustering of world cuisines" (Zhou et al., 2018), delves into how the principles of computational gastronomy can be applied to classify cuisines across the globe. Pattern mining and hierarchical clustering techniques were employed, identifying unique culinary clusters based on ingredient use and preparation methods, thus corroborating the idea that food pairing and bridging principles could vary significantly across different cultures and regions, which appears to align with geopolitical distributions. This notion of variance across cuisines is reiterated in the study, "Data-driven investigations of culinary patterns from traditional recipes across the world" (Singh et al., 2018). They outlined a macro-level analysis of food pairing principles, repeating the work of Ahn et al. (2011), uncovering both contrasting and uniform examples across 22 regions worldwide. One important finding to note was that while the methods were borrowed from Ahn et al. (2011), the flavour dataset used was FlavorDB (Varshney et al., 2018) and with their own curated recipe set (which was released as CulinaryDB), the food pairing scores of the results different from that of Ahn et al. (2011) and Jain et al. (2015). One of the most notable differences was the classification of the cuisine of “Indian Subcontinent” as showing positive food pairing patterns, which runs contrary to the results of Ahn et al. (2011) and Jain et al. (2015), among others. This raises an important question: On the assumption that the quality of flavour data used increased from what was available previously, were the old datasets fundamentally flawed, thus leading to incorrect labelling of cuisines? If true, then the findings in Ahn et al. (2011) and Jain et al. (2015), among others, would need to be reexamined.

Aside, some interesting findings from the paper suggest linkages between geographically disparate regions that share cuisine characteristics due to similar cultural backgrounds, like Canadian and French cuisines (owed to their shared histories as

Canada being a French colony) and Indian and North African (through their abundant use of spices). World regions had an average of 321 unique ingredients, with a consistently shared average of around 9 ingredients per recipe.

The practice of food pairing and bridging aims to explain the patterns underlying food choices in human cultures while allowing us to explore new and innovative combinations of ingredients that may not traditionally be paired together. By considering the chemical compounds and molecular interactions between ingredients, chefs and researchers can create unique, unconventional, and yet harmonious flavour combinations. These works hint towards the need for investigating higher-order ingredient combinations, a largely unexplored area that could yield further insights into the principles of cuisine construction. Are there shared higher-order ingredient combinations among cuisines? Are there unique groupings? Do the pairing and bridging indexes of these higher-order ingredient combinations adhere to the overall index observations of the parent cuisine(s)?

EXPLORATORY DATA ANALYSIS

This Exploratory Data Analysis report aims to provide an overview of the FlavorDB, CulinaryDB and Flavor Network datasets, sourced from "FlavorDB: a database of flavor molecules" (Varshney et al., 2018), CulinaryDB from “Data-driven investigations of culinary patterns in traditional recipes across the world.” sourced by Singh et al., (2018), and Flavor Network from “Flavor network and the principles of food pairing” by Ahn et al. (2011) respectively. These datasets serve as valuable resources for understanding flavour compounds and culinary ingredients, offering a wealth of information for various research and practical applications in the field of gastronomy and food science.

FlavorDB is a database that houses a diverse collection of identified flavour molecules, providing detailed descriptions of their chemical properties and associated natural or synthetic sources. It encompasses a wide range of compounds, including natural and artificial flavours, aromatic compounds, and additives. FlavorDB plays a crucial role in exploring the composition and characteristics of flavours, aiding in the development of new food products, recipes, and flavour pairing.

CulinaryDB, on the other hand, focuses on culinary ingredients and their relationships. It encompasses a broad array of ingredients, such as fruits, vegetables, herbs, spices, and more. CulinaryDB provides a comprehensive repository of ingredients, recipes and the cuisines they belong to. This database serves as a valuable resource for understanding the culinary world, facilitating recipe development, ingredient substitutions, and flavour pairings.

The Flavor Network ingredient-compound data contains both identified molecules and culinary ingredients, albeit with fewer attributes than the other two datasets. This dataset will fill gaps in data left by the other two for this research.

Data Description

CulinaryDB

The CulinaryDB database consists of 4 CSV (comma-separated values) files, provided as a bundle in a .zip file.

- **01_Recipe_Details.csv**:

- Details of all recipes in the database.
- Structure: A unique 'Recipe ID', 'Recipe Title', 'Source' (from which it was fetched), and the 'Regional Cuisine' to which it belongs.
- Contains 45,565 records.

Table 1

First two records of the 01_Recipe_Details.csv file.

Recipe ID	Title	Source	Cuisine
1	5 spice vegetable fried rice	TARLA_DALAL	Indian Subcontinent
2	aachar aaloo	TARLA_DALAL	Indian Subcontinent

Singh, N., & Bagler, G. (2018). Data-driven investigations of culinary patterns in traditional recipes across the world. 2018 IEEE 34th International Conference on Data Engineering Workshops (ICDEW), 157-162.

- **02_Ingredients.csv**:
  - List of all ingredients present in the database.
  - Structure: 'Name of Aliased Ingredient', its 'Synonyms', a unique 'Entity ID' and the 'Category' to which the ingredient belongs.
  - Contains 930 records.

Table 2

First two records of the 02_Ingredients.csv file.

Aliased Ingredient Name	Ingredient Synonyms	Entity ID	Category
Egg	egg	0	Meat
Bread	Bread; bun	2	Bakery

Singh, N., & Bagler, G. (2018). Data-driven investigations of culinary patterns in traditional recipes across the world. 2018 IEEE 34th International Conference on Data Engineering Workshops (ICDEW), 157-162.

- **03_Compound_Ingredients.csv**:
    - Description of "Compound Ingredients".
    - Structure: name of the 'Compound Ingredient', its 'Synonyms', a unique 'Entity ID', 'Components' that constitute a compound ingredient and the 'Category' of each compound ingredient.
    - Contains 103 records.

Table 3

First two records of the 03_Compound_Ingredients.csv file.

Compound Ingredient Name	**Compound Ingredient Synonyms**	**entity_id**	**Constituent Ingredients**	**Category**
Garam Masala	garam masala	2000	black pepper, mace, cinnamon, clove, cardamom,...	Spice
Ginger Garlic Paste	ginger garlic paste	2001	ginger, garlic	Spice

Singh, N., & Bagler, G. (2018). Data-driven investigations of culinary patterns in traditional recipes across the world. 2018 IEEE 34th International Conference on Data Engineering Workshops (ICDEW), 157-162.

- **04_Recipe-Ingredients_Aliases.csv**:
    - Enumerates the mapping of each ingredient of each recipe to one of the entities in our database.
    - Structure: A 'Recipe ID', 'Original Ingredient Name' (as in the recipe), 'Aliased Ingredient Name (from FlavorDB or Compound Ingredient)' and its 'Entity ID'.
    - Contains 456,279 records

Table 4

First two records of the Ingredients_Aliases.csv file.

Recipe ID	Original Ingredient Name	Aliased Ingredient Name	Entity ID
1	capsicum	capsicum	362
1	Green bell pepper	pepper bell	362

Singh, N., & Bagler, G. (2018). Data-driven investigations of culinary patterns in traditional recipes across the world. 2018 IEEE 34th International Conference on Data Engineering Workshops (ICDEW), 157-162.

As this dataset is a cleaner, though a smaller version of the RecipeDB dataset, no major preprocessing is required to get the data into a ready-to-use state. Aside from combining the ingredients and recipes into convenient data frames for later querying, some light cleaning was employed to remove the ‘Miscellaneous’ cuisines, as limited records were provided for those. This left a total of 22 world cuisines included in the dataset. One feature was engineered in order to perform dietary analysis later on; an identification of the dietary style of each recipe. In the larger RecipeDB dataset, a dietary style based on the ingredients used in a recipe was present, but not in CulinaryDB. To replicate the feature, rules were formed to associate a recipe with a specific dietary style. Firstly, the recipe should not include any ingredients from the restricted categories associated with the dietary style. Additionally, it should contain one or more ingredients from each of the mandatory categories defined for that dietary style. Regarding the 'bakery' category, which corresponds to the original feature in RecipeDB, any recipe containing an ingredient classified under this category is assumed to include eggs. Following were the rules used:

- Vegan
    - Exclude: meat, eggs, dairy, fish, seafood, dish
- Pescetarian
    - Include: fish or seafood

    - Exclude: meat, dairy, dish
- Lacto-Vegetarian
    - Include: dairy
    - Exclude: meat, eggs, fish, seafood, dish
- Ovo-Vegetarian
    - Include: egg
    - Exclude meat, fish, seafood, dairy, dish
- Ovo-Lacto Vegetarian
    - Include: egg and dairy
    - Exclude: meat, fish, seafood, dish

A recipe was not mapped to any dietary style if it had an ingredient in the ‘dish’ category as the underlying ingredients could not be determined. Any remaining recipes were labelled as ‘Non-Vegetarian’.

Table 5

First two records of the dietary file DataFrame.

Recipe ID	**Dietary Style**
1	Vegan
2	Lacto-Vegetarian

FlavorDB

The FlavorDB dataset is not provided in an easy-to-access downloadable format and requires scraping the hosting website for the data. A Python script was run to request the FlavorDB entities from the website one at a time, which was downloaded in JSON format. 939 entities were downloaded in this manner. Each entry contains a vast number of attributes, including category, alias’, molecules, flavour profiles, and detailed information on the chemical compositions, most of which are beyond the scope of what is required for this project. The total number of attributes is significant and will not be listed here exhaustively, but among the essential qualities were: 'entity_id', 'entity_alias_readable', 'entity_alias_synonyms', 'natural_source_name',

'category_readable', 'molecules', ‘alias’, ‘synonyms’, ‘scientific name’, ‘category’, ‘molecules’, ‘pubchem id’, ‘common name’, ‘flavor profile’. In addition, some of the attributes were contained within incorrect or unwieldy data types and needed to be adjusted. The following attributes were re-cast:

- 'entity id', 'pubchem id' cast to type int
- ''common name', alias', 'category', 'scientific name' cast to type str
- 'flavor profile', 'synonyms', cast to type set(str)
- 'molecules' cast to type set(int)

Ingredients and their names were mapped to synonyms and molecules for later searching, and common names of molecules to their PubChem IDs and related flavour profiles, yielding ‘intermediate’ data sets in a more malleable format for further processing. Table 6 shows the head of the Intermediate ‘Flavor’ DataFrame.

Table 6

First three records of the intermediate FlavorDB entity file.

entity id	**alias**	**synonyms**	**scientific name**	**category**	**molecules**
0	egg	{egg}	chicken	animal product	{6274,5311110 …}
1	bakery products	{bakery products}	poacceae	bakery	{27457, 7976…}
2	bread	{bread}	poacceae	bakery	{1031, 1032…}

Varshney, K. R., Pinel, F., Varshney, L. R., Bhattacharjya, D., Schelldorfer, J., Korhonen, J. H., & Freiwald, C. (2018). FlavorDB: a database of flavor molecules. IBM Journal of Research and Development

This intermediate data set was then subject to additional cleaning to correct outstanding issues in the dataset, which included the deduplication of records, removal of erroneous categories, and adjustments of outliers.

One feature was engineered for this dataset: ingredient ‘group’. This included merging food categories and tagging each entity with a group using the logic in Table 7. The base of this logic was taken from a GitHub repository, ‘Food Pairing and Data

Science' (Choo, V.), of which the author processed an older (version 1) of the FlavorDB dataset.

Table 7

Category mapping rules.

Input Category	Input Food	Output Group
bakery, vegetable tuber, cereal		grain
flower, fungus, plant, cabbage, vegetable fruit, herb, gourd, vegetable		vegetable
fruit-berry, berry, fruit, fruit citrus		fruit
legume, nut, seed, seafood, fish, meat		protein
dairy		dairy
fruit essence, additive, spice, essential oil		seasoning
beverage alcoholic		alcohol
beverage		beverage
plant derivative	sauce, vinegar, cocoa, creosote, storax	seasoning
plant derivative	seed, peanut butter	protein
plant derivative	butter, oil	fat
plant derivative	fermented tea	beverage
plant derivative	honey, chocolate, chocolate spread	sugar
plant derivative	macaroni	grain
plant derivative	jute, tofu	vegetable
plant derivative	soy yogurt	dairy
additive	sugar, fruit preserve, syrup, icing, molasses	sugar
additive	margarine, cooking oil, shortening	fat
additive	sauce, gelatin dessert, spread,	[no classification]

	topping, water	
additive	stuffing	grain

Additionally, some exceptions were coded as better aligned with a different category, as shown in Table 8.

Table 8

Food group exception rules.

Input food	**Output Group**
Soybean oil, cooking oil, fish oil, peanut oil, canola oil, corn oil, butter, ghee, margarine	fat
sugar, honey, molasses, agave, dulce de leche	sugar
irish moss, kelp, kombu, wakame	vegetable
butternut squash, winter squash, japanese pumpkin	vegetable
sweet custard, candy bar, chocolate mousse, fudge	sugar
cocoa	seasoning

In the end, 35 categories were reduced to 31, and each entity was labelled with a group.

Final Groups: grain, beverage, alcohol, vegetable, fat, dairy, seasoning, fruit, protein, sugar.

Final Categories: Meat, Bakery, Beverage, Beverage Alcoholic, Cereal, Maize, Dairy, Essential Oil, Fruit, Seafood, Fish, Flower, Fungus, Herb, Dish, Nuts & Seed, Legume, Plant, Spice, Vegetable, Additive.

Table 9

First three records of the post-processed FlavorDB entity file after group tagging.

entity id	alias	synonyms	scientific name	category	molecules	group
0	egg	{egg}	chicken	animal product	{6274,5311110…}	grain
1	bakery products	{bakery products}	poacceae	bakery	{27457, 7976…}	grain
2	bread	{bread}	poacceae	bakery	{1031, 1032…}	grain

Varshney, K. R., Pinel, F., Varshney, L. R., Bhattacharjya, D., Schelldorfer, J., Korhonen, J. H., & Freiwald, C. (2018). FlavorDB: a database of flavor molecules. IBM Journal of Research and Development

The molecule DataFrame shown in Table 10 did not require further cleaning.

Table 10

First three records of the post-processed FlavorDB molecules DataFrame.

pubchem id	common name	flavor profile
4	1-Aminopropan-2-ol	{fishy}
49	3-Methyl-2-oxobutanoic acid	{fruity}
58	2-oxobutanoic acid	{creamy, lactonic, sweet, brown, caramel}

Varshney, K. R., Pinel, F., Varshney, L. R., Bhattacharjya, D., Schelldorfer, J., Korhonen, J. H., & Freiwald, C. (2018). FlavorDB: a database of flavor molecules. IBM Journal of Research and Development

It is worth noting that the processed FlavorDB data, while significant, is not fully comprehensive. ‘Urchin’, for example, a popular type of seafood, is not present. Same with ‘cayenne’, an extremely popular spice. This will invariably impact analysis down the line.

Flavor Network

The Flavor Network ingredients-compound database consists of 3 TSV (tab-separated values) files, provided as a bundle in a .zip file.

- **comp_info.tsv**:
    - Details of all compounds in the database.
    - Structure: A unique ‘id', 'Comound name, 'CAS number'
    - Contains 1,107 records.

Table 11

First two records of the comp_info.tsv file.

# id	Compound name	CAS number
0	jasmone	488-10-8
1	5-methylhexanoic_acid	628-46-6

Ahn Y.-Y., Ahnert, S.E., Bagrow, J.P. & Barbasi, A.-L. Flavor network and the principles of food pairing. *Sci. Rep*. 1, 196; DOI:10.1038/srep00196 (2011)

- **ingr_comp.tsv**:
    - List of all associations between ingredient id and compound id.
    - Structure: '# ingredient id', 'compound id'.
    - Contains 36,781 records.

Table 12

First two records of the ingr_comp.tsv file.

# ingredient id	compound id
1392	906
1259	861

Ahn Y.-Y., Ahnert, S.E., Bagrow, J.P. & Barbasi, A.-L. Flavor network and the principles of food pairing. *Sci. Rep*. 1, 196; DOI:10.1038/srep00196 (2011)

- **ingr_info.tsv**:
    - List of ingredients and their categories.
    - Structure: '# id', 'ingredient name', ‘category’.
    - Contains 1,530 records.

Table 13

First two records of the ingr_info.tsv file.

# id	ingredient name	category
0	magnolia tripetala	flower
1	calyptranthes parriculata	plant

Ahn Y.-Y., Ahnert, S.E., Bagrow, J.P. & Barbasi, A.-L. Flavor network and the principles of food pairing. *Sci. Rep*. 1, 196; DOI:10.1038/srep00196 (2011)

The Flavor Network dataset, though not used as a source of recipes, served as a useful source of compound-ingredient combinations that were found missing from FlavorDB. Owing to the presence of the ‘CAS number’ feature present in both datasets, merging of records was possible.

First, the list of CAS IDs was flattened from the pre-processed molecules DataFrame and merged together with a normalized frame of the Flavor Network ingredient-compounds dataset. Ingredients from the main dataset used for this research were then enriched with the molecular data from the Flavor Network ingredient-compounds dataset where the existing data was missing. 55 essential ingredients were filled in using this method, including ingredients like ‘cayenne’, ‘coconut oil’, ‘lemon juice’, ‘salmon’, and so on.

Data Overview

CulinaryDB

A stats table was assembled, shown in Table 14, combining the data and performing some simple calculations showing the recipe count, avg ingredient count, total ingredient count, recipe percentage each attributed, and ingredient count attribution for each of the cuisines from the processed CulinaryDB set. We can see that the USA cuisine contributed the majority of recipes to the total dataset at around 35%, equating to 16106 recipes, lowest from Korea at 301 recipes, while the median count sat at 875. Interestingly, there seems to be a significant amount of uniformity around the average ingredient count within recipes within cuisines, sitting at around 10 worldwide. Similarly, the number of unique ingredients within each cuisine averages around 320, with a median value of 303 ingredients. Figure 1 and Figure 2 show similar findings, though better visualized; Figure 2 shows that recipe size distribution is bounded as very few recipes contain very few or very many ingredients. These findings would point towards a basis of palatability and/or nutrition in human dietary choices.

Table 14

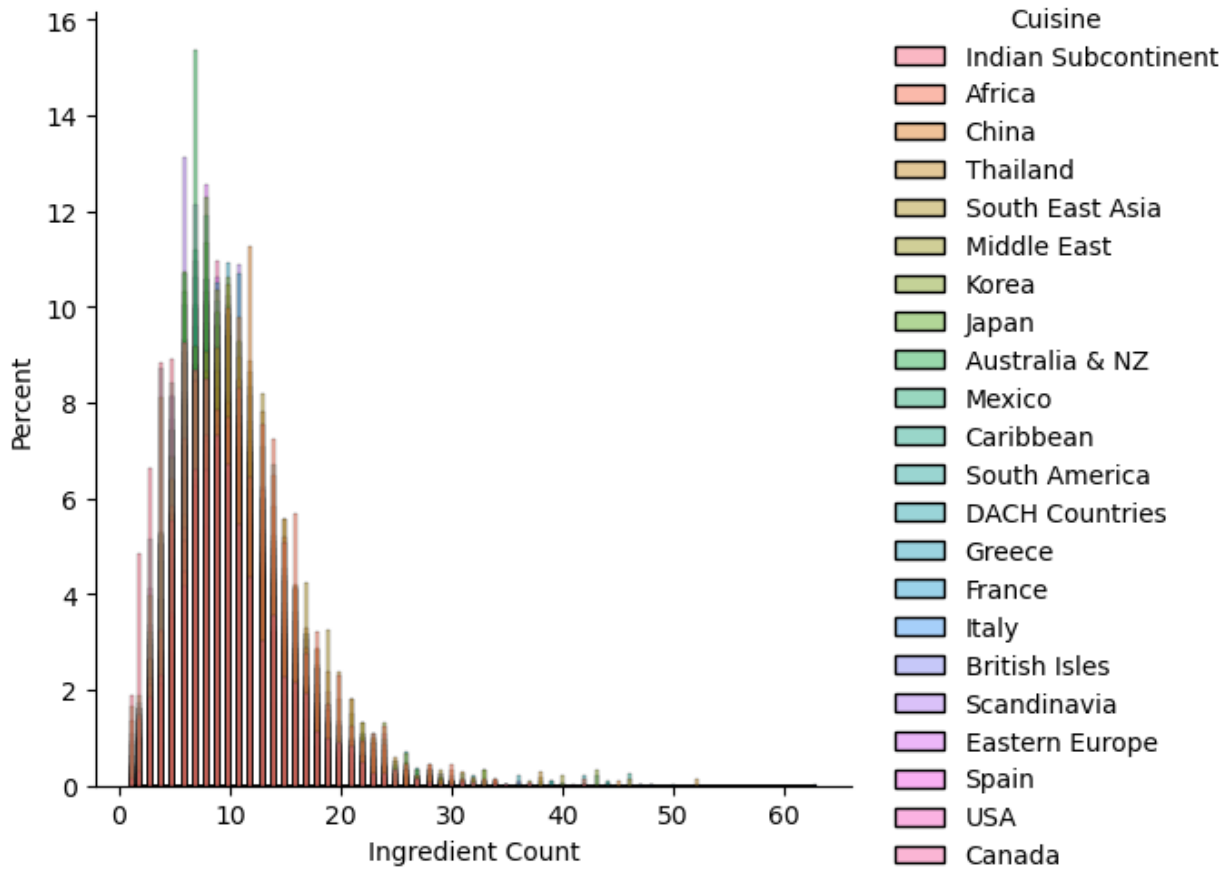
Cuisine stats table.

Cuisine	Recipe Count	Avg Ingredient Count	Tot. Ingredient Count	STD Ingredient Count
Indian Subcontinent	4057	8.5	383	5
Africa	650	12	307	5.5
China	941	11.5	299	4.9
Thailand	666	12.2	264	5.9
South East Asia	611	11.6	270	5.1
Middle East	993	10.6	313	4.9
Korea	301	10.7	199	5.5
Japan	578	9	280	4.7
Australia & NZ	494	8.8	293	3.4
Mexico	3138	9	372	4.1
Caribbean	1103	11.3	339	5.9
South America	310	9.3	220	4.6
DACH Countries	487	10.3	259	5.1
Greece	934	10.6	279	4.5
France	2701	10	418	5.4
Italy	7502	10	445	4.8
British Isles	1074	9.4	343	4.3
Scandinavia	404	9.3	247	3.8
Eastern Europe	565	10	252	4.2
Spain	815	11	312	5.3
USA	16106	10.1	595	5.4
Canada	1112	8.9	362	3.8
Average	2070.1	10.2	320.5	4.8
Median	874.5	10.05	303	4.9
SD	3551	1	87	1

Singh, N., & Bagler, G. (2018). Data-driven investigations of culinary patterns in traditional recipes across the world. 2018 IEEE 34th International Conference on Data Engineering Workshops (ICDEW), 157-162.

Figure 1

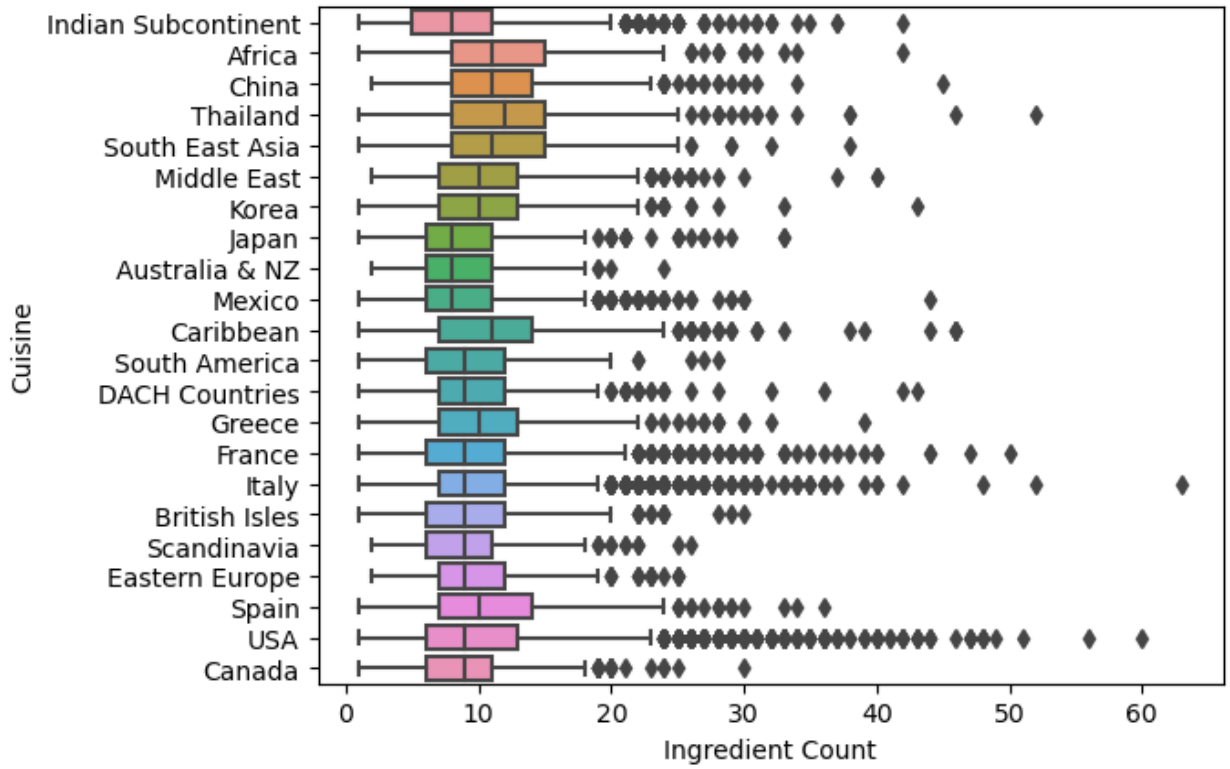
Distribution plot of the number of ingredients per recipe per cuisine.



Singh, N., & Bagler, G. (2018). Data-driven investigations of culinary patterns in traditional recipes across the world. 2018 IEEE 34th International Conference on Data Engineering Workshops (ICDEW), 157-162.

Figure 2

Box plot of the number of ingredients per recipe per cuisine.



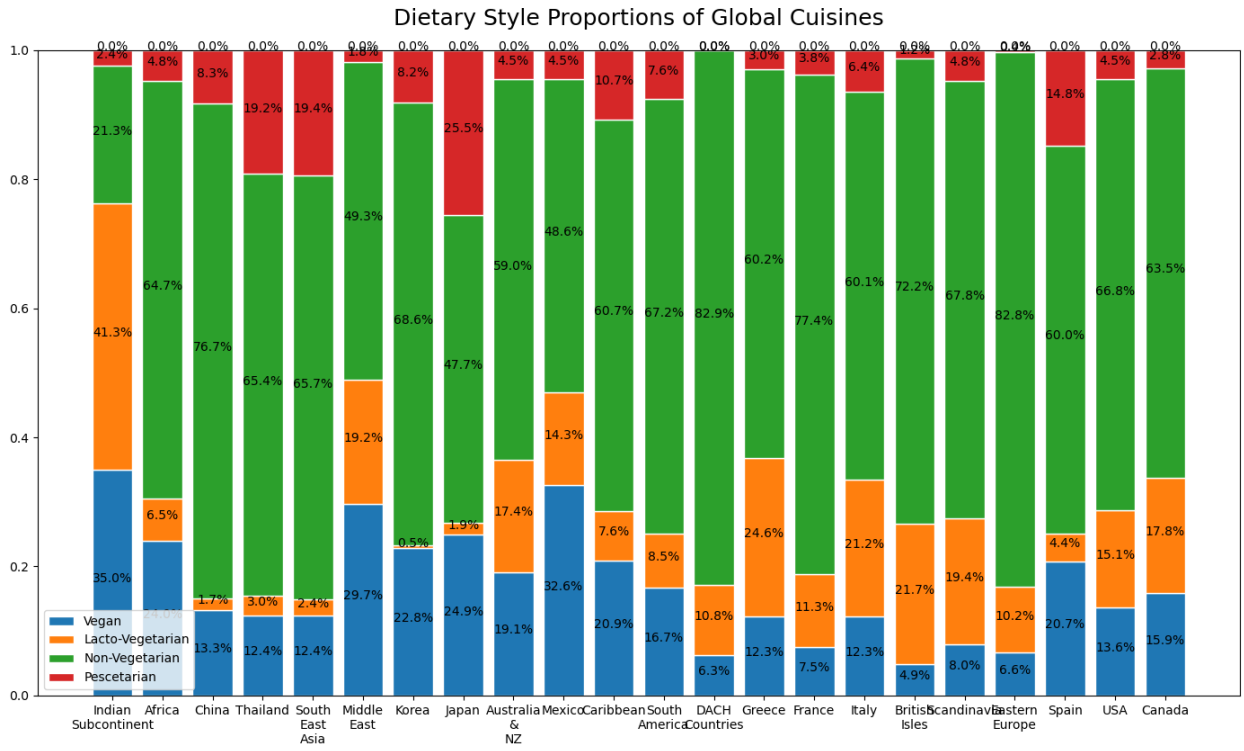
Singh, N., & Bagler, G. (2018). Data-driven investigations of culinary patterns in traditional recipes across the world. 2018 IEEE 34th International Conference on Data Engineering Workshops (ICDEW), 157-162.

Using the feature created during pre-processing to tag the recipes with dietary style, a proportional stacked bar graph shown in Figure 3 was generated to visualize the feature. The Indian Subcontinent contained by far the highest amount of vegan and vegetarian recipes; surveys cited by the USDA estimate some 40% of the Indian population as being vegetarian. (Landes, M.) This contrasted sharply with the Eastern European and DACH Countries, whose dietary styles appeared more non-vegetarian. Cuisine from Japan seemed to contain the highest proportion of pescatarian recipes, perhaps owing to the long cultural and geographic influences present in and around the island country. The presence of lacto-vegetarian recipes, in contrast to the rest of the cuisine regions, was almost non-existent in Korean cuisine. Those with Korean ancestry

are classified as a high-incidence race on the topic of genetic lactose intolerance which may explain such a correlation. (Bahk et al.)

Figure 3

Dietary Style Proportions of Global Cuisines.

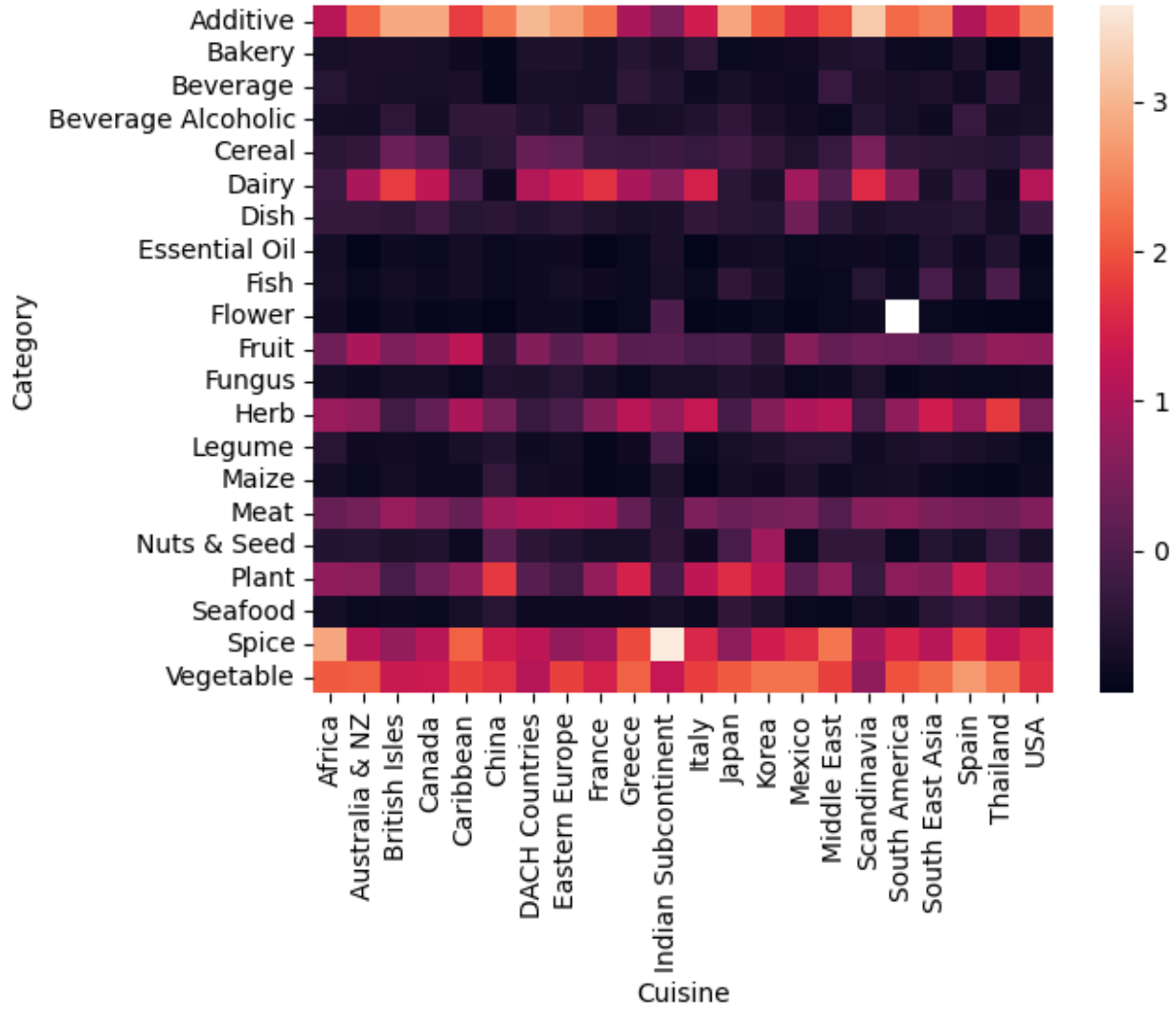


Singh, N., & Bagler, G. (2018). Data-driven investigations of culinary patterns in traditional recipes across the world. 2018 IEEE 34th International Conference on Data Engineering Workshops (ICDEW), 157-162.

Figure 4 shows a heatmap showing compositions of recipes in terms of various ingredient categories for all cuisines. From this visualization, we can clearly see the dominance of ingredients of the Spice category within the Indian Subcontinent cuisine and the Flower category within South American cuisine. Overall, Vegetable, Spice, Plant, Herb and Additive were used most frequently. France, Scandinavia, and the British Isles appear to use more Diary than Vegetable. Spice appears most popular in Indian Subcontinent, Middle East, Caribbean and Africa cuisines.

Figure 4

Heatmap showing compositions of recipes in terms of various ingredient categories for all cuisines.



Singh, N., & Bagler, G. (2018). Data-driven investigations of culinary patterns in traditional recipes across the world. 2018 IEEE 34th International Conference on Data Engineering Workshops (ICDEW), 157-162.

FlavorDB

For the FlavorDB dataset, Figure 5 shows the category distribution of the entities. As noted earlier in the data overview, some categories could be combined, and the result was the manufactured “Group” feature in the dataset. Figure 6 shows the distribution of the streamlined “Group” feature after the rules were applied.

Figure 5

Barchart of the count of FlavorDB entity categories.

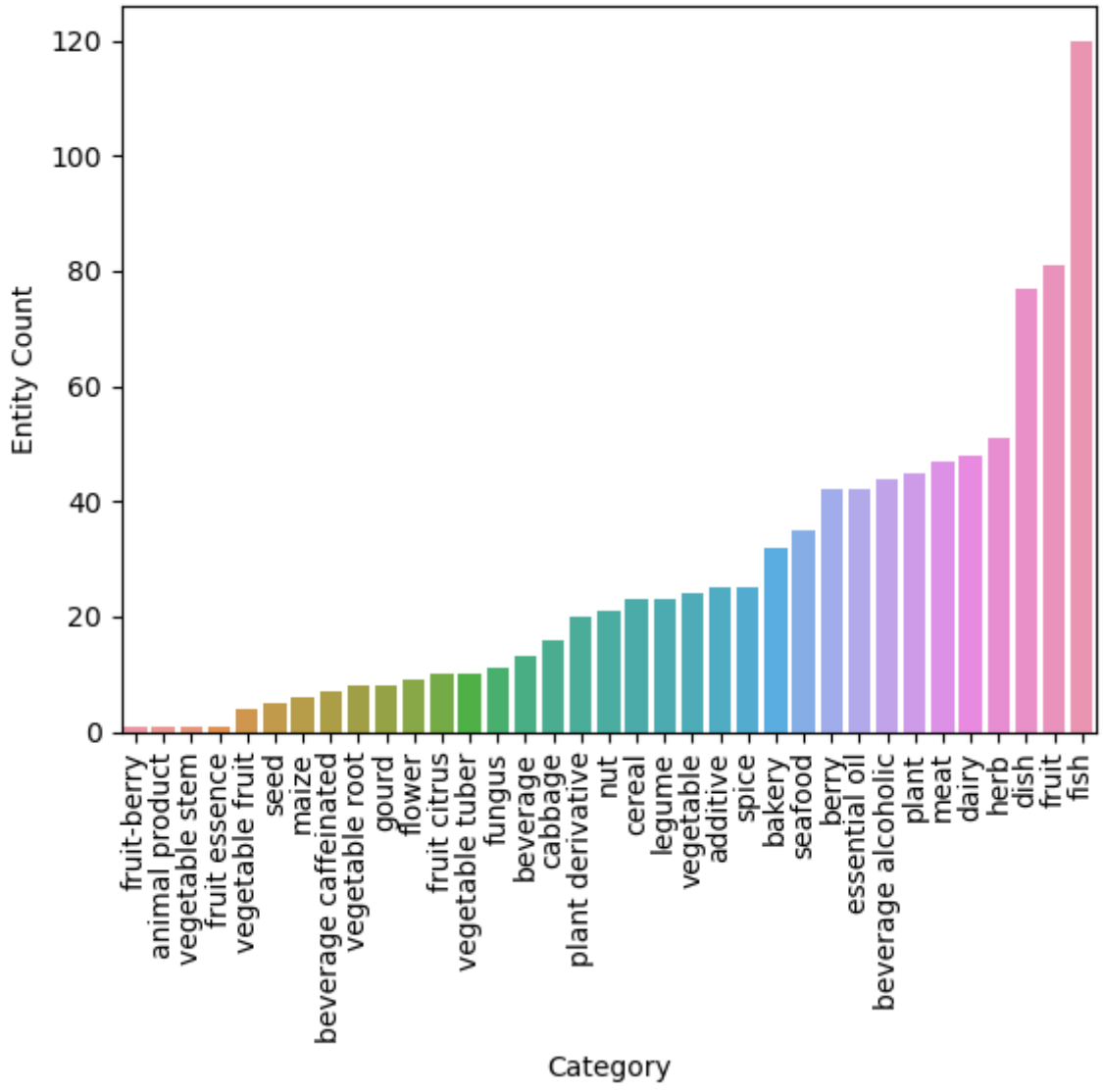
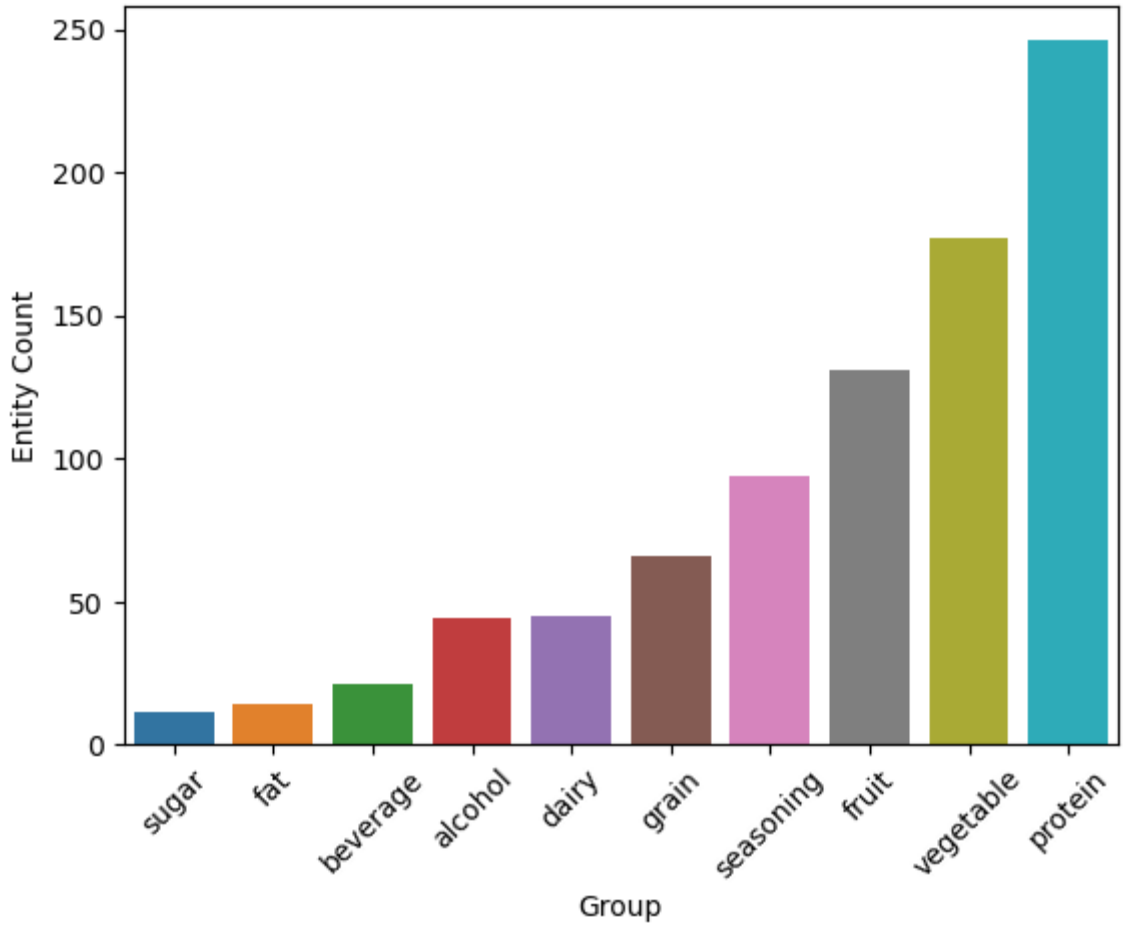


Figure 6

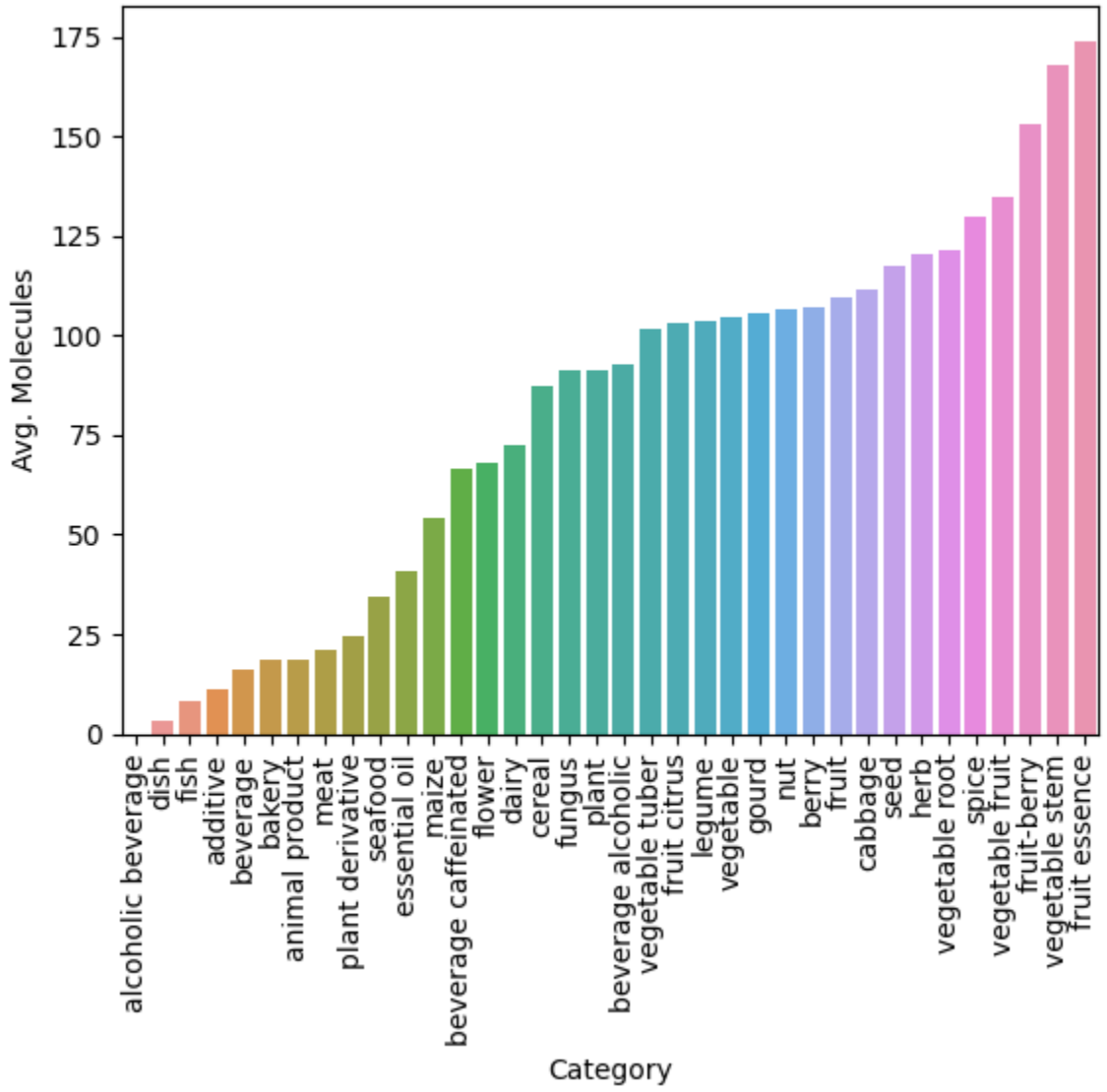
Barchart displaying the count of post-processed FlavorDB feature: 'Group'.



Protein is the dominant entity group type in the FlavorDB dataset, followed by vegetable and fruit. When examining the average number of flavour molecules within each Category and Group, seen in Figure 7 and Figure 8, respectively, one can see how vegetables and fruit provide the most flavour compounds per entity. Owing to the abundance and location of plant matter in the food chain, this seems evident. Heribores and omnivores, like humans, possess more taste buds compared to strict carnivores to better determine which among the plethora of plants in their environment are suitable for consumption. (Tyrrell, J. 2018)

Figure 7

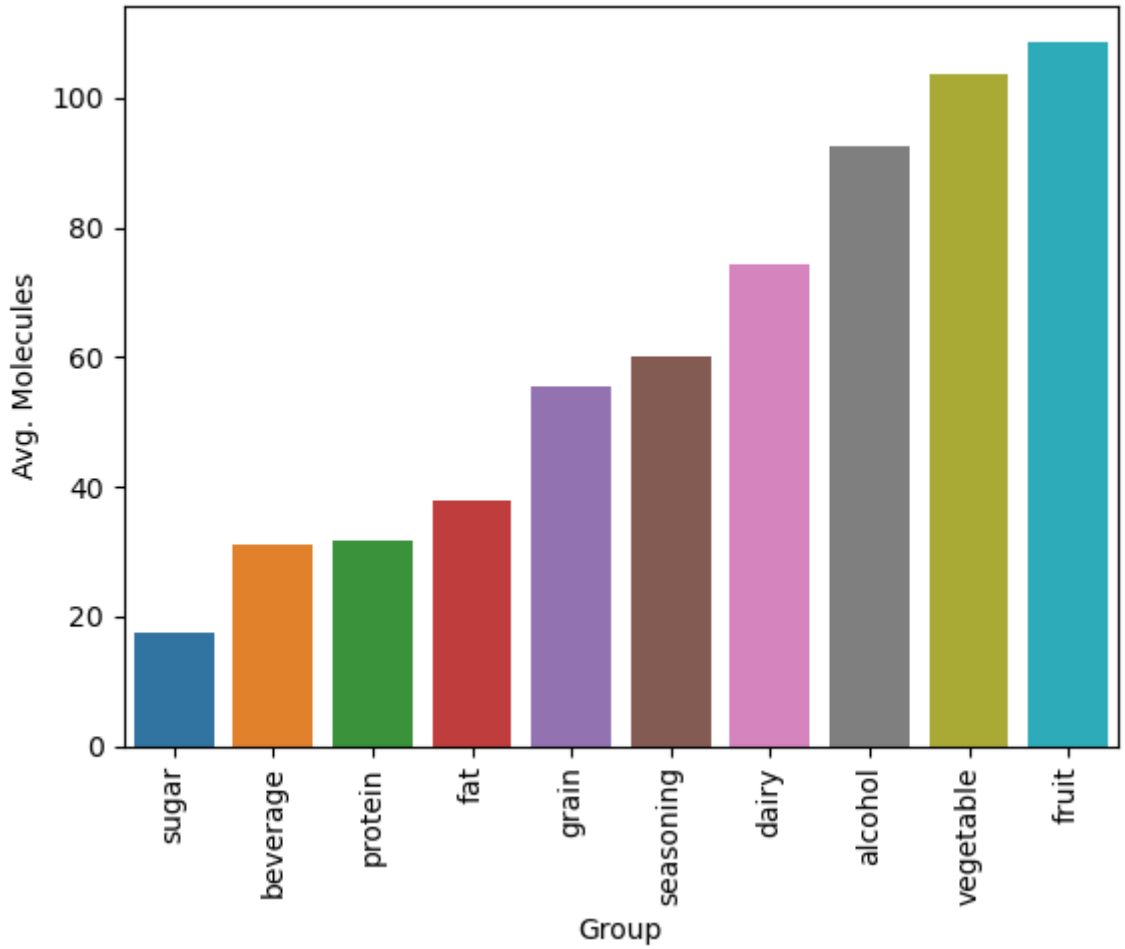
Barchart displays the average count of flavour molecules per entity within each entity category.



Varshney, K. R., Pinel, F., Varshney, L. R., Bhattacharjya, D., Schelldorfer, J., Korhonen, J. H., & Freiwald, C. (2018). FlavorDB: a database of flavor molecules. IBM Journal of Research and Development, 59(2/3), 1-14.

Figure 8

Barchart displays the average count of flavour molecules per entity within each entity group.



Varshney, K. R., Pinel, F., Varshney, L. R., Bhattacharjya, D., Schelldorfer, J., Korhonen, J. H., & Freiwald, C. (2018). FlavorDB: a database of flavor molecules. IBM Journal of Research and Development, 59(2/3), 1-14.

METHODOLOGY & EXPERIMENTS

Food Pairing

Food pairing is a concept in culinary science that assesses the affinity of ingredients based on their shared flavour molecules. According to the food pairing hypothesis, when a pair of ingredients overlap in flavour compounds, they are more likely to produce a pleasing taste when used together in a recipe. This approach to recipe creation considers the flavour profiles of its constituent ingredients, where each flavour profile embodies a unique set of volatile chemical compounds that confer the distinct taste and smell to an ingredient.

As defined by Ahn et al. (2011) and Singh, N., & Bagler, G. (2018), given a set of ingredients *I = {i1, i2, ..., in}* in a given recipe *R*, and each ingredient *i* has a set of flavour compounds *F = {f1, f2, ..., fm}*, the food pairing score for a pair of ingredients *(i, j)* can be represented as the overlap of their flavour compounds: $O(i, j) = |F_i \cap F_j|$. The food pairing score of a recipe with *n* ingredients would be the average overlap of all pairs of ingredients, written as:

$$N_s^R = \frac{2}{n(n-1)} \sum_{i,j \in R, i \neq j} |F_i \cap F_j|$$

Quantifying the overall food pairing ($N(S|C)$) of a cuisine *C* involves calculating the average food pairing score of its constituent recipes, written as:

$$N_s^C = \frac{1}{n_r} \sum_R N_s^R$$

This numerical measure is then compared to the score from a cuisine created with randomly generated recipes using shuffled ingredients $N(S|rand)$ to assess the bias in food pairing. The degree of deviation of a cuisine's food pairing score from this random cuisine indicates whether the cuisine demonstrates a positive or negative food pairing. A positive value indicates a trend for the real cuisine to exhibit positive food pairing,

while a negative value indicates the opposite. The greater the deviation, the stronger the positive or negative bias in food pairing. Formally, this statistic can be computed as:

$$\Delta N_s^C = N_s^C - N_s^{rand}$$

To further test what factors influenced the food pairing trend in a given cuisine, four types of random control models were created, with each set of randomly generated recipes following the size distribution of the real cuisine:

- Uniform Control Model
    - Ingredients used to create the random cuisine were selected uniformly from the set of ingredients used in the real cuisine.
- Ingredient Frequency Model
    - The frequencies of ingredients used in the real cuisine were preserved.
- Category Frequency Model
    - The categories of ingredients used in the real cuisine were preserved.
- Group Frequency Model
    - The groups of ingredients used in the real cuisine were preserved.

Each random control model had the number of recipes of its real cuisine generated. To determine which model most closely contributed to the cuisine's delta food pairing score, a Z-score is used:

$$Z = \sqrt{n_r^{rand}} \frac{N_s^C - N_s^{rand}}{\sigma_{rand}}$$

Where *nr rand* denotes the number of recipes in the random cuisine and *sigma rand* denotes the standard deviation of the food pairing scores of the random cuisine.

Food Bridging

Food bridging is a concept in culinary science which states that ingredients that do not share or have fewer flavour compounds in common can still combine well in a recipe if there's a "bridging" ingredient(s) that shares flavour compounds with both. This is contrary to the food pairing hypothesis, which posits that ingredients with shared flavour compounds are more likely to taste good together.

To determine the food bridging score of a cuisine or between any of its constituent ingredients, a flavour network must first be created. In this created graph, ingredients are the nodes, and the edges between them are the shared flavour compounds. Normally, the network weights lie on a non-normalized interval of Zij ∈ [*a, b*] ⊂ R, however, normalized weights of the edges can be computed as initially shown in Simas and Rocha (2015) and then modified by Simas et al. (2017) with:

$$w_{ij} = \frac{(1-2\epsilon)Z_{ij} + (2\epsilon - 1)\cdot MIN(Z_{ij})}{MAX(Z_{ij}) - MIN(Z_{ij})} + \epsilon$$

Where ε is used to avoid merging or isolating nodes with weights at the boundaries of Zij ∈ [a, b]. ε is normally set to 0.01.

Once the flavour network is created, analysis of the metric and semi-metric paths between ingredients can be performed. An edge in a weighted graph is considered metric if there is a direct connection between two nodes. If no direct path exists and there is at least one alternate path that involves other nodes, it is considered semi-metric. The extent of food bridging in a recipe is, therefore, the average of all semi-metric paths in a recipe. In further referencing Simas and Rocha (2015) as well as Simas and Suckling (2016), the graph semi-metric percentage of a cuisine is given as:

$$SMP = \frac{\sum_{i,j} \delta(s_{i,j} > 1 \wedge s_{i,j} < +\infty)}{\sum_{i,j} \delta(s_{i,j} \geq 1 \wedge s_{i,j} < +\infty)}$$

Where δ is the *discrete-Kronecker* function, i.e. δ(condition) = 1 if logically true and δ(condition) = 0 if logically false, and *s(i,j)* is the semi-metric ratio between ingredients *i* and *j* in the flavour graph. The numerator of this formula counts all semi-metric edges, and the denominator counts all edges.

Similar to the approach taken in the Food Pairing experiments, the numerical measure is then compared to the score from a cuisine created with randomly generated recipes using shuffled ingredients SMP(S|rand) to assess the bias in food bridging. Additionally, the same four random models were constructed and had z-scores computed to measure the deviation of the food bridging scores of the random cuisines.

Frequent Pattern Mining

Frequent pattern mining (or frequent itemset mining) is a technique used to uncover patterns in data sets. The aim is to discover sets of items (in this case, ingredients) that frequently appear together in the data. These commonly co-occurring items are referred to as 'frequent itemsets' or 'frequent patterns'. In the context of recipes, frequent pattern mining can be used to uncover ingredient combinations that often appear together. A recipe is an unordered mix of ingredients, which is ideal for this methodology. Each recipe is treated as a transaction, and each ingredient as an item within that transaction. By applying frequent pattern mining algorithms to our collection of recipes, one can identify common ingredient combinations. Several algorithms can be used for frequent pattern mining, including the Apriori algorithm, Eclat algorithm, and FP-Growth algorithm. Due to its speed, an implementation of the FP-Growth Algorithm can be utilized for this purpose, as it's an effective and scalable approach for mining frequent patterns through the growth of pattern fragments. A minimum support level of 0.05 is used to ensure the extraction of patterns across a substantial number of recipes.

Once frequent itemsets are generated, the respective MSC and SMP scores of the tuples can be measured, and some comparisons can be made against equivalent-sized generated orders of tuples within each cuisine to gauge the significance of the calculated scores. Combinations of tuples can be created using:

$$C = \{c \mid c \subseteq r, \forall\ r \in R, |c| = set\_size\}$$

Where *R* is the set of recipes, where each recipe *r* is a set of ingredient IDs for the real cuisine. *R* is taken as an input along with a *set_size* parameter, which specifies the size of the tuples in the combinations. The output is a set of combinations *C* of size *set_size* that contain non-repeating ingredient IDs, where each combination *c* is a set of

ingredient IDs. In this representation, $c \subseteq r$ denotes that combination *c* is a subset of recipe *r*, and |*c*| denotes the cardinality (size) of combination *c*, which is equal to *set_size*. Once tuples are generated for set sizes of 3 to 6, an average food pairing score and their standard deviation can be computed. Due to the extremely large number of combinations (upwards of 10 million combinations for groups of 6 ingredients) and the computation requirements to compute SMP scores through semi-metric analysis, comparisons will not be able to be performed for food bridging analysis, but food pairing analysis is feasible. Once the tuples have been identified and enriched with quantitative data about the set, correlation analysis can be performed against the features to uncover possible alignments. Spearman's correlation will likely be the best approach since data is likely not to be distributed among the set and is better suited for the monotonic relationships between the food pairing/bridging scores.

To further probe the results of the correlation analysis, an ingredient tuple's influence on a cuisine's pattern pairing/bridging direction can be tested using a classification task. Binary classification can be undergone to gauge whether the tuple's MSC and SMP scores could be used to predict whether the parent cuisine scored positive or negative in food pairing and food bridging. A decision tree classifier is a good choice for the parameters of this experiment, chosen for its simplicity and interpretability, allowing us to see the importance of different features in making predictions. To address the diminishing number of tuples as they ascend in size and to address the unbalanced number of positive food-pairing cuisines over negative ones, for each N-tuple, a balanced dataset can be created by undersampling the majority class to match the minority class size. Stratified k-fold cross-validation on the balanced dataset can then be performed, with a holdout of 20% for the test data set. For each iteration, a decision tree classifier can be trained, run against the held-out test set, then have the test's precision, recall, and F1-score calculated.

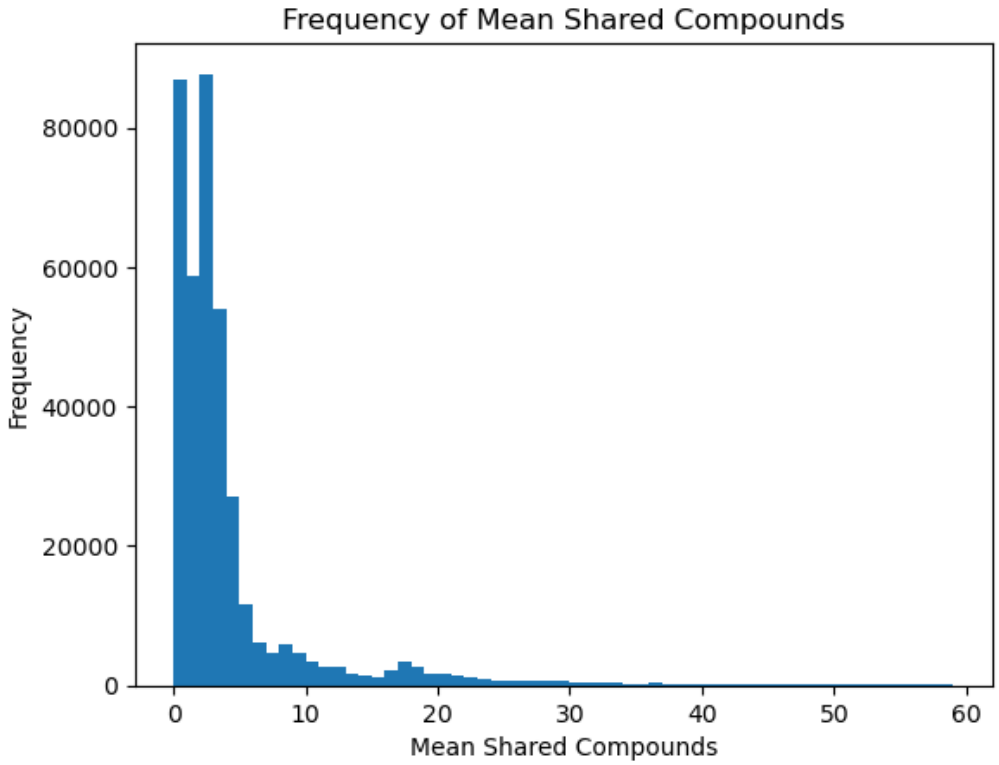
RESULTS

Food Pairing

A preliminary analysis was undertaken by computing the food pairing score (which is used interchangeably with mean shared compounds) between every ingredient pair available across all cuisines. An average of 18 flavour molecules were found to be shared, with a median value of 2. When plotted, Figure 1 appears to follow a power law distribution which shows that a few ingredients share a large number of flavour compounds with other ingredients (these are the very popular ingredients used across many dishes), and many ingredients share a small number of flavour compounds with other ingredients (these are less common, more unique ingredients).

Figure 9

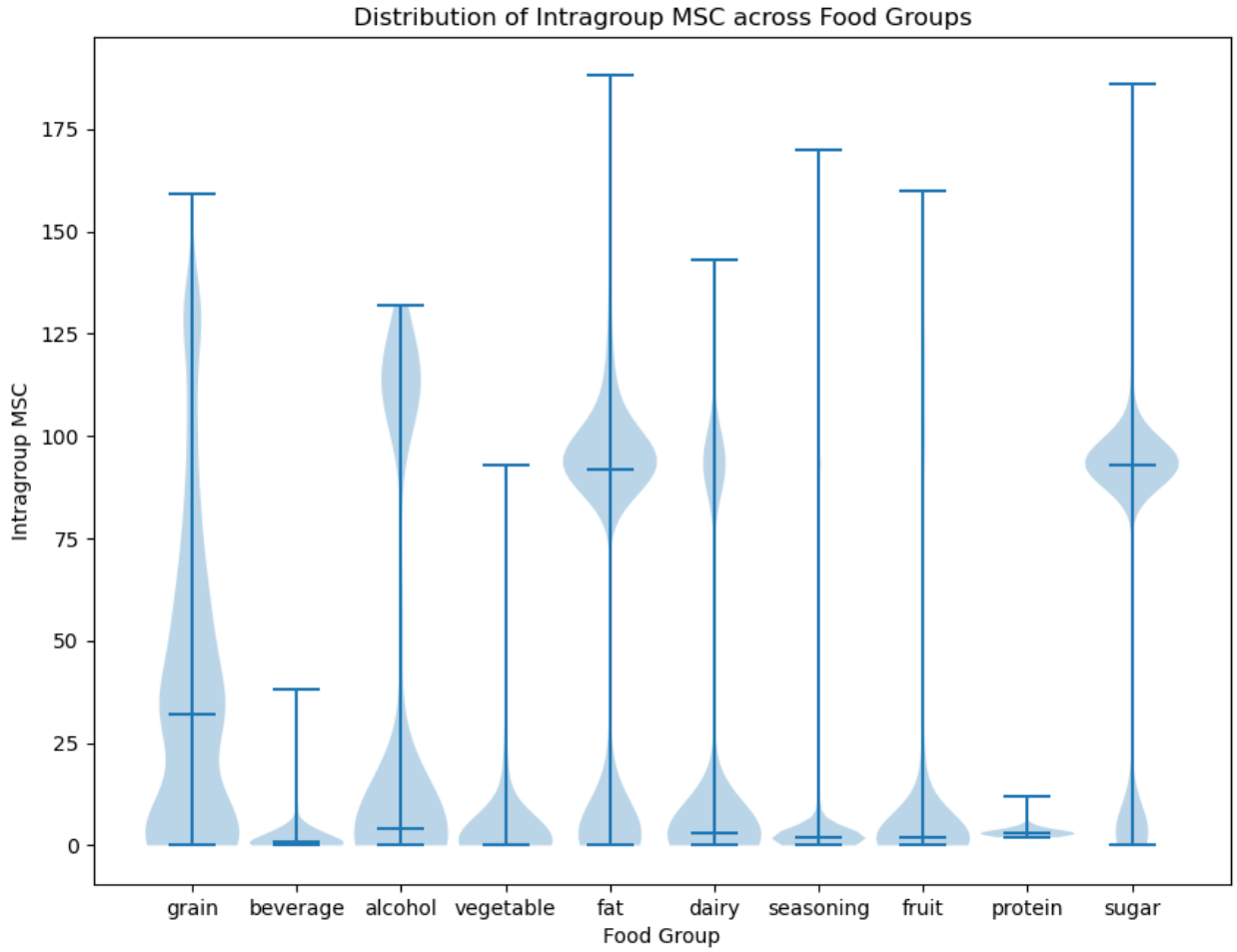
Distribution plot of the number of the frequency of mean shared compounds across the superset of ingredients.



By transforming the visualization, one can see how similar ingredients taste similar within their groups. Figure 2 illustrates how small groups of grains, alcohols, fats, dairy, and sugars taste similar within their respective categories.

Figure 10

Violin plot showing the distributions of intragroup mean shared compound scores. The density curves help visualize denser pockets of shared scores among each group.



The mean shared compound score represents the flavour pairing, enumerating the extent of overlap in flavour profiles within ingredients with a cuisine's recipes. When computed against each cuisine and compared to that of a cuisine built from randomly generated recipes based on the size distribution and ingredients of the real cuisine, one can note a characterization of a bias towards a uniform blend of flavours, contrasting

blends of flavours, or being indistinguishable from its random counterpart. Figure 3 shows that 17 cuisines showed a bias toward uniform (or positive) food pairing, which is consistent with the hypothesis proposed by Chef Heston Blumenthal (2008). These cuisines include China, South East Asia, Korea, Canada, France, South America, Thailand, Australia & NZ, Mexico, Middle East, USA, Greece, Caribbean, Spain, Italy, Indian Subcontinent, and Africa. The remaining 5 regions showed a bias toward negative food pairing, which includes cuisines Scandinavia, Japan, DACH Countries, Eastern Europe and the British Isles. These results are contrary to previous observations conducted by Ahn et al. (2011), Jain et al. (2017) and Jain et al. (2015) but similar to the more recently published work by Singh, N., & Bagler, G. (2018) since similar datasets of recipes and flavours were sourced.

Four types of random control models were created: uniform, frequency preserving, category preserving, and group preserving, to assess their contribution to the observed biases. Z-scores were then calculated between the actual food pairing score of the cuisine and their random model's counterpoint, as shown in Figure 4 and Table 1. From averaging the z-scores in (Table 2 and Table 3), It was found that the ingredient frequency model was by far the best indicator of both positive and negative food pairing biases within cuisines, with category composition or group composition less so. This signifies that specific ingredient popularities within each cuisine drive the unique flavours associated with each, intensified by the compounds shared with other ingredients.

When plotted against the average recipe size of the cuisine, as shown in Figure 5, there appeared to be no real correlation between the average food pairing score and the average recipe size.

Figure 11

Scatter plot showing the mean shared compound scores of each cuisine compared to its counterpart of randomly generated recipes and the difference between them.

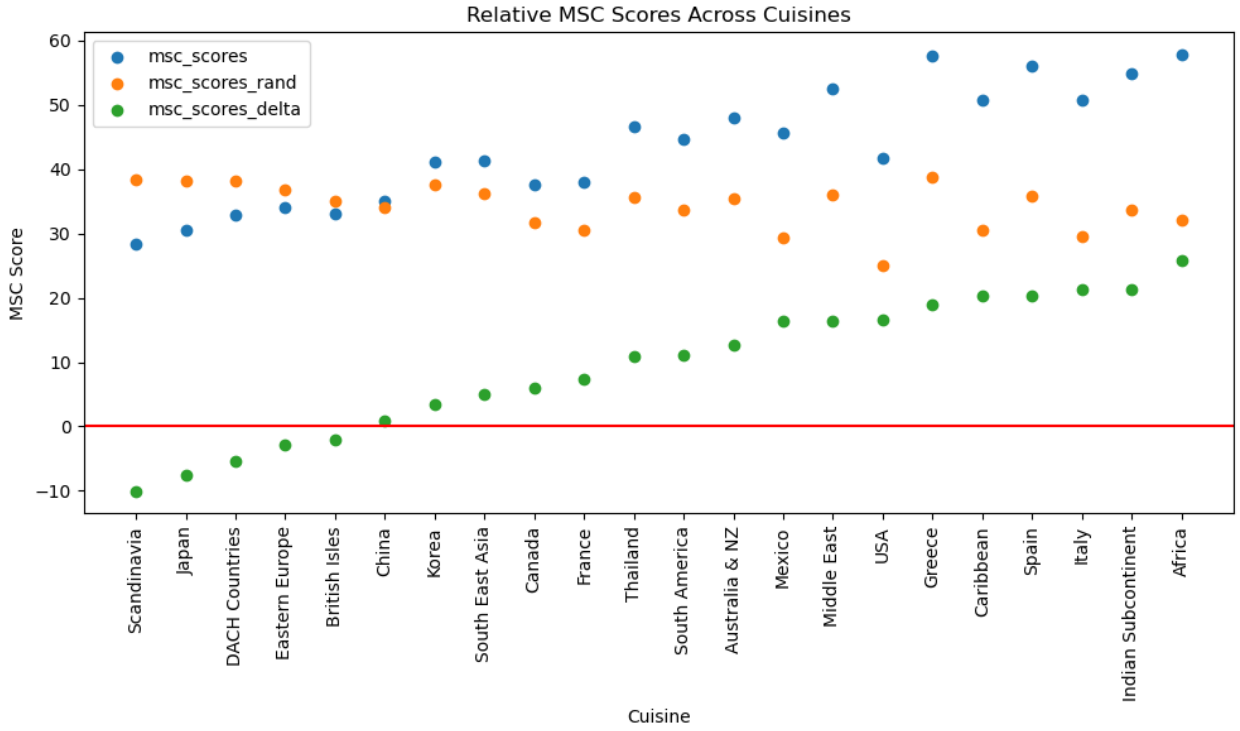


Table 15

Table of the mean shared compound scores of each cuisine, their respective random counterpart, and the delta between the two.

Cuisine	msc_scores	msc_scores_rand	msc_scores_delta
Scandinavia	28.309	38.426	-10.117
Japan	30.565	38.097	-7.532
DACH Countries	32.845	38.222	-5.378
Eastern Europe	34.053	36.850	-2.796
British Isles	33.052	35.099	-2.047
China	35.028	34.084	0.944
Korea	41.083	37.584	3.499
South East Asia	41.259	36.234	5.025
Canada	37.616	31.687	5.929
France	37.907	30.592	7.315
Thailand	46.592	35.662	10.930
South America	44.664	33.662	11.002
Australia & NZ	47.961	35.380	12.581
Mexico	45.634	29.287	16.348
Middle East	52.464	35.993	16.471
USA	41.659	25.019	16.640
Greece	57.680	38.662	19.018
Caribbean	50.815	30.564	20.251
Spain	55.978	35.720	20.258
Italy	50.716	29.501	21.215
Indian Subcontinent	54.874	33.655	21.219
Africa	57.860	32.030	25.831

Figure 12

Scatter plot of the z-scores of the random control cuisines.

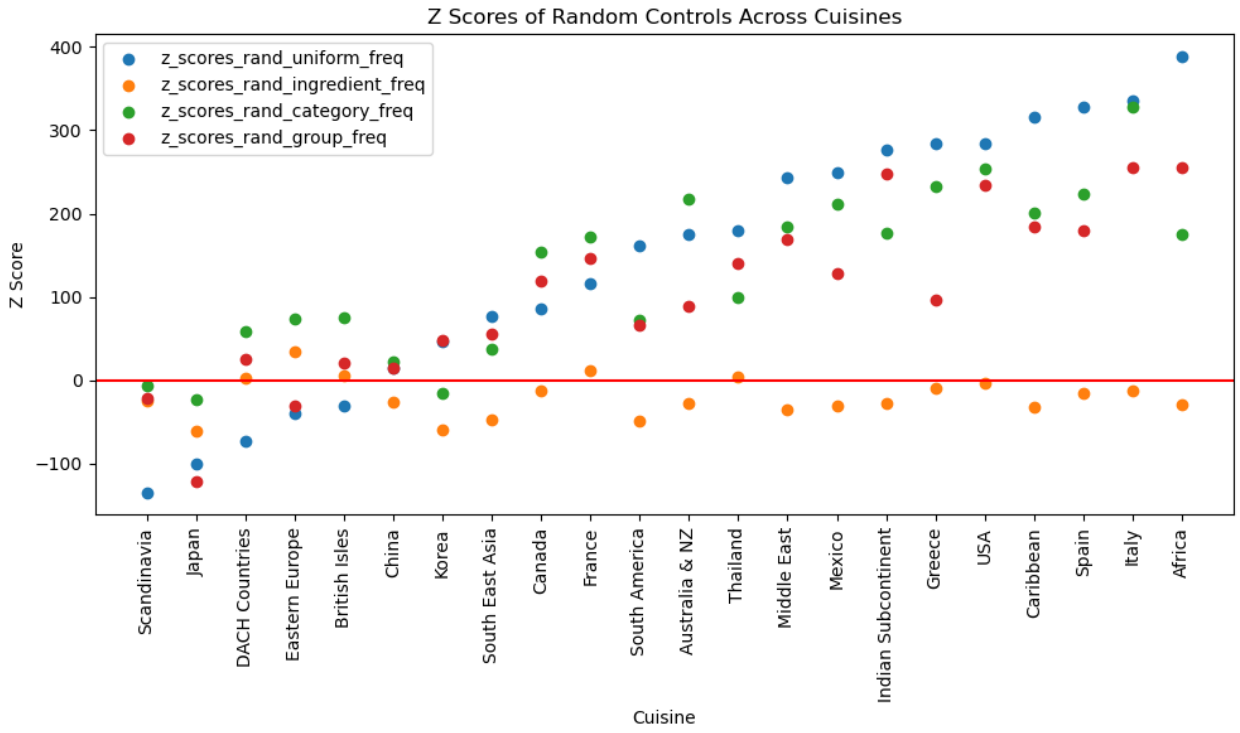


Table 16

Table of the mean shared compound z-scores of the random control versions of each cuisine.

Cuisine	z_scores_rand_uniform_freq	z_scores_rand_ingredient_freq	z_scores_rand_category_freq	z_scores_rand_group_freq
Scandinavia	-134.811	-24.312	-5.573	-21.774
Japan	-99.717	-61.292	-22.672	-121.674
DACH Countries	-72.917	2.965	58.294	25.782
Eastern Europe	-39.862	34.312	73.620	-30.287
British Isles	-29.982	6.410	74.707	20.245
China	14.711	-26.284	23.083	15.600
Korea	47.202	-58.938	-15.821	47.531
South East Asia	76.184	-46.969	37.597	55.166
Canada	86.089	-11.778	154.216	119.285
France	116.591	12.601	172.817	146.177

South America	161.909	-49.094	71.848	65.521
Australia & NZ	175.525	-27.056	216.902	89.663
Thailand	180.394	4.635	98.914	140.094
Middle East	243.508	-35.691	184.764	169.533
Mexico	249.619	-29.906	211.556	128.744
Indian Subcontinent	276.819	-27.181	176.694	246.967
Greece	284.134	-9.450	232.490	96.705
USA	284.450	-2.567	254.166	234.572
Caribbean	315.637	-31.826	200.211	184.552
Spain	328.279	-15.600	224.233	180.228
Italy	334.767	-12.130	327.749	255.291
Africa	388.771	-28.755	175.791	255.208

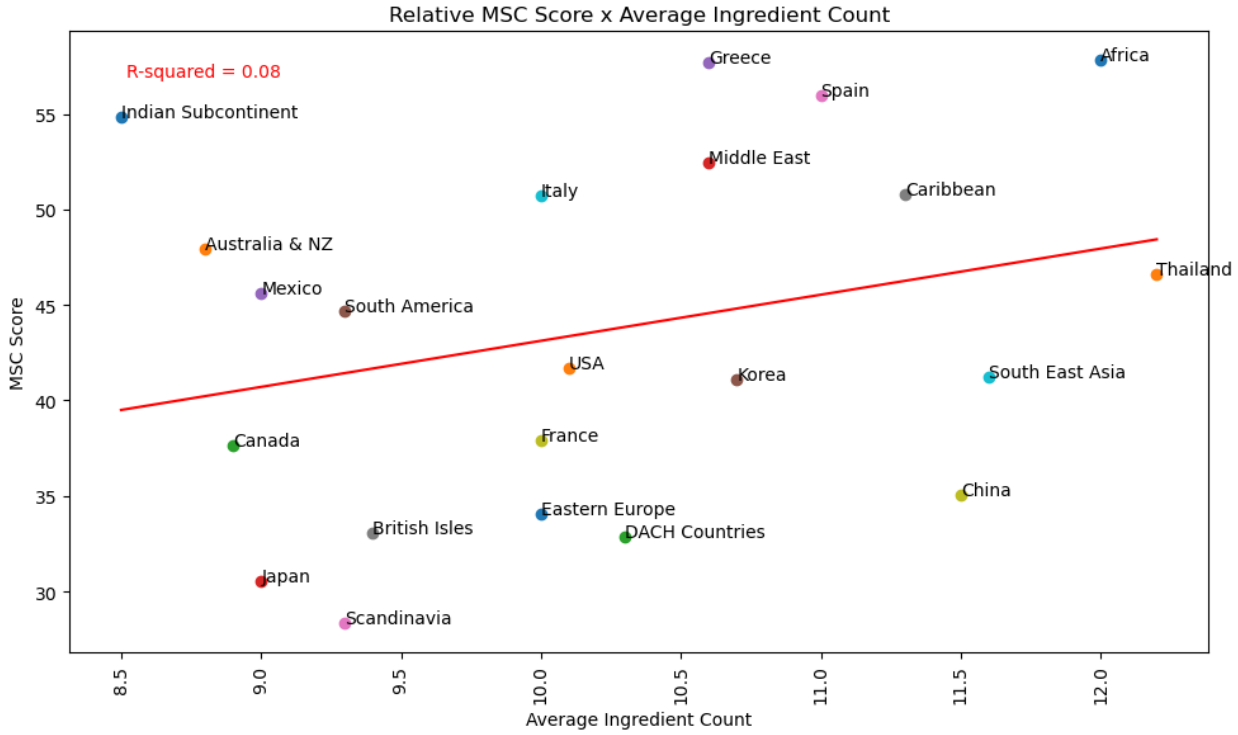
Table 17

Mean z-scores of each random control cuisine set. Random ingredient frequency control cuisine showed the closest alignment with the real cuisine food pairing score by a significant margin.

Set	**Mean z-score**
z_scores_rand_uniform_freq	3941.87
z_scores_rand_ingredient_freq	559.75
z_scores_rand_category_freq	3013.71
z_scores_rand_group_freq	2650.59

Figure 13

The average cuisine mean shared compound score and respective average recipe ingredient count are plotted. R-squared value of a regression function is very weak and suggests little to no correlation.



Food Bridging

The semi-metric percentage score represents the degree of flavour bridging, enumerating the extent of alternative flavour paths between ingredients in a cuisine's recipes. Values ranged from 57% (Africa) on the low end to 71% (Scandinavia) on the high end. When computed against each cuisine and compared to that of a cuisine built from randomly generated recipes based on the size distribution and ingredients of the real cuisine, as shown in Figure 7, there appears a split between cuisines with a bias towards a greater, lesser or equivalent degree compared to their random counterpart. Africa, Caribbean, South East Asia, China, Australia & NZ, Spain, Canada, Italy, Korea and USA exhibited lower scores, Mexico exhibited an equivalent score, and France, South America, Thailand, Indian Subcontinent, Middle East, Greece, Japan, Eastern Europe, British Isles, DACH Countres and Scandinavia exhibited a greater percentage than their random counterpart.

Repeated from the flavour pairing investigation, four types of random control models were created: uniform, frequency preserving, category preserving, and group preserving, to assess their contribution to the observed biases. Z-scores were then calculated between the actual food pairing score of the cuisine and their random model's counterpoint, as shown in Figure 8 and Table 5. From averaging the z-scores in Table 6, it was found that the category-preserving random model aligned closest with the real-world cuisine SMP score.

When SMP scores are plotted against the average ingredient count per recipe, a medium strength correlation appears between the two, whereby as the average recipe size increases, the SMP score decreases. (r-squared value of 0.58 as shown in Figure 9) This trend seems counterintuitive, as one would assume that as the average number of ingredients increases, the opportunities for semi-metric paths between ingredients would also increase, given the greater pool of flavours available in a recipe.

Figure 14

Scatter plot of the calculated relative average semi-metric path scores across cuisines.

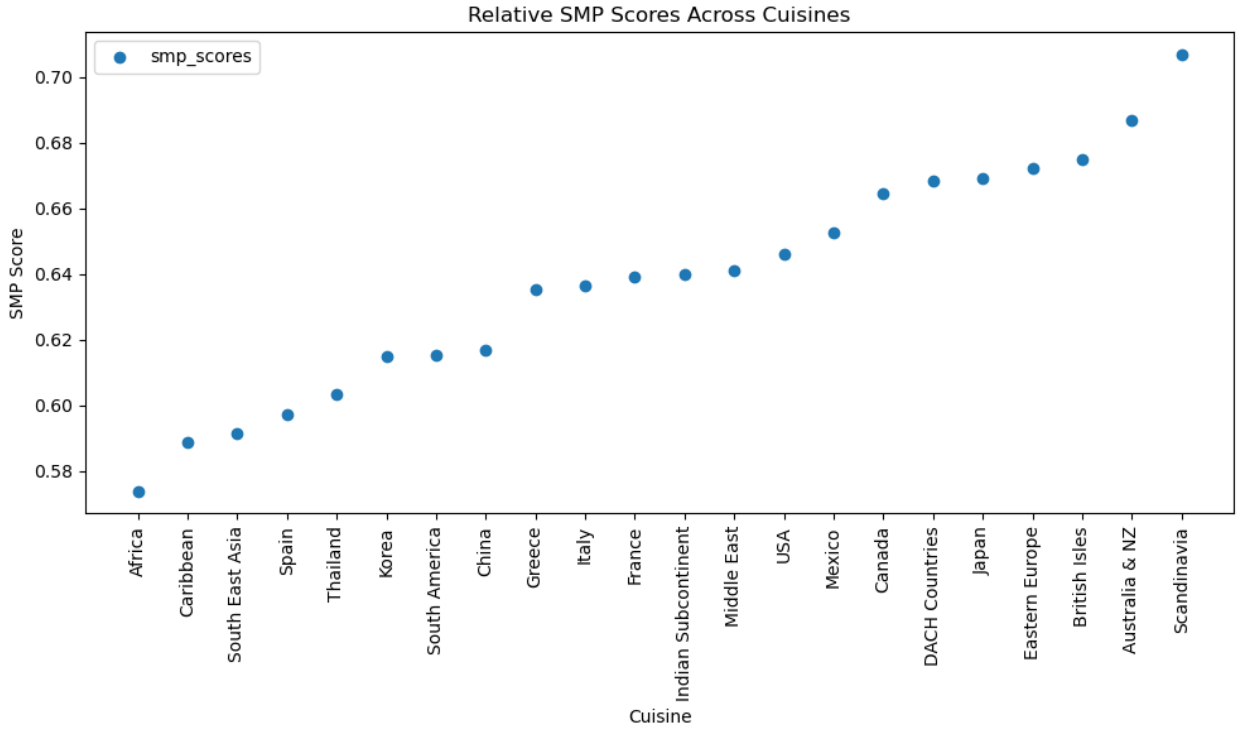


Figure 15

Scatter plot of the average semi-metric path scores of each cuisine compared to its counterpart of randomly generated recipes and the difference between them.

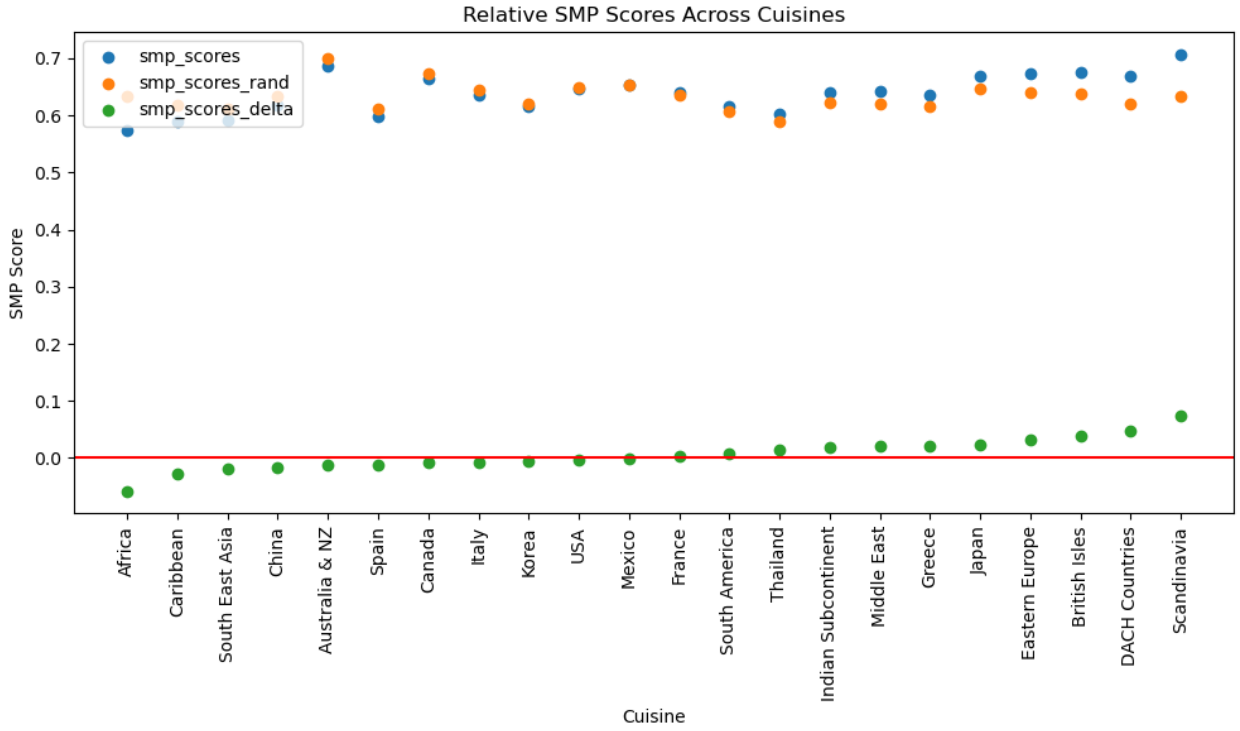


Table 18

Shows the average semi-metric path scores of each cuisine compared to its counterpart of randomly generated recipes and the difference between them.

Cuisine	smp_scores	smp_scores_rand	smp_scores_delta
Africa	57.40%	63.20%	-5.80%
Caribbean	58.90%	61.70%	-2.80%
South East Asia	59.10%	61.00%	-1.90%
China	61.70%	63.40%	-1.70%
Australia & NZ	68.70%	70.00%	-1.30%
Spain	59.70%	61.00%	-1.30%
Canada	66.50%	67.30%	-0.80%
Italy	63.70%	64.40%	-0.70%
Korea	61.50%	62.00%	-0.50%
USA	64.60%	64.90%	-0.30%
Mexico	65.30%	65.30%	0.00%
France	63.90%	63.60%	0.30%
South America	61.50%	60.70%	0.80%
Thailand	60.30%	58.80%	1.50%
Indian Subcontinent	64.00%	62.20%	1.80%
Middle East	64.10%	62.10%	2.00%
Greece	63.50%	61.50%	2.00%
Japan	66.90%	64.60%	2.30%
Eastern Europe	67.20%	64.10%	3.10%
British Isles	67.50%	63.70%	3.80%
DACH Countries	66.80%	62.10%	4.70%
Scandinavia	70.70%	63.40%	7.30%

Figure 16

Scatter plot of the average semi-metric percentage z-scores of the random control cuisines.

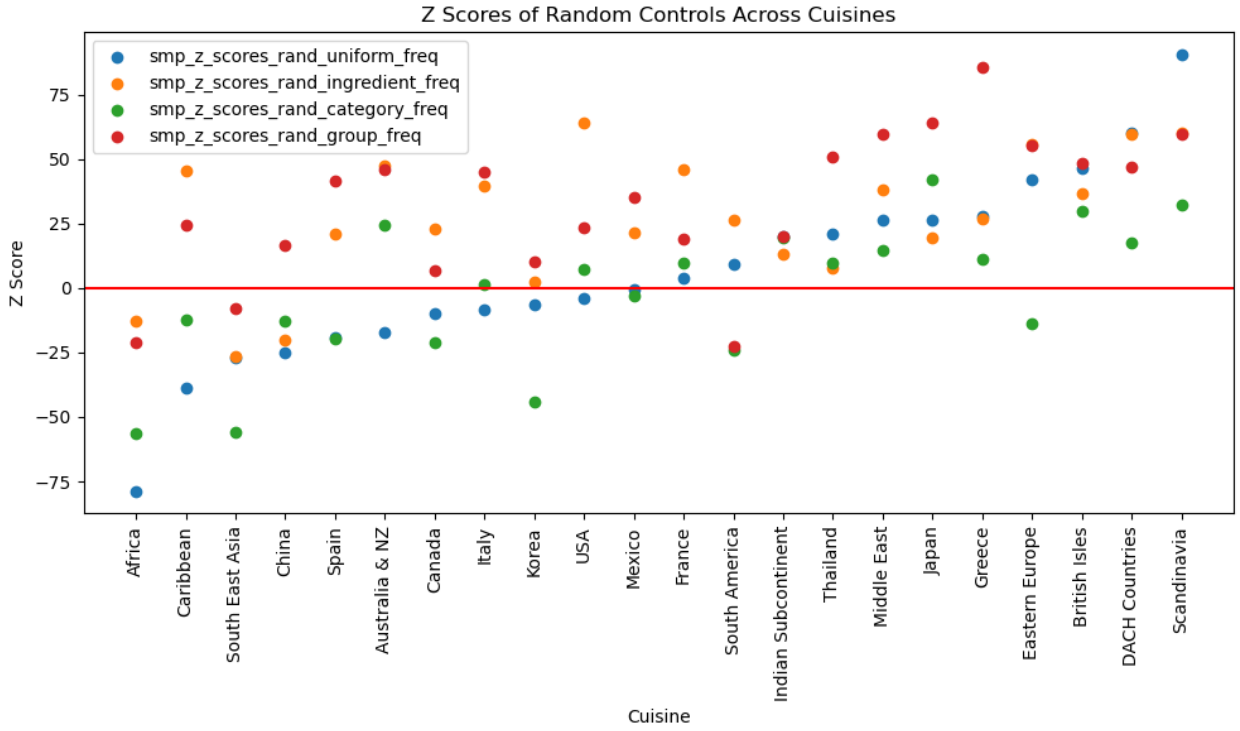


Table 19

Table of the semi-metric percentage z-scores of the random control versions of each cuisine.

Cuisine	z_scores_rand_uniform_freq	z_scores_rand_ingredient_freq	z_scores_rand_category_freq	z_scores_rand_group_freq
Africa	-78.899	-12.710	-56.563	-20.893
Caribbean	-38.893	45.534	-12.488	24.242
South East Asia	-27.005	-26.731	-55.854	-7.747
China	-25.131	-19.958	-12.798	16.497
Spain	-19.180	21.080	-19.881	41.497
Australia & NZ	-17.444	47.618	24.572	46.054
Canada	-9.820	23.079	-21.022	6.988
Italy	-8.465	39.440	1.533	44.942
Korea	-6.462	2.578	-44.229	10.182

USA	-4.129	64.278	7.435	23.534
Mexico	-0.557	21.471	-3.084	35.013
France	3.859	46.131	9.513	19.170
South America	9.374	26.281	-24.259	-22.747
Indian Subcontinent	19.986	13.236	19.513	19.763
Thailand	21.146	7.520	9.917	50.928
Middle East	26.197	38.295	14.410	59.571
Japan	26.557	19.411	42.024	64.046
Greece	28.050	26.791	11.181	85.352
Eastern Europe	41.901	55.764	-13.790	55.242
British Isles	46.372	36.412	29.876	48.335
DACH Countries	60.298	59.777	17.643	47.137
Scandinavia	90.656	60.188	32.353	59.680

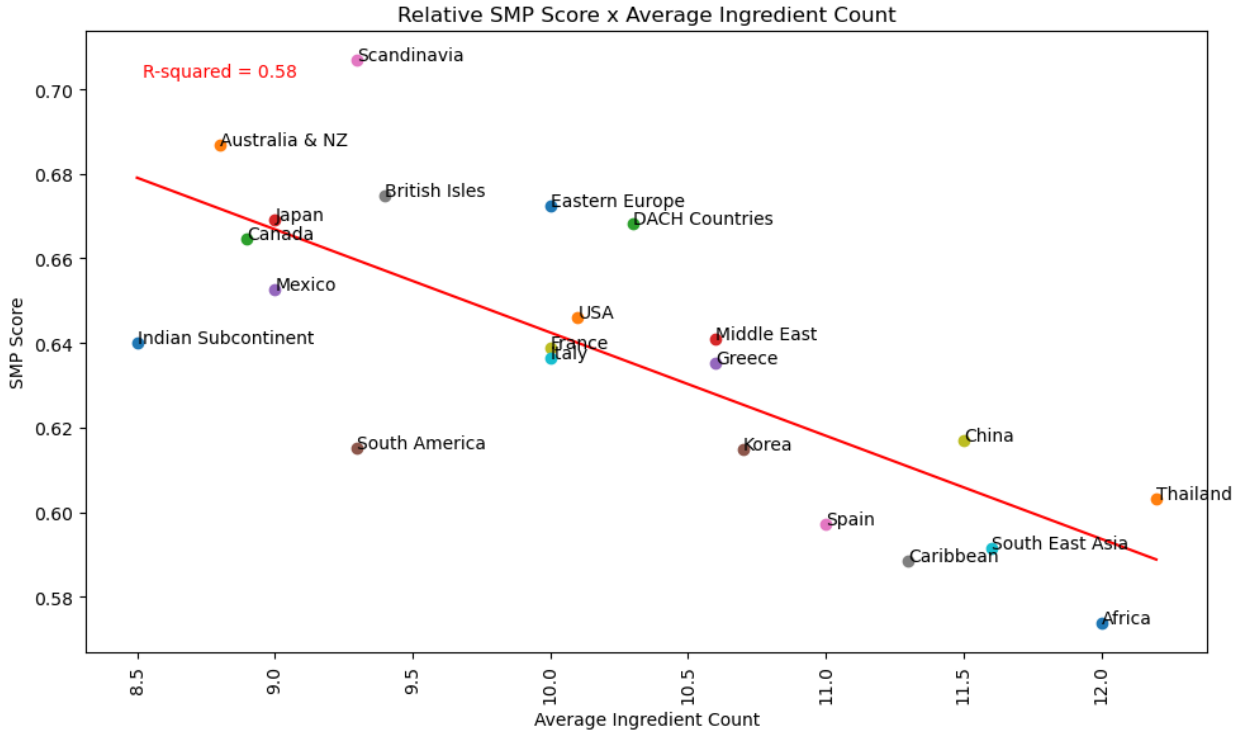
Table 20

Mean z-scores of each random control cuisine set. Random category frequency control cuisine showed the closest alignment with the real cuisine food pairing score.

Set	**Mean z-score**
z_scores_rand_uniform_freq	610.38
z_scores_rand_ingredient_freq	714.28
z_scores_rand_category_freq	483.93
z_scores_rand_group_freq	809.55

Figure 17

The average cuisine semi-metric percentage score and respective average recipe ingredient count are plotted. The R-squared value of a regression function is listed as 0.58, a moderate relationship.



Food Pairing and Food Bridging

In Figure 10, the relationship between cuisine SMP and MSC is plotted, providing a glance arrangement of scores between respective cuisines. At the extremes of the scale, Scandinavia simultaneously possesses the highest food bridging and lowest food pairing score, while Africa with the highest food pairing and lowest food bridging score. The delta scores of food pairing and bridging can be similarly plotted, as shown in Figure 11, which, when split on the axis where the cuisines equate to being indistinguishable from their random counterparts, subdivide the grid into four zones or classes:

- Class 1: positive pairing + positive bridging
- Class 2: Positive pairing + negative bridging
- Class 3: Negative pairing + positive bridging
- Class 4: Negative pairing + negative bridging

France, South America, Thailand, Middle East, Greece and Indian Subcontinent fall into class 1, where the ingredients strongly pair and bridge their flavours. China, Korea, Canada, South East Asia, Australia & NZ, Mexico, USA, Italy, Spain, Caribbean, and Africa into class 2, where ingredients mainly pair their flavour compounds with fewer indirect chains. Scandinavia, DACH Countries, British Isles, Eastern Europe and Japan fall into class 3, where flavour compounds are mainly bridged through indirect chains. Interestingly, and perhaps tellingly, there are no cuisines with negative food pairing and negative food bridging in class 4. From these observations, a dichotomy is apparent in that ingredients less suited to food pairing tend to use food bridging and vice versa.

Figure 18

The combined plot of the average food pairing (N(R)) and food bridging (SMP(R)) scores of each cuisine.

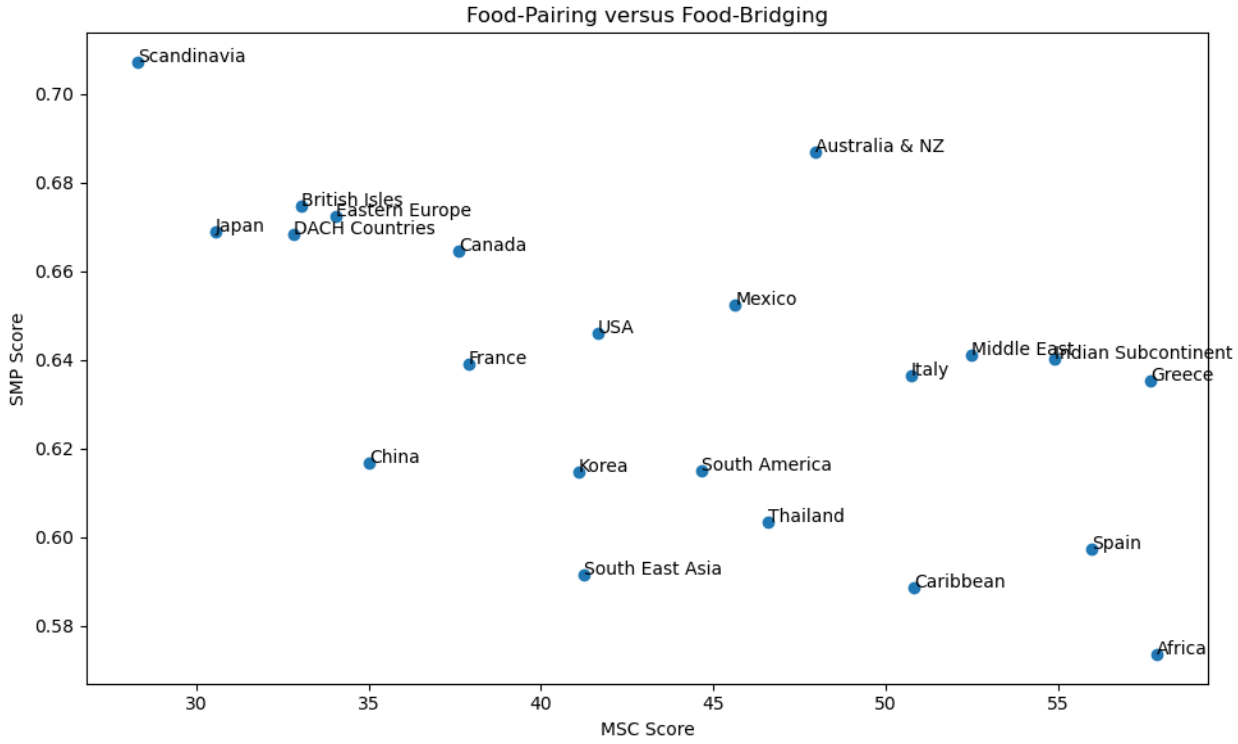
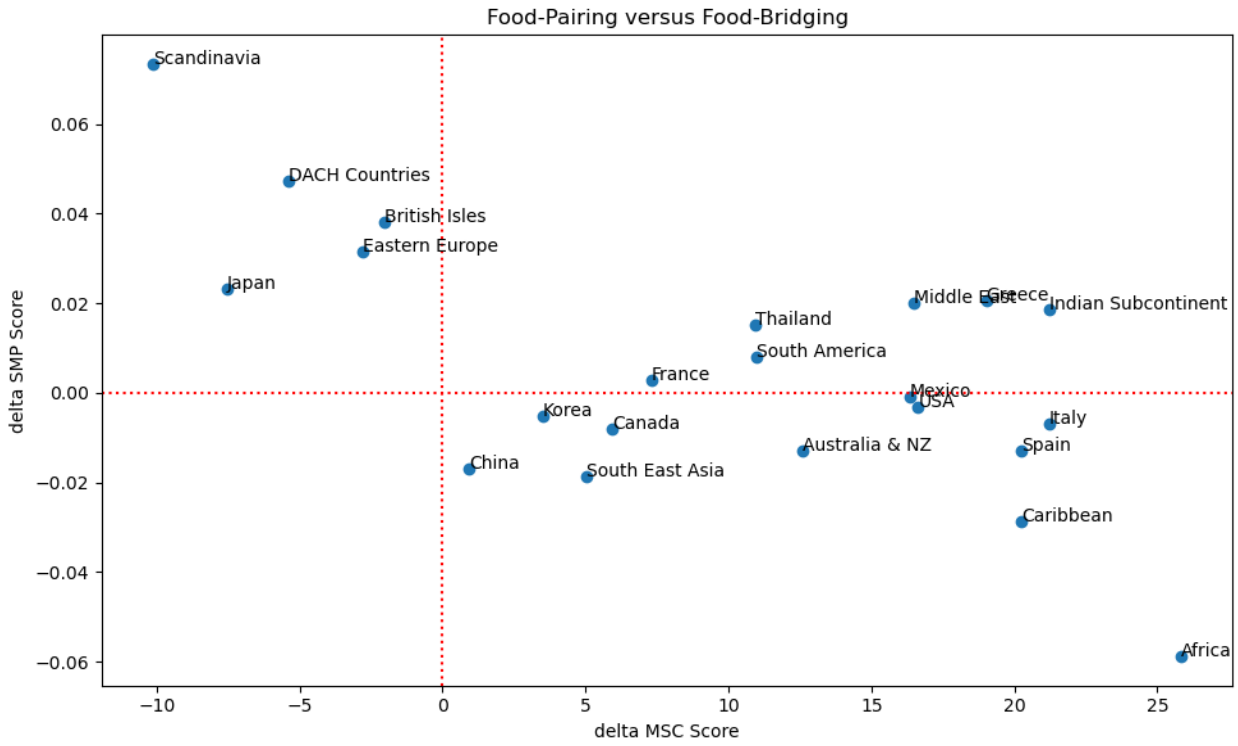


Figure 19

The combined plot of the delta food pairing and food bridging scores of each cuisine.



Frequent Pattern Mining

An FP-Growth Algorithm implementation was used to generate frequent itemsets for each cuisine. A minimum support level of 0.05 is used to ensure the extraction of patterns across a substantial number of recipes. Tuples of sizes ranging from 3 to 6 inclusive were considered from that subset (as we do not wish to consider singles or pairs of ingredients). The most common intragroup ingredient tuples from this set across all cuisines, based on support, are listed in Table 7.

Table 21

The 10 most common ingredient tuples across all cuisines based on intragroup support scores.

Cuisine	Combination	Support
Korea	Soy Sauce, Garlic, Sesame	0.389
Korea	Soy Sauce, Pepper, Sesame	0.342
Scandinavia	Butter, Flour, Sugar	0.342
Korea	Garlic, Sesame, Pepper	0.332
Korea	Soy Sauce, Sesame, Sugar	0.332
Korea	Soy Sauce, Garlic, Pepper	0.329
Korea	Soy Sauce, Garlic, Sugar	0.322
Korea	Garlic, Sesame, Sugar	0.319
Scandinavia	Flour, Egg, Sugar	0.304
DACH Countries	Egg, Flour, Sugar	0.3

It's interesting to note that Korean and Scandinavian cuisines heavily dominate the most common combinations. In particular, combinations involving soy sauce, garlic, and sesame are very common in Korean cuisine, while combinations involving basic baking ingredients like butter, flour, sugar, and egg are very common in Scandinavian and DACH (Germany, Austria, Switzerland) cuisines. The most common intergroup ingredient tuples from the set across all cuisines are listed in Table 8.

Table 22

The 10 most common ingredient tuples across all cuisines based on the sum of intergroup *support scores. The percentile of the tuple against the average MSC of all tuples of that size is also listed. For instance, a group of Garlic, Pepper and Onion shares many compounds relative to other 3-group tuples within cuisines and sits at the 92nd percentile relatively. This group also shares the largest proportion of world cuisines, followed by Salt, Garlic, and Pepper.*

Combination	Sum of Support Scores	Avg Tuple MSC Percentile
garlic, pepper, onion	2.251677	91.772222
salt, garlic, pepper	2.1136	37.611765
salt, pepper, onion	1.853193	35.85
olive, garlic, pepper	1.820229	89.058333
salt, garlic, onion	1.747947	42.423529
olive, garlic, onion	1.565498	91.05
flour, egg, sugar	1.510941	14.677778
butter, flour, egg	1.509926	19.06
salt, olive, garlic	1.505524	35.490909
salt, egg, flour	1.468048	13.87

1. Garlic, Pepper, Onion: This is the most common ingredient combination across all cuisines, with a total support value of 2.25. These ingredients are staples in many cuisines and are often used together in a wide variety of dishes.
2. Salt, Garlic, Pepper: This combination, with a total support of 2.11, is also quite common. Salt is a basic seasoning used in virtually all cuisines, while garlic and pepper add flavor and heat.
3. Salt, Pepper, Onion: With a total support of 1.85, this combination is also prevalent. Onion adds sweetness and depth to dishes, while salt and pepper are basic seasonings.
4. Olive, Garlic, Pepper: This combination, with a total support of 1.82, is common in Mediterranean cuisines, where olive oil is a staple.

5. Salt, Garlic, Onion: This combination, with a total support of 1.74, combines the flavour-enhancing properties of salt, the heat and flavor of garlic, and the sweetness and depth of onion.
6. Olive, Garlic, Onion: This combination, with a total support of 1.56, is common in cuisines that use olive oil as a base, such as Italy and Greece.
7. Flour, Egg, Sugar: This combination, with a total support of 1.51, is a basic ingredient set for many types of baked goods and desserts.
8. Butter, Flour, Egg: This combination, also with a total support of 1.51, is another basic set for baking. The presence of butter suggests these recipes may be from cuisines with a strong tradition of butter-based pastries, like France or USA.
9. Salt, Olive, Garlic: With a total support of 1.50, this combination is common in Mediterranean cuisines, where these ingredients are staples.
10. Salt, Egg, Flour: This combination, with a total support of 1.46, is versatile and could be used in a wide variety of dishes, from pastas and breads to cakes and pastries.

Once tuples were identified, rows were enriched by computing the MSC and SMP scores for the group. Additionally, an average MSC score for the tuple size (avg tuple MSC score) within each cuisine was added for comparison purposes. (i.e. MSC score for every combination of 3, 4, 5, 6 ingredients in each cuisine) An average tuple SMP score was found to be impractical to compute due to the immense time required to compute SMP scores against such a large number of combinations, so it was skipped. Correlation analysis was run against the tuple MSC, tuple SMP, avg tuple MSC, parent MSC and parent SMP scores, as shown in Table 8. The following were observed between the pairs:

- **MSC Score and Parent MSC Score**: These two variables have a strong positive correlation (0.506048), which suggests that as the MSC Score increases, the Parent MSC Score also tends to increase.
- **Avg Tuple MSC Score and Parent MSC Score**: These variables show a very strong positive correlation (0.851566), indicating that higher Tuple MSC Scores are associated with higher Parent MSC Scores. This seems to indicate that if the bulk of tuples score higher on bridging, then the cuisine as a whole will too.

- **Parent MSC Score and Parent SMP Score**: These two variables have a strong negative correlation (-0.699593), which suggests that as the Parent MSC Score increases, the Parent SMP Score tends to decrease. This appears to align with the earlier observations in Food Pairing and Food Bridging, where a cuisine, on average, veered towards an either-or between pairing and bridging.

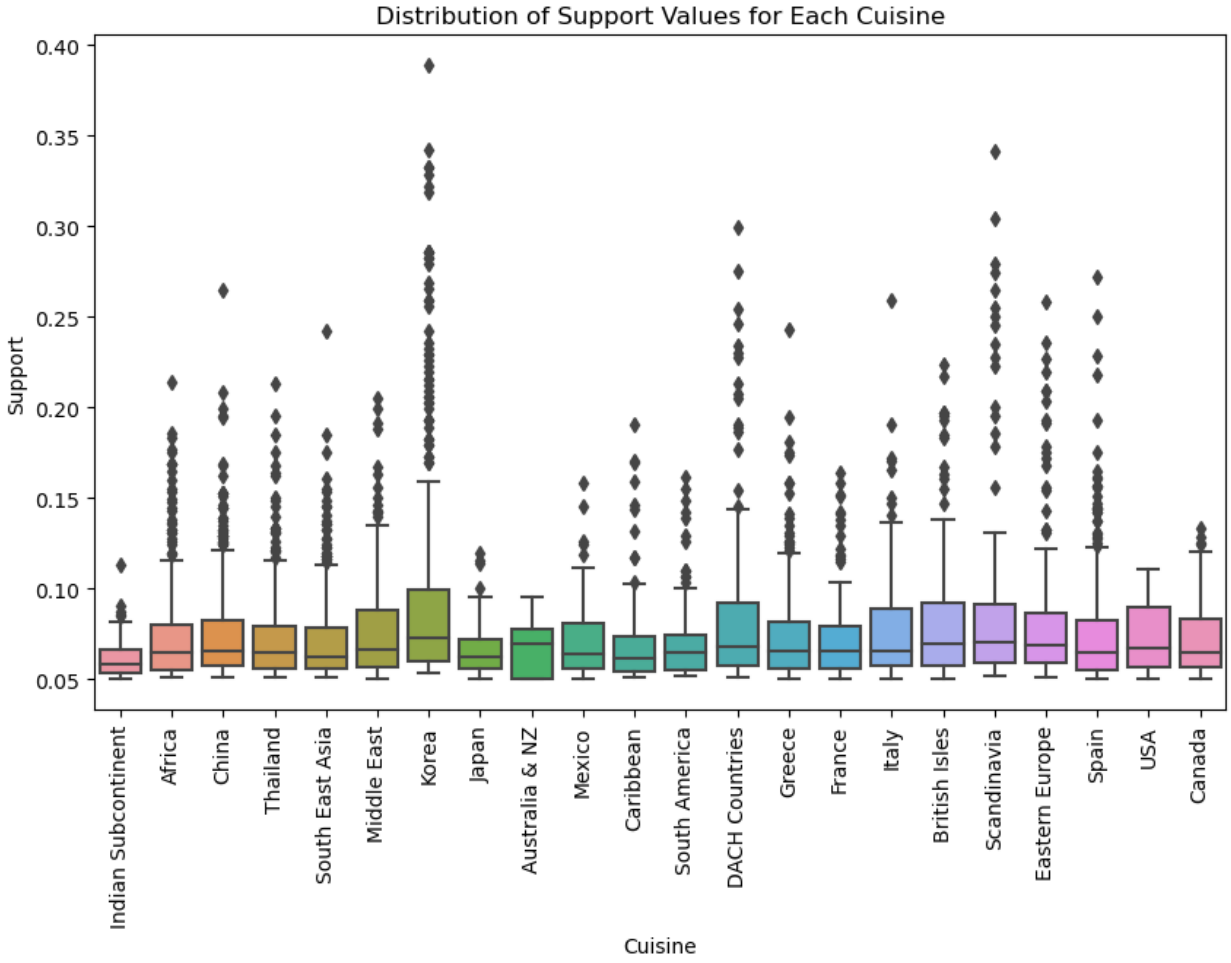
Table 23

Spearman correlation matrix between tuple MSC, tuple SMP, avg tuple MSC, parent MSC, parent SMP scores, parent MSC score delta and parent SMP score delta.

	MSC Score	SMP Score	Avg Tuple MSC Score	Parent MSC Score	Parent SMP Score	Parent MSC Score Delta	Parent SMP Score Delta
MSC Score	1	0.118873	0.37747	0.506048	-0.334535	0.491599	-0.278688
SMP Score	0.118873	1	0.036213	-0.043123	0.089291	-0.024836	0.094459
Avg Tuple MSC Score	0.37747	0.036213	1	0.851566	-0.505171	0.844115	-0.291956
Parent MSC Score	0.506048	-0.043123	0.851566	1	-0.699593	0.969995	-0.515539
Parent SMP Score	-0.334535	0.089291	-0.505171	-0.699593	1	-0.663494	0.826251
Parent MSC Score Delta	0.491599	-0.024836	0.844115	0.969995	-0.663494	1	-0.548951
Parent SMP Score Delta	-0.278688	0.094459	-0.291956	-0.515539	0.826251	-0.548951	1

Figure 20

The boxplot shows the distribution of support values for each cuisine when support exceeds 0.05. The support value represents the relative frequency of each ingredient tuple within a given cuisine. Note the extensive outliers for Korea and Scandinavia, whose most extreme tuple support values approach 0.4 and 0.35, respectively.



From Figure 13, we can observe that there is a wide variation in support values across different cuisines, implying that the commonality of ingredient tuples can vary greatly from one cuisine to another. Certain cuisines, like Korea, Scandinavia, and DACH Countries, tend to have higher support values for their most common ingredient tuples. This suggests that these cuisines might have a more defined set of common ingredients or that the data for these cuisines contains less variety. The cuisines of France, Italy, and China show a wider spread of support values, suggesting a greater diversity of ingredient combinations.

Figure 14 shows a different perspective based on the total count of higher-order tuples per cuisine with a minimum support of 5%. Combined with the recipe counts available to each (Table 9), we can make some interesting observations about the complexity of world cuisines. Cuisines with a higher tuple count compared to their recipe count (Korea, Africa, China) could indicate a higher prevalence of specific culinary ‘fingerprints’ that define common flavours to that cuisine (exemplified with the tuple of Soy Sauce, Garlic, and Sesame in the Korea cuisine). It could indicate that these cuisines focus on certain signature flavours or traditional combinations. Cuisines with a low tuple count relative to their recipe count (USA, Italy) might rely more on simpler ingredient combinations, have a stronger emphasis on specific key ingredients, or have a very diverse set of dishes.

Figure 21

Heatmap of the count of N-sized tuples per cuisine with a cutoff support value of 0.05.

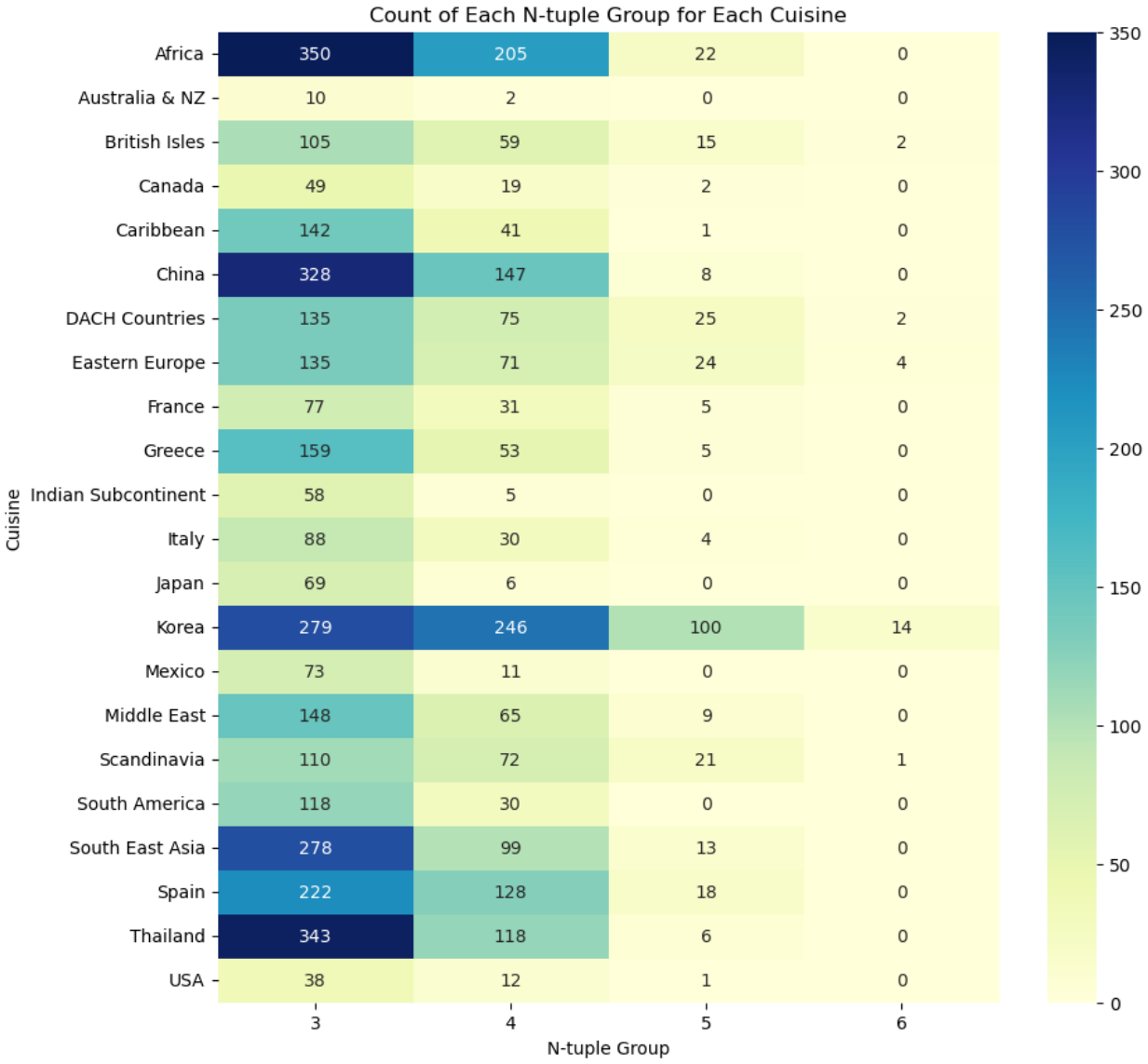


Table 24

Count of ingredient tuples per Cuisine (with support >= 0.05 and ingredient size of 3-6), the recipe count of the cuisine in the dataset, and the ratio between the two.

Cuisine	Tuple Count	Recipe Count	Ratio
Africa	577	650	0.888
Australia & NZ	12	494	0.024
British Isles	181	1074	0.169
Canada	70	1112	0.063
Caribbean	184	1103	0.167
China	483	941	0.513
DACH Countries	237	487	0.487
Eastern Europe	234	565	0.414
France	113	2701	0.042
Greece	217	934	0.232
Indian Subcontinent	63	4057	0.016
Italy	122	7502	0.016
Japan	75	578	0.130
Korea	639	301	2.123
Mexico	84	3138	0.027
Middle East	222	993	0.224
Scandinavia	204	404	0.505
South America	148	310	0.477
South East Asia	390	611	0.638
Spain	368	815	0.452
Thailand	467	666	0.701
USA	51	16106	0.003

To further probe ingredient tuple influences on a cuisine's pattern of positive or negative pairing and bridging (first glimpsed at with the correlations between tuples and the parent MSC scores in Table 8), a classification task was created to gauge whether the tuple MSC and SMP scores of a cuisine could be used to predict whether the parent cuisine scored positive or negative in food pairing and food bridging respectively. A decision tree classifier was chosen for its simplicity and interpretability, allowing us to see the importance of different features in making predictions. To address the diminishing number of tuples as they ascend in size and to address the unbalanced number of positive food-pairing cuisines over negative ones, for each N-tuple, a balanced dataset was created by undersampling the majority class to match the minority class size. Stratified 5-fold cross-validation on the balanced dataset was then performed, with a holdout of 20% for the test data set. In each iteration, a decision tree classifier was trained, and run against the held-out test set, then precision, recall, and F1-score were calculated. Finally, a classifier on the entire balanced dataset was run. The results of the classifier against each N-tuple, against the binary classification of the parent MSC and SMP class, are listed in Table 10.

From the results in Table 10, the classifier had good to excellent predictions of the parent's food pairing class from the constituent ingredient tuples at every size. For the Parent MSC Score Class, the model performs progressively better as the N-tuple increases. This suggests that more complex ingredient combinations might be more indicative of the parent cuisine's MSC Score Class. For the Parent SMP Score Class, there's a significant increase in performance from N-tuple 3 to N-tuple 4 and then again from N-tuple 4 to N-tuple 5, with the model achieving perfect performance with N-tuple 6. This again suggests that more complex ingredient combinations might be more indicative of the parent cuisine's SMP Score Class. Comparing the two, the model seems to have performed better overall at predicting the Parent MSC Score Class than the Parent SMP Score Class. This might suggest that the MSC score is more strongly tied to ingredient combinations than the SMP score, at least in this dataset. Due to the scarcity of tuple data at an N-tuple size of 6, this could be skewed by the lack of data at this level, but for the smaller tuple sizes where data is more plentiful, this is significant. The same is less so claimed by the classifiers predicting the SMP class of the parent,

which ranged from poor (N-tuple = 3) to good (N-tuple = 5). Once again, the lack of data at size 6 could skew the results at that level. There is an additional caveat when predicting the SMP class: as a consequence of the measurement used for food-bridging, SMP scores disproportionately higher with a greater number of items in a set (i.e. more connections are available between more ingredients, which is unlike the MSC score is calculated), so one would expect that a more accurate SMP score would result from a higher N-tuple ingredient combination, which means smaller tuple sizes would not be representative.

Table 25

Results of the binary classification task, with the dataset split based on N-tuple size and the associated average precision, recall and F1 scores of the 5-fold cross-validation.

Target	N-tuple	Average precision	Average recall	Average F1-score
Parent MSC Score Class	3	0.95	0.67	0.79
Parent MSC Score Class	4	0.99	0.85	0.91
Parent MSC Score Class	5	0.99	0.94	0.97
Parent MSC Score Class	6	1.00	1.00	1.00
Parent SMP Score Class	3	0.50	0.43	0.46
Parent SMP Score Class	4	0.60	0.65	0.62
Parent SMP Score Class	5	0.81	0.82	0.81
Parent SMP Score Class	6	1.00	1.00	1.00

Due to the choice of using the Decision Tree algorithm for classification, of which transparency is a major advantage, the feature importances for each classifier can be extracted, along with the precise rules the classifier has learned to split the data. It was found that for all classifiers, the input MSC Score was by far the more significant of the inputs, which consistently contributed to 83% or greater importance to the classification results.

DISCUSSION

The ability to predict the parent cuisine's food pairing or bridging score, based on the scores of its constituent ingredient groups, provides valuable insights into the structure and characteristics of cuisines. The results suggest that frequent ingredient groups significantly influence a cuisine's overall food pairing score, indicating that a cuisine's culinary tradition is shaped by the combination of its ingredients rather than individual elements.

Understanding these common ingredient combinations offers a framework for creating new recipes or modifying existing ones. For example, cuisines with a positive food pairing score can benefit from choosing ingredient groups with positive food pairing scores when crafting new recipes. Additionally, identifying recurring ingredient combinations can help capture a cuisine's culinary 'fingerprint', particularly in cuisines that strongly support traditional ingredients or flavor combinations. This knowledge can inspire new food pairings based on individual preferences for harmonious or contrasting flavors, supporting culinary enthusiasts, chefs, and automated recipe generators in creating innovative and appealing dishes.

Moreover, examining the correlations between the MSC and SMP scores of ingredient tuples and the overall MSC and SMP scores of cuisines can guide the selection of ingredients for new recipes. This understanding of the balance between food pairing (where ingredients share flavor compounds) and food bridging (where ingredients do not share flavor compounds but have intermediary ingredients that do) allows recipe creators to make informed decisions about ingredient combinations. The developed decision tree classifier can predict the potential pairing or bridging tendencies of a new recipe based on its tuple MSC and SMP scores, indicating the likely palatability before physical testing and saving time and resources. Incorporating these findings into AI recipe generators can empower them to create innovative recipes that respect the culinary traditions they draw inspiration from, particularly useful for dietary restrictions or personalized meal planning with nutritional balance and diverse flavors.

In conclusion, this study provides data-driven insights that can revolutionize recipe creation in traditional and computational gastronomy. It enriches recipe development with a deeper understanding of the fundamental building blocks of different world cuisines, offering the potential for diverse and enjoyable culinary experiences.

FUTURE WORK

While this study has provided valuable insights into the significance of ingredient groups in determining the overall food pairing/bridging score of cuisines, there is potential for further refinement and expansion of the analysis. One avenue for future work is to investigate the impact of ingredient quantities and proportions within a recipe. Analyzing how varying amounts of specific ingredients influence the cuisine could provide a more nuanced understanding of the interplay between flavors. There is a need for analysis to take into account the concentrations of the molecular compounds, as proportions play a large part of the final flavours of a dish.

Perhaps just as important are the consideration for the effects of heat and preparation techniques on ingredients. As a prime example, Wok hei, the so-called “breath of a wok” coined in “The Wisdom of the Chinese Kitchen: Classic Family Recipes for Celebration and Healing” (Young 1999), outlines a fundamental concept in Cantonese stir-frying. Wok hei involves cooking at very high temperatures, using a traditional round-bottomed wok made of materials with excellent heat conductivity. Rapid cooking through continuous tossing and stirring of ingredients, even heat distribution, minimal sauce, and skillful technique are essential in creating the characteristic flavors: a distinct smoky and charred depth that is not easily replicable by alternative means. Such quantitative measurements of the resulting compounds created by the cooking process would further lend a more accurate understanding of food pairing or bridging analysis.

APPENDIX A

GitHub Repository

https://github.com/JamesMarcogliese/mrp-higher-order-pairings

REFERENCES

1. Ahn, Y. Y., Ahnert, S., Bagrow, J., et al. (2011). Flavor network and the principles of food pairing. Scientific Reports, 1, 196.
2. Ahn, Y.-Y., & Ahnert, S. E. (2013). The flavor network. Leonardo, 46, 272-273.
3. Bahk, Y. W., & Ahn, G. S. (1977). Lactose intolerance in the Koreans. Modern medicine of Asia, 13(3), 7–8.
4. Blumenthal, H. (2008). The big fat duck cookbook. Bloomsbury.
5. Choo, V. (2018, December 20). Vchoo/vchoo.github.io. GitHub. https://github.com/vchoo/vchoo.github.io
6. Batra, D., Diwan, N., Upadhyay, U., Kalra, J. S., Sharma, T., Sharma, A. K., Khanna, D., Marwah, J. S., Kalathil, S., Singh, N., Tuwani, R., & Bagler, G. (2019). RecipeDB: A Resource for Exploring Recipes and Their Nutritional Profiles. Retrieved from http://cosylab.iiitd.edu.in/recipedb/
7. Issa, L., Alghanim, F., & Obeid, N. (2018). Analysis of food pairing in some Eastern Mediterranean countries. In 2018 8th International Conference on Computer Science and Information Technology (CSIT), Amman, Jordan (pp. 167-172).
8. Jain, A., Rakhi, N. K., & Bagler, G. (2015). Spices form the basis of food pairing in Indian Cuisine. PLOS ONE, 10(3), e0118697.
9. Jain, A., Rakhi, N. K., & Bagler, G. (2017). Analysis of food pairing in regional cuisines of India. PLOS ONE, 12(8), e0180939.
10. Kim, J. Y., Glober, N., & Varshney, L. R. (2019). RecipeDB: a resource for exploring recipes. IEEE Transactions on Knowledge and Data Engineering.
11. Landes, M. (n.d.). USDA Economic Research Service - the Elephant is jogging: New pressures for agricultural reform in India. USDA ERS. https://web.archive.org/web/20161018015249/http://www.ers.usda.gov/amber-waves/2004-february/the-elephant-is-jogging-new-pressures-for-agricultural-reform-in-india.aspx#.WAWAdOzP1qs

12. Simas, T., Ficek, M., Diaz-Guilera, A., Obrador, P., & Rodriguez, P. R. (2017). Food-bridging: a new network construction to unveil the principles of cooking. Frontiers in ICT, 4, 14.
13. Simas, T., & Suckling, J. (2016). Commentary: Semi-Metric Topology of the Human Connectome: Sensitivity and Specificity to Autism and Major Depressive Disorder. Frontiers in neuroscience, 10, 353. https://doi.org/10.3389/fnins.2016.00353
14. Singh, N., & Bagler, G. (2018). Data-driven investigations of culinary patterns in traditional recipes across the world. 2018 IEEE 34th International Conference on Data Engineering Workshops (ICDEW), 157-162.
15. Young, G. (1999). The Wisdom of the Chinese Kitchen: Classic Family Recipes for Celebration and Healing. Simon & Schuster.
16. Tyrrell, J. (2018, January 10). How does meat taste to carnivores? - londolozi blog. Blog. https://blog.londolozi.com/2017/12/24/how-does-meat-taste-to-carnivores/
17. Varshney, K. R., Pinel, F., Varshney, L. R., Bhattacharjya, D., Schelldorfer, J., Korhonen, J. H., & Freiwald, C. (2018). FlavorDB: a database of flavor molecules. IBM Journal of Research and Development, 59(2/3), 1-14.
18. Zhu, Y.-X., Huang, J., Zhang, Z.-K., Zhang, Q.-M., Zhou, T., & Ahn, Y.-Y. (2013). Geography and similarity of regional cuisines in China. PLOS ONE, 8(11), e79161.
19. Zhou, W. X., & Yan, H. Y. (2018). Hierarchical clustering of world cuisines. Physical Review E, 98(3), 032413.