Chapter- III

Materials and Methods

Chapter-III

3. Materials and Methods

This chapter describes the procedure followed in conducting a systematic investigation to address the research questions of this study. It encompasses a description of the materials and methods of their preparation, the methodology for their data collection including a description of the algorithms used, and the framework for data interpretations used in the pursuit of fulfilling the objectives.

The chapter is arranged to describe the steps involved in the data collection process, and pre-processing of the data, under the section 3.1. The statistical analyses used in the study are described briefly in the section on theoretical backgrounds, section 3.2. This is followed by a brief description of the methodology followed for fulfilling the four stated objectives of the study.

3.1 Materials

3.1.1 Data collection

Our study involved the collection of information on traditional recipes at three levels, namely, recipes, ingredients and flavour compounds.

3.1.1.1 Recipe

The initial step involves an extensive data curation process considering various sources such as online databases and cookbooks. Cookbooks may be of varying types based on the type of meals designed for, a quick recipe meals or sophisticated meals for an expensive restaurant. These differences can result in the variations in the size of recipes as it is more likely that a cookbook focusing on sophisticated, high-level recipes will use richer ingredients per recipe as compared to simple cookbooks. However, traditional cookbooks, are considered to be more reliable since they are typically written in a more refined and sophisticated manner containing more ingredients in the recipes [5]. In addition to providing instructions on how to prepare a dish, a recipe serves as a description of the end product, describing the ingredients to be used, their quantities, and their transformations [87].

Considering the above statement, recipe data collection was collected from traditional cookbooks. Additionally, to overcome text parsing error and disparity in ingredient usage

which is mostly reported in the case of cookbooks we considered data from online database.

The food recipes data of the eight regional cuisines of the Northeast were obtained from cookbooks, *Essential North-East Cookbook*, *The seven sisters: kitchen tales from the north east* [33, 71] in addition to an online database of *Assamese cuisine tastes real freshness* [40]. A total of 620 recipes were obtained from these sources which were used for the purpose of the study. The name of each recipe and the list of ingredients were extracted for each recipe as shown in Table 3.1. The recipes collected consist of mostly traditional recipes which are known to represent the culinary tradition of the region.

Table 3.1 Format for tabulation of recipe data

Name of recipe	**Ingredient names**					
Oying (vegetable stew)	Mustard leaves	Cabbage	French beans	Potatoes	Green chillies	Ginger paste
Sana thongba (cottage cheese cooked in milk)	Mustard oil	Bay leaves	Potatoes	Green peas	Turmeric powder	Milk

3.1.1.2 Ingredients

The list of ingredients was obtained from the recipe dataset after undergoing a series of pre-processing. This process was carried out to remove the irrelevant phrases and quantifiers. The final dataset consists of a recipe and its accompanying ingredients.

a) *In names*: The ingredients were aliased to their source ingredients. For example, green chilli paste was aliased as green chilli. In addition, derived ingredients which are a set of ingredients combined such as garam masala (black pepper, mace, cinnamon, clove, cardamom, nutmeg), panch phoron seeds (fenugreek, nigella seed, cumin, fennel, mustard seed) and ginger garlic paste (ginger, garlic) were separated into its constituent individual ingredients.

   Some ingredients may be renamed to match the names of the ingredients available in Ahn's dataset [5]. For example, bay leaf was renamed as bay laurel. Ingredient variations were also taken into account in terms of linguistics and spelling. For example, diced onion and sliced onion; both are onion, jeera and cumin etc. were regarded as similar entities.

b) *In specifiers of size quantity form and condition*: Quantifiers such as '1 cup' or '5ml' and words referring to consistency or temperature, e.g., thick or hot were removed.

Further, as the current study was related to flavour pairing, only the names of each ingredient used in the recipe were retained to map the ingredients to their flavour profile (list of volatile compounds present in that ingredient). In doing so, special characters, culinary stop words, and punctuation marks were removed. All words were converted into their singular form and saved in lower case (Fig.3.1). Finally, the ingredients were divided into their constituent categories based on their nature and origin. These steps were carried out with the help of NLTK (Natural language toolkit), python packages Fig. 3.2. In some cases, it has been carried out manually, in the case of cookbooks, where manual entry of the recipe and ingredients is required.

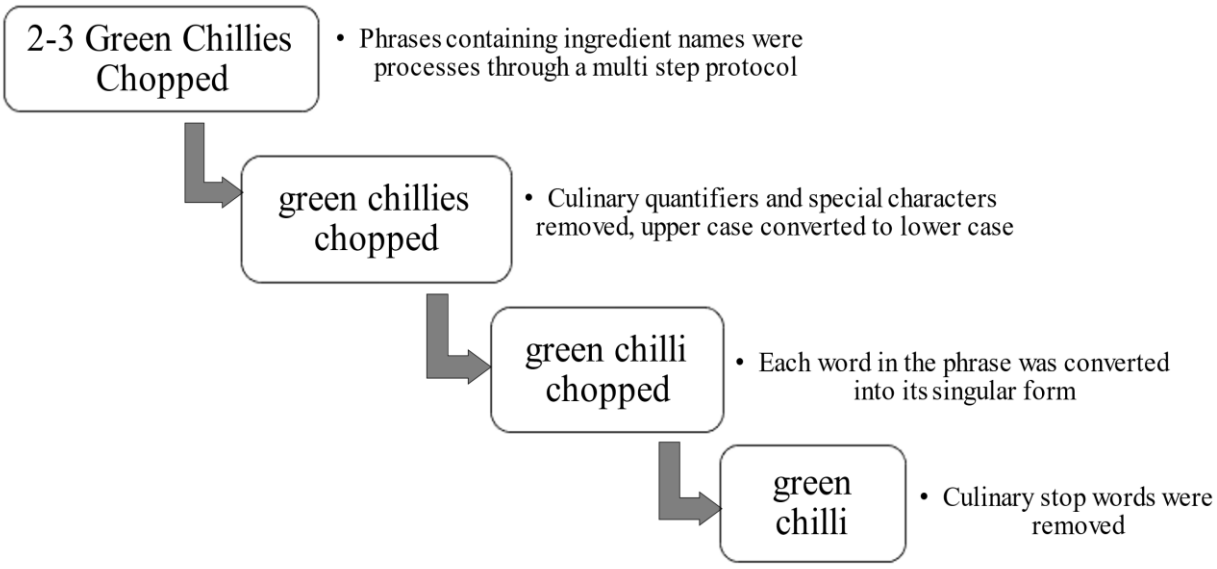

Fig. 3.1 Steps for data pre-processing

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt

In [2]: import seaborn as sns
        from nltk.stem import WordNetLemmatizer
        from nltk.stem.snowball import SnowballStemmer
        from nltk.tokenize import word_tokenize
        import re
        import pickle
        import itertools
        from collections import Counter
```

Fig. 3.2 Python libraries used for pre-processing of recipe data

3.1.1.3 Flavour compound

The ingredients flavour compound data for the regional cuisine was collected from the archived data made available by Ahn at al. [5] and Jain et al. [39]. The flavour compound dataset consists of its unique ID along with its corresponding compound name and chemical abstract service (CAS) number, an example shown in Table 3.2.

Ahn et al. [5] tabulated the ingredient flavour compound data from Fenaroli's Handbook of Flavour Compounds 5th edition [23]. The archived data is available as supplementary information of Ahn's [5]. Three datasets were available in the form of comma-separated value (CSV) files and tab-separated value (TSV) files. The dataset consists of (i) compound ID mapped against the compound name along with its CAS number (ii) ingredient ID mapped against its corresponding flavour compounds ID (iii) compound ID mapped against ingredient ID. In addition, Jain et al. [39] updated the dataset of Ahn's with the additional fifty ingredients of the Indian regional cuisines. The same dataset was made available in the form of CSV and TSV files.

Table 3.2 Format for tabulation of flavour compound data

Compound ID	Compound Name	CAS Number
0	Jasmone	488-10-8
1	5-methylhexanoic_acid	628-46-6
3	1-methyl-3-methoxy-4-isopropylbenzene	1076-56-8

3.1.2 Software implementation

A few standard software was used for the computation of data integration, analysis and visualization tasks (Table 3.3).

Table 3.3 List of software used

S.N.	Name of software
1	Microsoft Excel
2	Python 3.9.6
3	Cytoscape 3.8.2

3.1.2.1 Cytoscape as a tool for flavour network visualization

Cytoscape is an open-source software project, which allows the visualization of biomolecular interactions and high-throughput expression data (Fig.3.3). Despite its origins as a biological research tool, Cytoscape has evolved into a general network visualization and analysis tool. Using the Cytoscape core programme, data integration, analysis, and visualization are all made possible. Users can arrange and query their networks, integrate them with various expression patterns, ontologies, and other molecular variables, and connect them to functional annotation databases. Extensions to the Core are easily achieved through a plug-in architecture, which greatly facilitates the rapid development of additional analytical and computational capabilities. There are available apps for network analysis, molecular profiling, screen layouts, additional file formats, scripts, and databases [68].

3.1.2.1.1 Core features of cytoscape

3.1.2.1.1.1 Data input format

GraphML, KGML (KEGG XML), SBML, OBO, Gene Association, SIF (Simple Interaction Format), GML, XGMML, BioPAX and PSI-MI are just some of the network and annotation formats that are supported by Cytoscape. Additionally, we can import data files, such as expression profiles or GO annotations, generated by other programs or spreadsheets, such as delimited text files and MS Excel workbooks (Fig. 3.4). This feature allows for arbitrary attributes to be loaded and saved on nodes, edges, and networks. For our study of the construction of the flavour network, the network file is imported as an

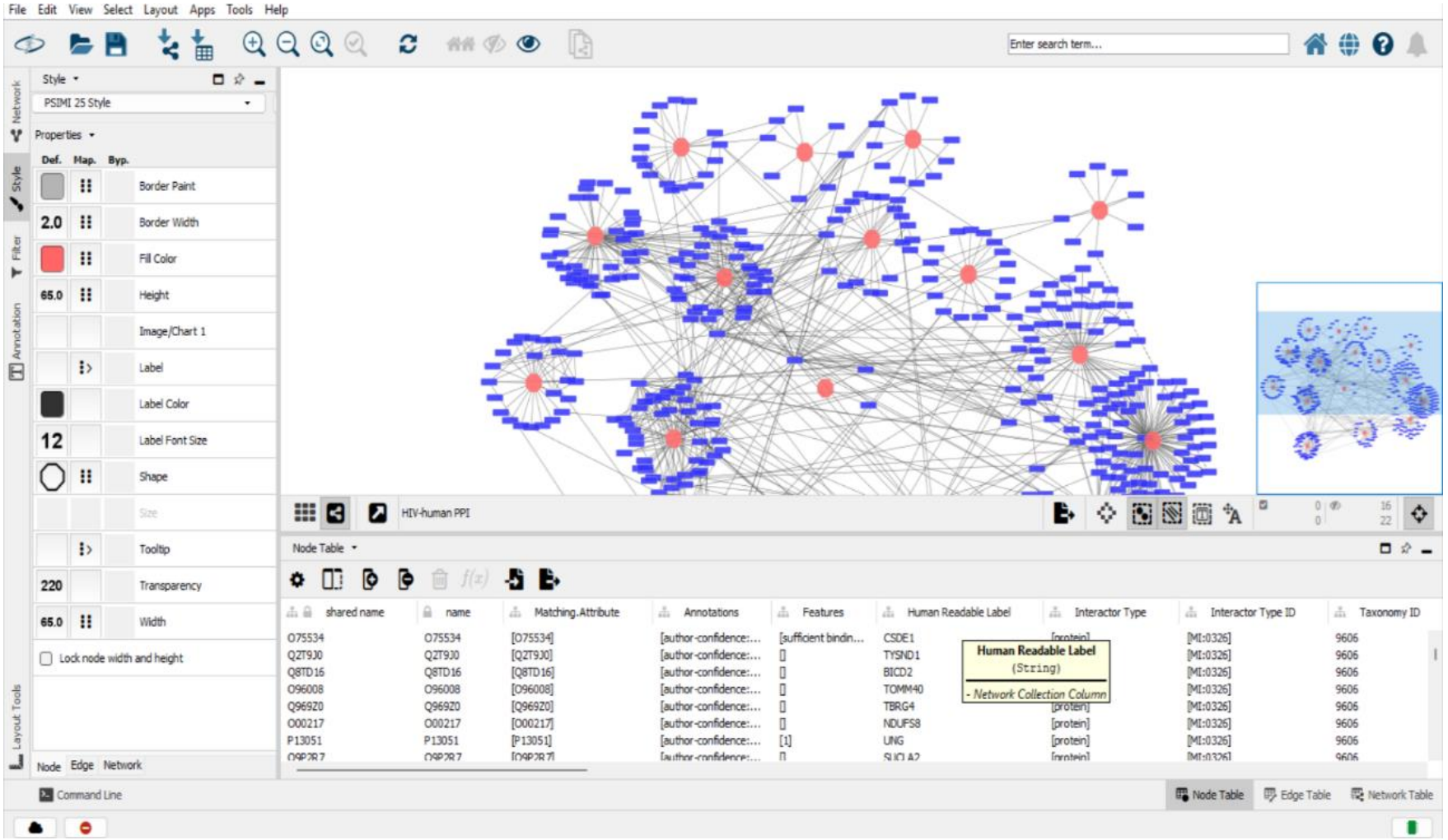

Fig 3.3 (a) Cytoscape 3.8.3 Desktop – View-1

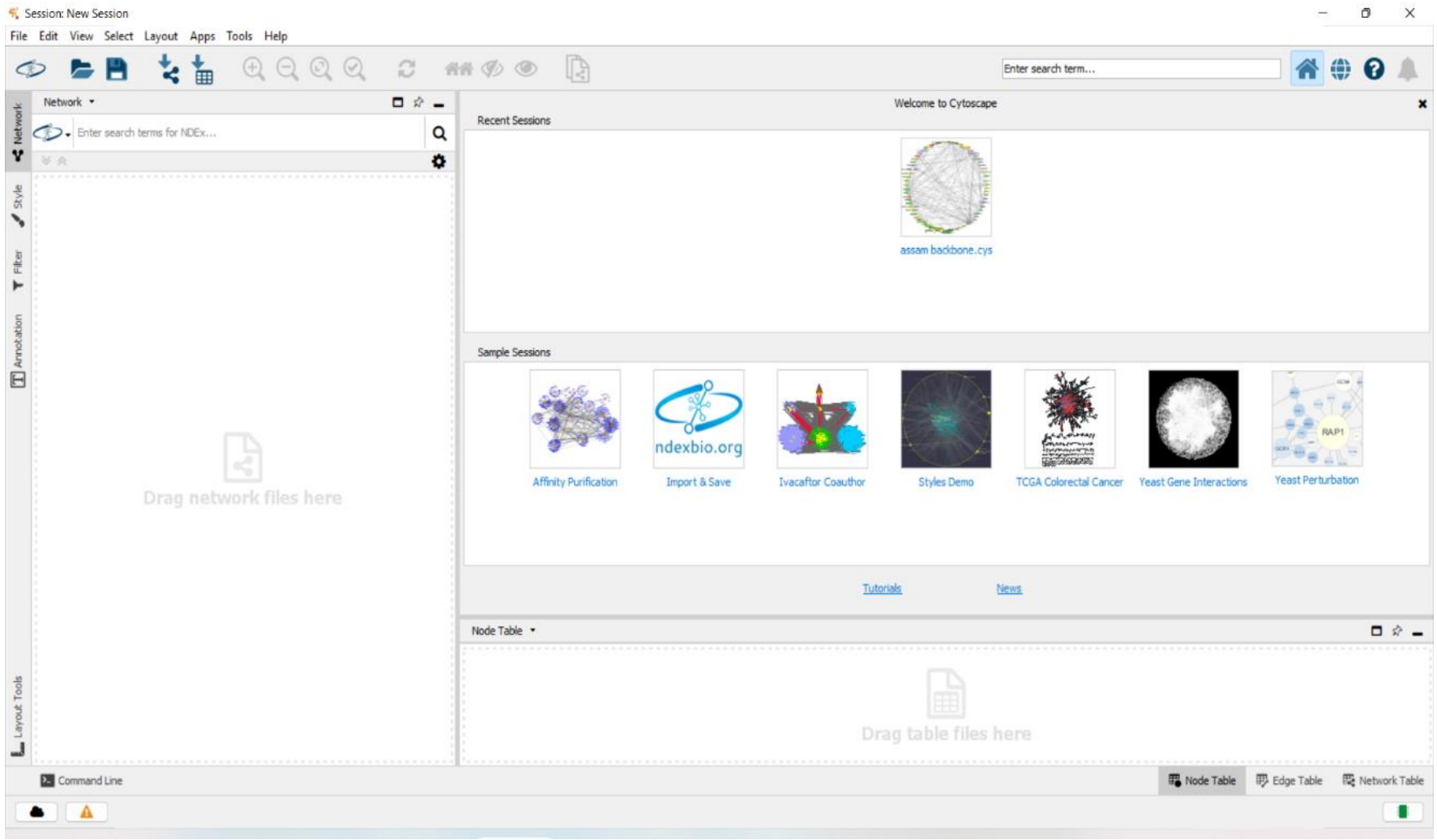

Fig 3.3 (b) Cytoscape 3.8.3 Desktop – View-2

excel workbook. Once the network data is imported, we assigned the function such as source node, target node, source node attribute, target node attribute and edge attribute depending on the requirements of the network to be constructed.

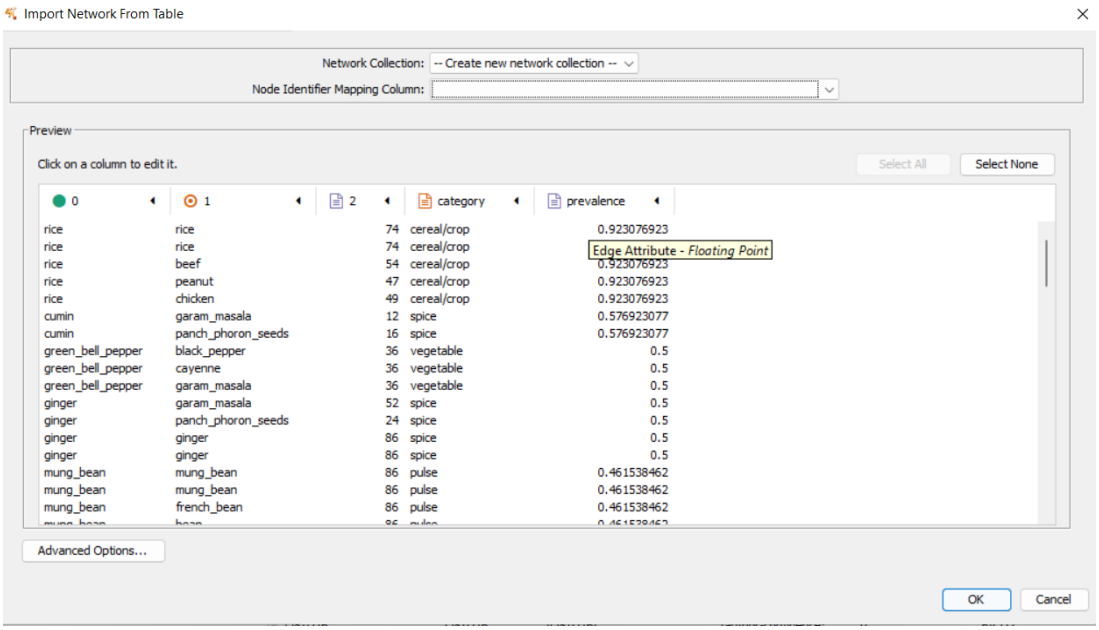

Fig. 3.4 Cytoscape 3.8.3 import network data files

3.1.2.1.1.2 Session file, layout and image export

We can save the settings, data, and visualizations as Cytoscape Session (.cys) files. Networks in Cytoscape are laid out in two dimensions. A variety of layout methods are available such as cyclic, tree, force-directed, edge-weight, and yFiles Organic layouts (Fig. 3.5). Additionally, tools for manual layout are also available and resemble the user interface of other graphics programmes. For our study, we use the layout style compound spring embedded with bundle edge 0 to reduce the density of edges. Furthermore, networks can be exported as publishable-quality images. It supports the following file types: JPEG, PNG, SVG, PDF, and PS (Fig. 3.6).

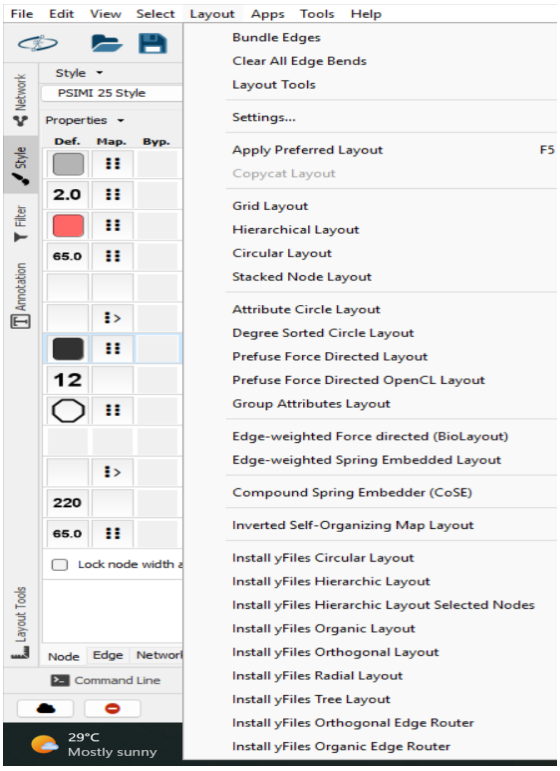

Fig. 3.5 Cytoscape 3.8.3 layout styles

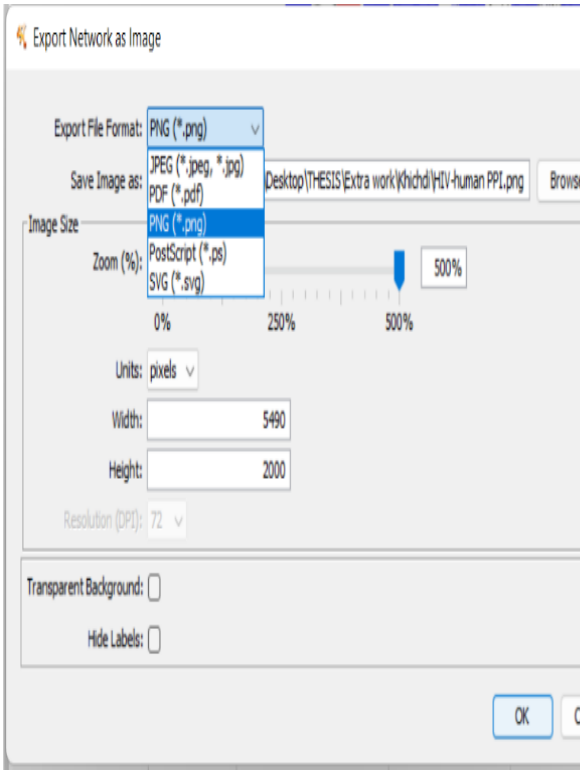

Fig. 3.6 Cytoscape 3.8.3 export images

3.1.2.1.1.3 Visualization styles in cytoscape

We can customise the network data display utilizing robust visual display functions. User-configurable colours and visualization schemes can be applied to expression data, such as node colour, node size, label fonts, label colour, edge thickness, edge colour etc (Fig.3.7). For the construction of the flavour network, we customise our style where we assign different node colours for each ingredient category and the size of the node depending on the prevalence of ingredients. Further, the size or thickness of the edges between each node was customised according to the number of shared flavour compounds between the ingredients. The style of the node is assigned according to the target column that we want depending on that we either use the function discrete mapping, continuous mapping and passthrough mapping. For our study, we use the mapping style continuous mapping for assigning the node size, edge width and label font size (Fig. 3.8).

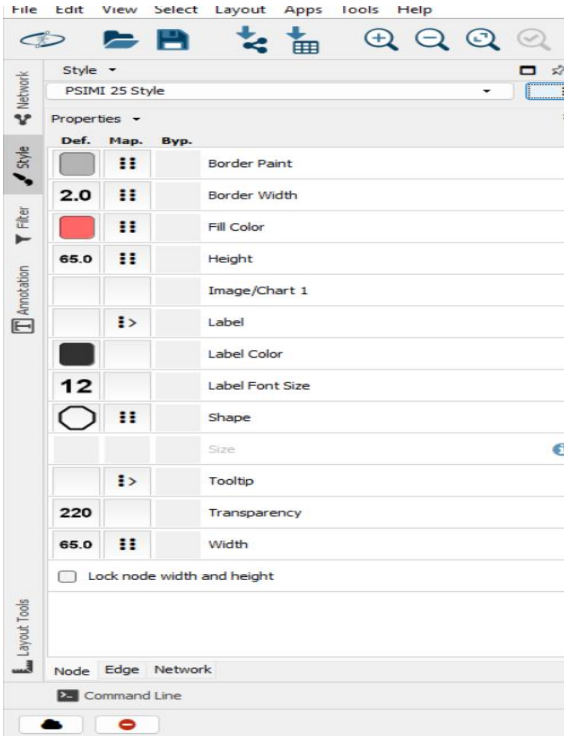

Fig. 3.7 Cytoscape 3.8.3 visual styles

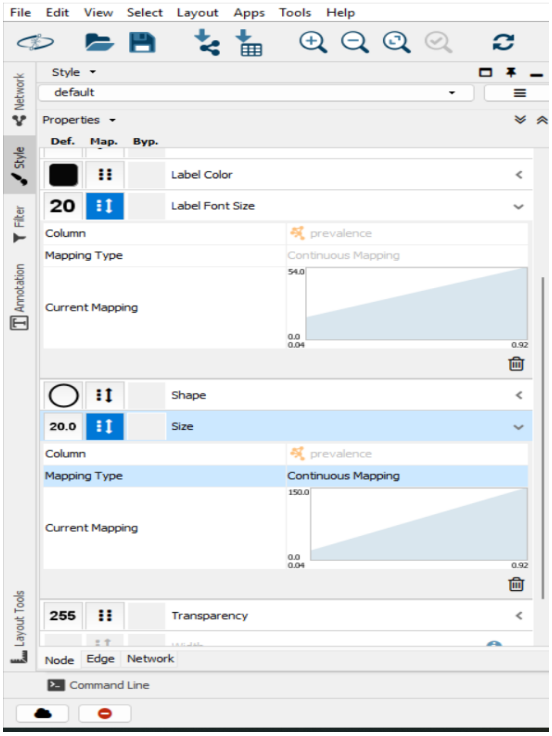

Fig. 3.8 Cytoscape 3.8.3 mapping styles

3.1.2.2 Data compilation and data analysis using python programming

Works related to data compilation, data integration and data analysis were done using python. Libraries such as *NumPy, pandas and networkX* were used for the backbone extraction of the flavour network and the creation of random recipes. In addition, libraries such as *matplotlib, seaborn, re, pickle, and itertools* were used for the analysis of cosine similarity and non-matrix factorization (NMF) and libraries such as *json, sys, and os* were used for the analysis of unigram and skipgram.

The data was stored in the format comma-separated value (CSV), and tab-separated value (TSV).

3.1.2.2.1 Data source

The data was gathered from three different sources,

1. Ingredient flavour compound data from Fenaroli's handbook of flavour compounds
2. Archived data from Ahn et al. [5]
3. Archived data from Jain et al. [39] (Ds1, Ds2, Ds3)

3.1.2.2.2 Data files

The datafile required for the study was compiled and stored in four different files,

1. Ds1- Ingredient IDS and Ingredient name
2. Ds2- Flavour compound ID and name of flavour compound
3. Ds3- Ingredient ID and Flavour compound ID
4. DS4- Recipes collected in present work

Further, the data frame was created in python and the illustration is shown in Fig 3.9

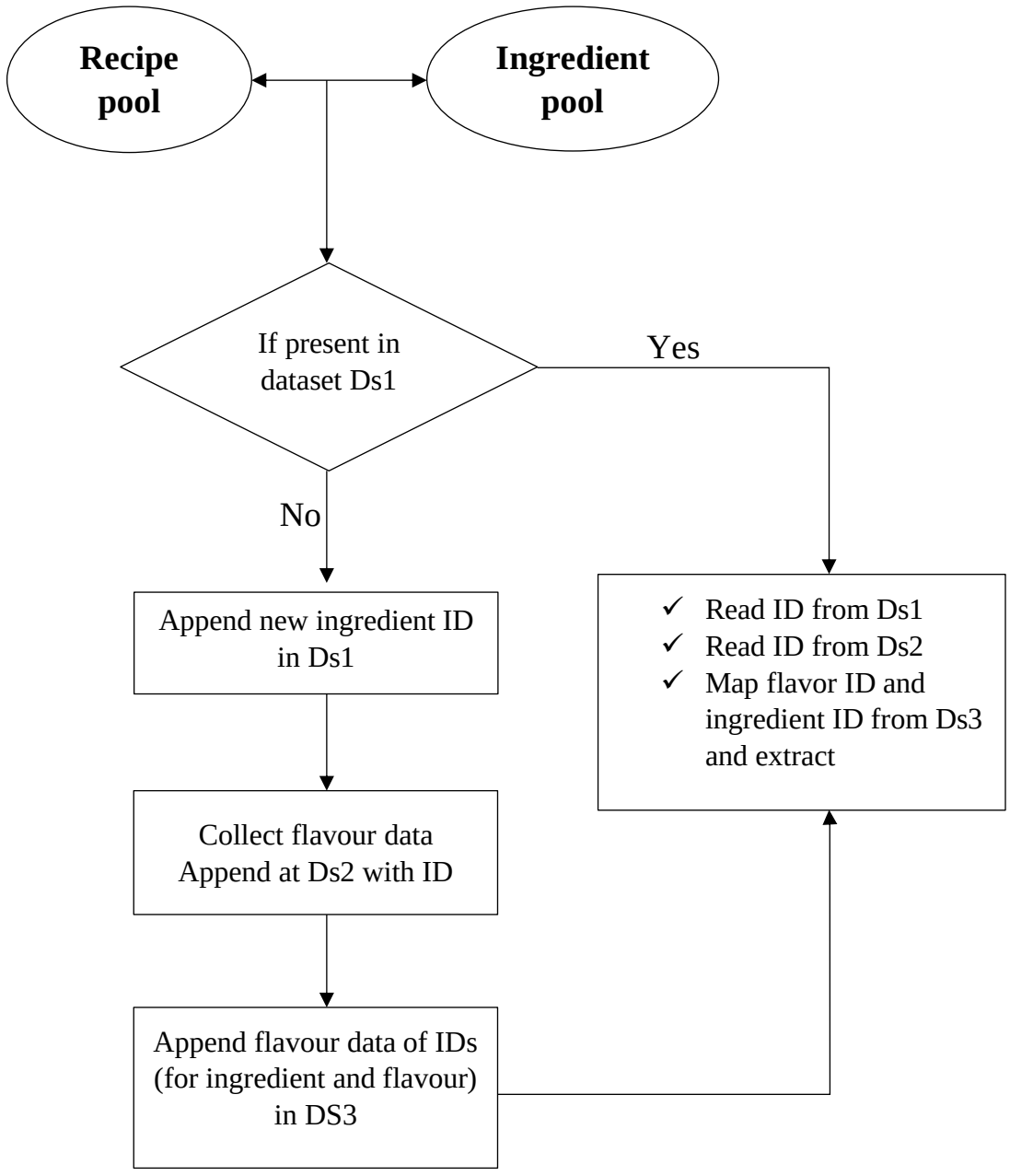

Fig. 3.9 Illustration of data file compilation

3.1.2.2.3 Data analysis

3.1.2.2.3.1 Shared compound hypothesis calculation

The various data analysis was carried out in jupyter notebook with the use of suitable libraries to carry out the computation task. Fig. 3.10 shows the detail illustration of the average shared compound calculation of the real and random recipes.

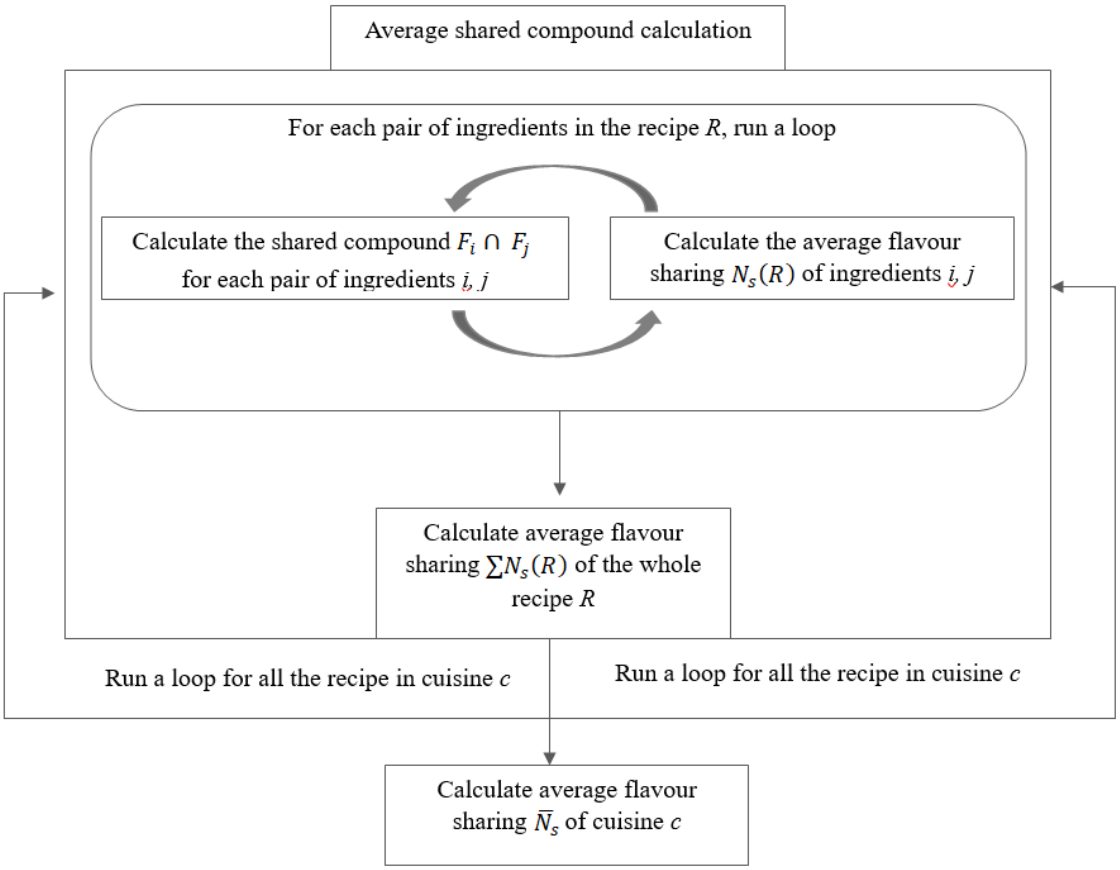

Figure 3.10 Average shared compound calculation for real and random cuisine

3.1.2.2.3.2 Random recipe generation

A reference model of a randomly constructed recipe is generated from a set of ingredients similar to the overall universe of ingredients considering a probability distribution. To examine the mechanism that contributes to food pairing, a set of four random controls was generated from the existing set of recipes (Fig.3.11). The uniform selection of ingredients provided the first model of random control. The second model was generated by selecting an ingredient while keeping its frequency in mind. The third model was generated by considering the ingredient while keeping the category in mind. The final model was generated by choosing ingredients while keeping the category and frequency

in mind. The python libraries used for creating random controls are shown in Fig. 3.12. Additionally, an illustration of a random recipe generation is shown in Fig.3.13.

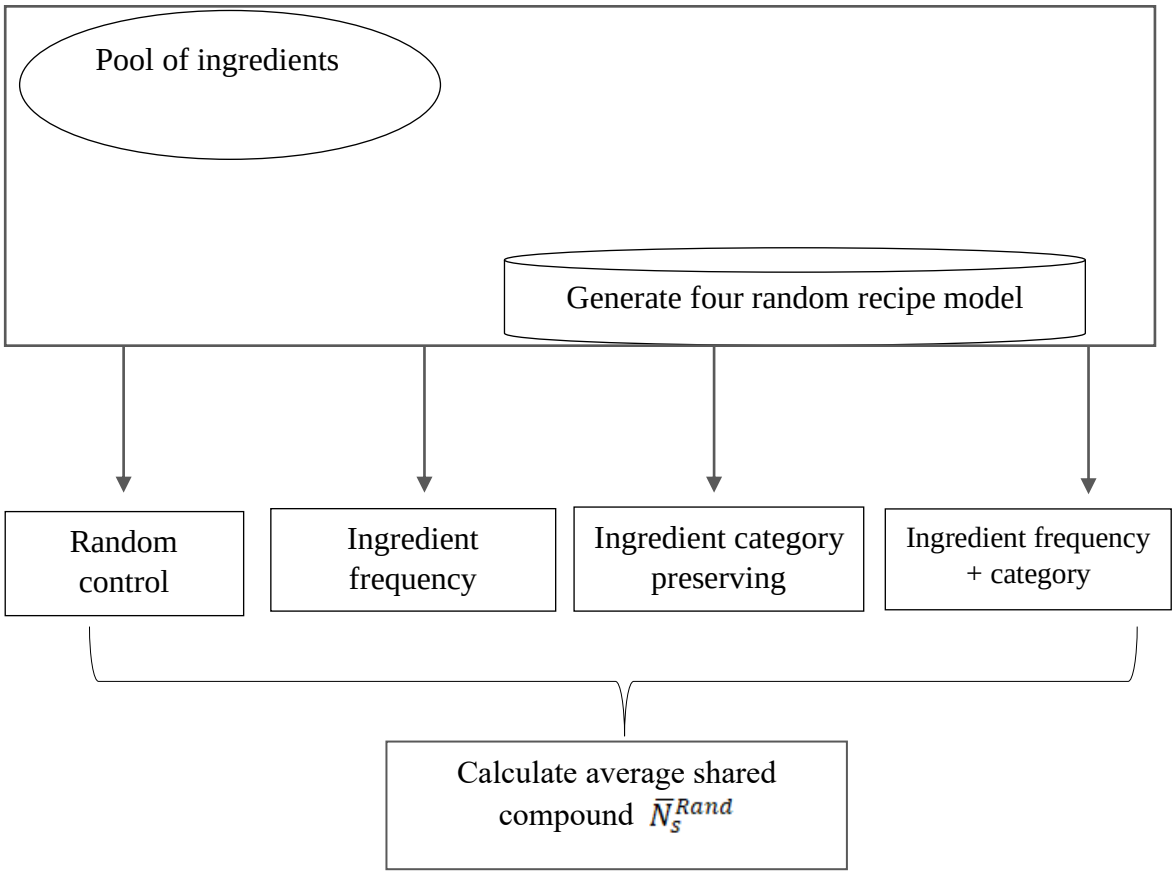

Figure 3.11 Random control generation

```
In [1]: import numpy as np
        import pandas as pd
        import networkx as nx
        import csv
```

Fig. 3.12 Python libraries used for creating random controls

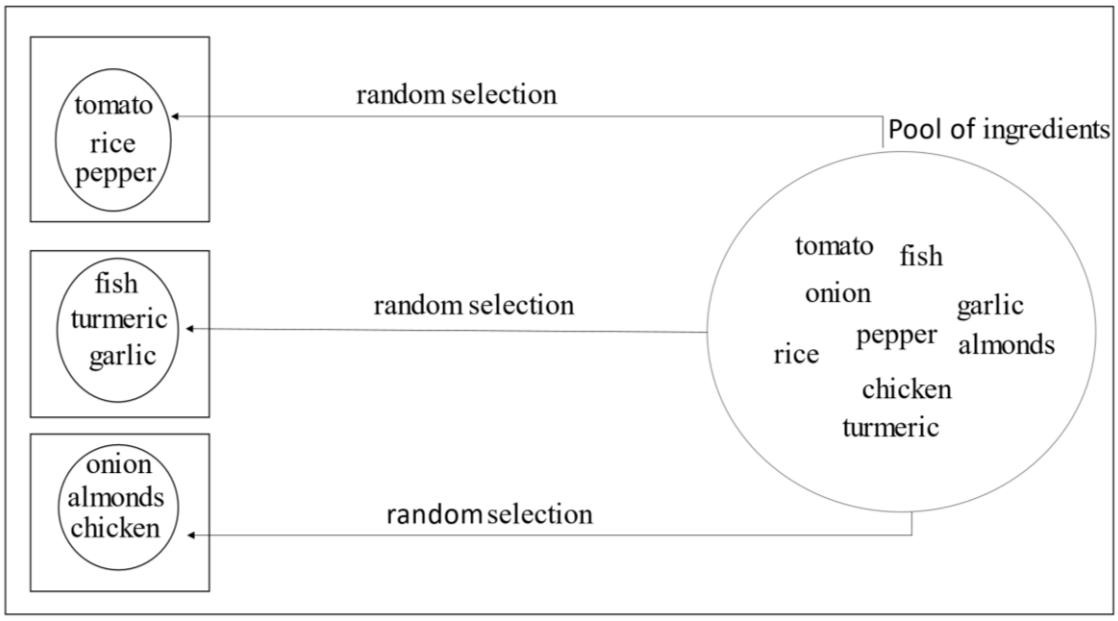

Fig. 3.13 Random recipe generation

3.1.2.2.3.3 Cosine similarity

Firstly, we imported the libraries listed required for the analysis Fig. 3.14,

```
In [25]: '''Make recommendations based on flavor profile

         '''

         import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt
         %matplotlib inline
         import seaborn as sns
         import pickle
         from sklearn.metrics.pairwise import cosine_similarity
```

Fig. 3.14 Python libraries used for cosine similarity analysis

We then convert our recipe text data into a vector using Sklearn. This is calculated using "Term Frequency Inverse Document Frequency (TF-IDF)" which is considered to be an index that measures the importance of words in a collection or corpus of documents. This allows us to compare similarities and differences between text vectors. In this calculation, the most common words are closest to 0 and the most unique words are closest to 1 [93].

Term frequency $tf_{t,d}$ refers to how many times the term *t* appears in document *d*. The term *t* in *d* can be defined by its log frequency weight shown in equation 1, if $tf_{t,d} > 0$

$$W_{t,d} = 1 + \log_{10} tf_{t,d}$$

Else,

$$W_{t,d} = 0$$

A document frequency *dft* is measured by the number of documents or texts containing a term *t*. Then for a given term *t* we can calculate the Inverse Document Frequency (idf) as,

$$idf_t = \log_{10}\left(\frac{N}{df_t}\right)$$

The product of tf and idf results in the weight of term *t*'s tf-idf. After we obtain the data frame our results will be returned once we run the cosine similarity function, example shown in Fig.3.15. This python code is made available in GitHub [47] and with few modifications we used it for our study.

```
In [10]: #normalize flavor matrix with tfidf method
         def make_tfidf(arr):
             '''input, numpy array with flavor counts for each recipe and compounds
             return numpy array adjusted as tfidf
             '''
             arr2 = arr.copy()
             N=arr2.shape[0]
             l2_rows = np.sqrt(np.sum(arr2**2, axis=1)).reshape(N, 1)
             l2_rows[l2_rows==0]=1
             arr2_norm = arr2/l2_rows

             arr2_freq = np.sum(arr2_norm>0, axis=0)
             arr2_idf = np.log(float(N+1) / (1.0 + arr2_freq)) + 1.0

             from sklearn.preprocessing import normalize
             tfidf = np.multiply(arr2_norm, arr2_idf)
             tfidf = normalize(tfidf, norm='l2', axis=1)
             print (tfidf.shape)
             return tfidf
```

Fig. 3.15 Python code for calculating TF-IDF

The cosine similarity between the two recipes may be compared by adding up each encoding vector. As a result, with this method, we can find the recipes which are closest to each other based on flavour similarity. If both recipes are close to 1, they are similar, while if they are close to 0, they are not.

3.1.2.2.3.4 t-Distributed Stochastic Neighbor Embedding (t-SNE)

The disadvantage of linear dimensionality reduction algorithms is that dissimilar data points are placed far apart in a representation that is of a lower dimension. Similar data points must be represented close together to represent high-dimension data on low-dimension manifolds, which is not the case when linear dimensionality reduction algorithms are employed. As a result, we used t-Distributed Stochastic Neighbor Embedding (t-SNE) which is a non-linear algorithm. To identify the structure within the

data t-SNE relies on probability distributions and random walks on neighbourhood graphs. The python libraries used for the analysis is shown in Fig. 3.16.

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        %matplotlib inline
        import seaborn as sns
        from scipy.spatial.distance import pdist, squareform
        from sklearn.manifold import MDS, TSNE
```

Fig. 3.16 Python libraries used for t-SNE analysis

We load the required data frame for the analysis and calculate the regional cuisine similarity based on the t-SNE clustering, Fig. 3.17.

```
In [4]: #take some regional cuisines, tsne clustering, and plotting
        def tsne_cluster_cuisine(df,sublist):
            lenlist=[0]
            df_sub = df[df['cuisine']==sublist[0]]
            lenlist.append(df_sub.shape[0])
            for cuisine in sublist[1:]:
                temp = df[df['cuisine']==cuisine]
                df_sub = pd.concat([df_sub, temp],axis=0,ignore_index=True)
                lenlist.append(df_sub.shape[0])
            df_X = df_sub.drop(['cuisine','recipeName'],axis=1)
            dist = squareform(pdist(df_X, metric='cosine'))
            tsne = TSNE(metric='precomputed').fit_transform(dist)
            print (df_X.shape, lenlist)
            palette = sns.color_palette("hls", len(sublist))
            plt.figure(figsize=(10,10))
            for i,cuisine in enumerate(sublist):
                plt.scatter(tsne[lenlist[i]:lenlist[i+1],0],\
                tsne[lenlist[i]:lenlist[i+1],1],c=palette[i],label=sublist[i])
        plt.legend('sublist')
Out[4]: <matplotlib.legend.Legend at 0x1f983a343d0>
```

Fig. 3.17 Python code for t-SNE clustering

```
In [8]: #select four cuisines and plot tsne clustering with flavor
        sublist = ['Assam','Arunachal','Manipur','Meghalaya','Mizoram','Nagaland','Tripura','Sikkim']
        df_flavor = yum_tfidf.copy()
        df_flavor['cuisine'] = yum_ingr['cuisine']
        df_flavor['recipeName'] = yum_ingr['recipeName']
        tsne_cluster_cuisine(df_flavor,sublist)
        plt.legend(sublist,loc='best', fontsize= 14)
        plt.title('flavor profile',fontsize=24)
        plt.savefig('image/TSNE2.jpg', dpi = 300)
```

Fig. 3.18 Python code for plotting t-SNE in flavour and ingredient space

Further, we select the cuisine and plot t-SNE clustering based on flavour and ingredient similarity across the cuisines shown in Fig. 3.18.

3.2 Methodology to characterize and analyse traditional food recipes for ingredient pairing behaviour

Traditionally, cultural cuisine refers to the food practices associated with any given culture that have been elaborated and transmitted over time [7]. The culinary culture has evolved similarly to the variation in regional languages. Typically, these changes are encoded in the signature ingredient combinations of a cuisine's recipes. The cuisines are analysed as a set of recipes, which itself is a set of ingredients. The analysis is started with an analysis of the ingredient pairing behaviour.

3.2.1 Statistical analysis for ingredient pairing behaviour

The recipes are set of ingredients that has a wider cultural acceptability, and are related to liking of the consumers. Analysis of ingredient pairing is started with an analysis of occurrence of a certain combination of ingredients in recipes.

3.2.1.1 Classification of cuisine

A cuisine is defined by its recipes, ingredients and flavour molecules. As a result, it is possible to uncover underlying patterns in traditional recipes by analysing recipes, ingredients, and relevant features [75]. A preliminary statistical investigation was carried out which involves, the analysis of recipe size and frequency rank distribution of the regional cuisine.

a) *Recipe size:* The recipe size of the cuisine was determined as a preliminary statistical investigation. The average recipe size is the average number of ingredients present in a recipe of the given cuisine or given set of recipes. The average size of the recipe was calculated to understand the basic trend of ingredients used in the cuisine.
b) *Frequency-rank distribution:* This process is carried out to determine if the ingredients used in the cuisine exhibit a uniform pattern. We assigned a ranking to the ingredient based on their decreasing frequency of use in the cuisine. In addition, we determine the degree distribution for the ingredients which is the probability distribution $P_I(k)$ to define and test the likelihood of a random ingredient appearing in k recipes. Further, the complementary cumulative degree

distribution $P_c(k)$ plots were compared to a pure power-law distribution to check if they are similar [44].

$$P_c(k) = 1 - \sum_k P_I(k) \tag{3.1}$$

3.2.1.2 Authenticity of ingredients

Every regional cuisine has its specific ingredients that represent the taste palette of the region. Such ingredients are considered to be uniquely placed in the cuisine which is considered to be authentic. As we explore the authenticity of ingredients across the regional cuisine, they help us to understand the similarities or dissimilarities between the various regional cuisines. We also examine the authenticity of each ingredient (p_i^c), ingredients pair (p_{ij}^c), and ingredient triplet (p_{ijk}^c) based on the frequency with which a particular ingredient appears in a particular cuisine's recipe. Each ingredient's prevalence P_i^c in a cuisine c is defined as,

$$P_i^c = {n_c^i}/{N_c} \tag{3.2}$$

Where N_c is the number of recipes in the cuisine and n_i^c is the number of recipes in the cuisine that contain the particular ingredient *i*. It is And expressed as a fractional value.

The relative prevalence $p_i^c = P_i^c - (P_i^{c'})_{\,c' \neq C}$ determines the authenticity of the ingredient *i* which gives the difference between the prevalence of *i* in a particular cuisine *c* and the overall average prevalence of *i* in all other cuisines [5].

3.2.1.3 Ingredient pairing behaviour

Ingredient pairing analysis is the preliminary process to examine and validate the flavour pairing hypothesis of the regional cuisines. It is carried out to determine if two ingredients have similar flavours based on their flavour overlap. It examines the recipe composition pattern by taking into account the similarities in flavour profiles of its constituent ingredients. This process is carried out to quantify the food pairing pattern of the eight regional cuisines belonging to the Northeast.

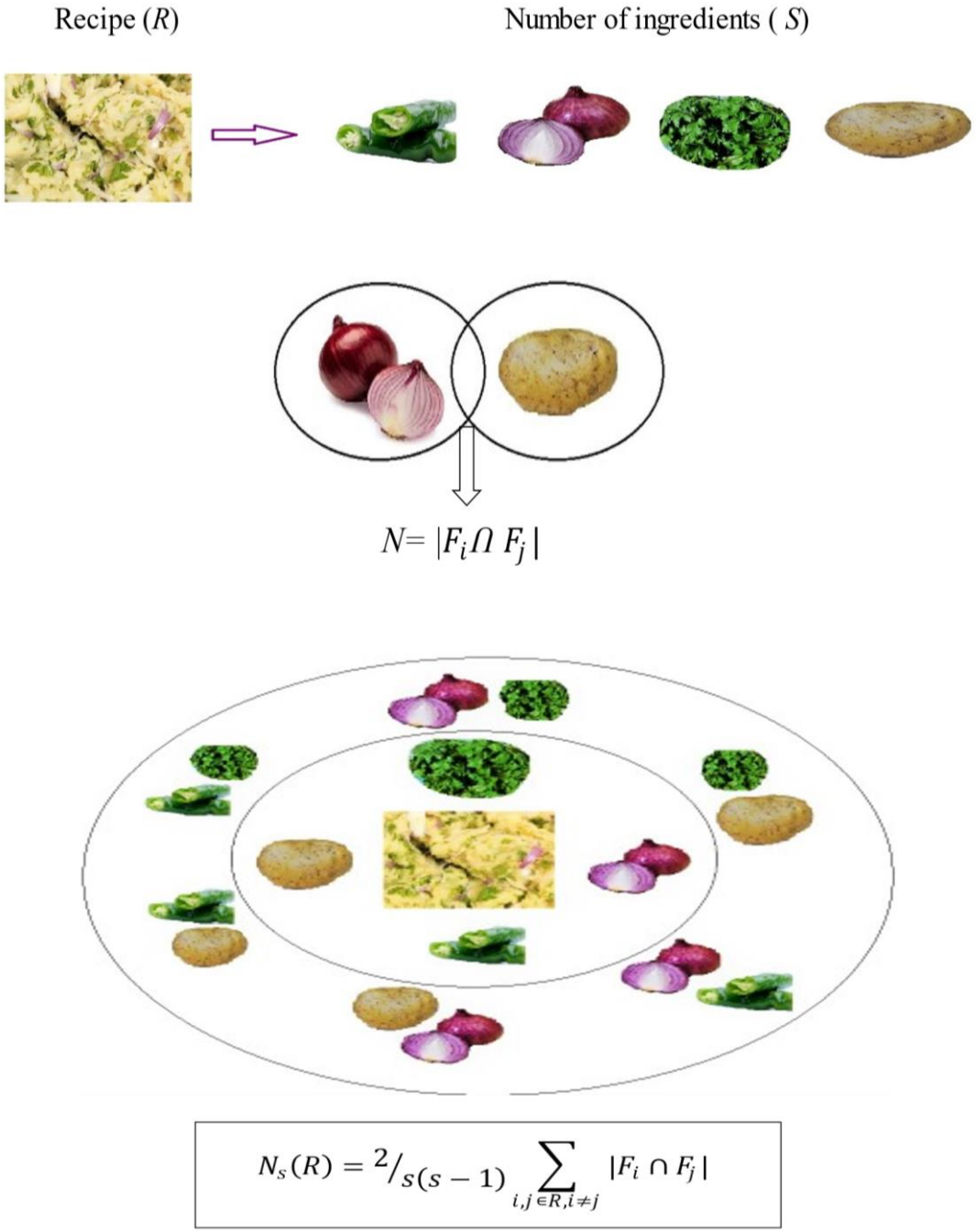

Fig. 3.19 Illustration for calculating average flavour sharing $N_s(R)$ on recipe level

The average number of shared flavour compounds $N_s(R)$ in the cuisine C is first calculated (Eq.3.3), where R indicates a culinary recipe, F_i and F_j represents the set of flavour compounds in ingredients i and j in the recipe R having S number of ingredients. An illustration for calculating the average flavour sharing $N_s(R)$ on recipe level is shown in Fig.3.19.

$$N_s(R) = {}^{2}/_{s(s-1)} \sum_{i,j \in R, i \neq j} |F_i \cap F_j| \tag{3.3}$$

Additionally, for cuisine with N_c recipes we calculated the mean of $N_s(R)$ to estimate the flavour-sharing index $\bar{N}_s$ as to which degree the flavour pairing exists overall (Eq. 3.4). An illustration for calculating flavour sharing index $\bar{N}_s$ on cuisine level is shown in Fig. 3.20

$$\bar{N}_s = {}^{1}/_{N_c} \sum_R N_s(R) \tag{3.4}$$

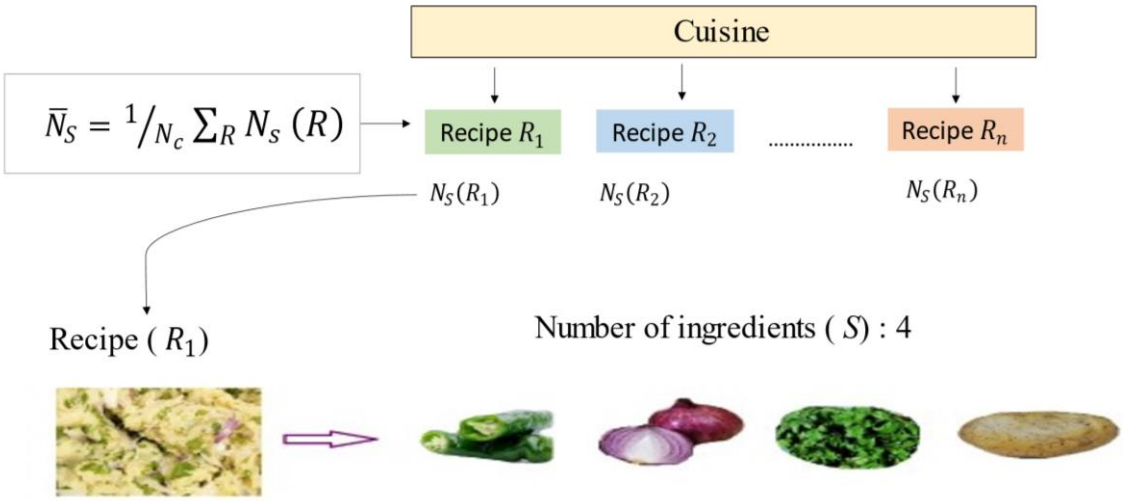

Fig. 3.20 Illustration for calculating flavour sharing index $\bar{N}_s$ on cuisine level

Finally, we compared the difference between $\bar{N}_s^{Real}$, degree of the flavour pairing of the original cuisine (*Real*) and the $\bar{N}_s^{Rand}$, degree of the flavour pairing of randomly constructed cuisine *(Rand*) by calculating ΔN_s (Eq. 3.5) to determine its statistical relevance.

$$\Delta N_s = \bar{N}_s^{Real} - \bar{N}_s^{Rand} \tag{3.5}$$

The variation in ΔN_S measure, if close to null/zero indicates that there is no significant relationship between the recipe and the flavour compounds. If positive, it indicates that the original recipe has a strong influence over the random recipe where ingredients tend

to share more flavour compounds in the recipe validating the shared compound hypothesis. If negative it indicates that the ingredients used in the recipe do not share flavour compounds.

3.2.1.4 Random recipe generation

The shared compound hypothesis was validated, indicative of whether ingredients sharing flavour compounds appear more often in the cuisine. The process involves considering a null hypothesis. Using appropriate null models, we built several random recipe datasets and compared the shared compounds $\overline{N}_s$ between real and randomized recipe datasets to test the accuracy of our results. A reference model of a randomly constructed recipe is generated from a set of ingredients similar to the overall universe of ingredients considering a probability distribution. To examine the mechanism that contributes to food pairing bias, a set of four random controls was generated from the existing set of recipes.

1) *Random model:* The uniform selection of ingredients provided the first model of random control where ingredients were chosen uniformly. The recipe is constructed by selecting at random an ingredient that is used at least once in the cuisine. There are chances of creating recipes with ingredients that are rarely found in random models.
2) *Ingredient frequency model:* The second model was generated by selecting an ingredient while preserving its frequency in mind. As a result, an ingredient is more likely to be selected if it is frequently used in the recipe.
3) *Ingredient category model:* The third model was generated by considering the ingredient while preserving the ingredient category, but without considering the frequency of the ingredients.
4) *Ingredient frequency and category model:* The final model was generated by choosing ingredients while preserving the category and frequency of ingredients to be selected. For example, a recipe with chicken and ginger will contain a meat and a spice. Probabilities of picking an ingredient depend on how prevalent it is in the cuisine.

3.2.1.5 Contribution of ingredients

The analysis of ingredient contribution determines the degree of contribution of each ingredient towards the food pairing effect. It can be obtained by calculating each ingredient's contribution χ_i (Eq. 3.6) to the measure of $\overline{N}_s$ and analysing it as follows,

$$\chi_i = \overline{N}_s(C) - \overline{N}_s\left(C^i\right) \tag{3.6}$$

Here, $\overline{N}_s(C)$ is the degree of the flavour pairing of a given cuisine *C*, and $\overline{N}_s\left(C^i\right)$ is the degree of the flavour pairing of the cuisine *C* without the ingredient of concern *i*. If an ingredient's contribution is positive, removing the ingredient from the cuisine would cause the $\overline{N}_s$ measure to decrease. Whereas, if an ingredient's contribution is negative, removing the ingredient from the cuisine would cause the $\overline{N}_s$ to increase. The differences in the measure of $\overline{N}_s$ results in determining to which degree it affects the overall food pairing behaviour. This defines the pattern of ingredient combinations in the cuisine, the higher the contribution, the more flavour-sharing ingredients are included in the cuisine, and vice versa [60].

3.3 Methodology to apply data-driven similarity analysis for intra- and inter-regional cuisine similarities

3.3.1 Framework for flavour network data analysis

3.3.1.1 Ingredient-compound bipartite network

The ingredient-compound association forms a bipartite network, which consists of two types of nodes with connections only between nodes of different types. The two types of nodes are food ingredients and flavour compounds and connection signifies that an ingredient contains a compound.

3.3.1.2 Extraction of the backbone of flavour network

The backbone extraction method involves the identification of statistically significant links of each ingredient, given the sum of weight characterizing the particular node that has significant connections to others. This method will be carried out to circumvent the high-density flavour network which is hard to visualize.

3.3.1.3 Principle of flavour network theory

3.3.1.3.1 Theory

The flavour network is a weighted network that is formed by the projection of a bipartite network. A bipartite network consists of two kinds of nodes (i) ingredients present in the recipes and (ii) ingredient flavour compound of each ingredient known to contribute to the flavour of ingredients. The construction of a flavour network for the regional cuisine was carried out to validate the food pairing hypothesis to examine whether we combine ingredients that share a significant link or we avoid them? The flavour network consists of a node and edges, in our study each node represents the ingredients and the edges/link represents the number of flavour compounds shared amongst the ingredients. Further, to circumvent the high density of the network we carry out a process of backbone extraction for clear visualization of the network. Only the statistically significant edges/links are retained in the network's extracted backbone (*p*-value 0.07).

3.3.1.3.2 Inference

The general theory behind ingredient blending in foods is the flavour pairing theory, which states that ingredients are more likely to blend when they have common flavours. To prove whether or not we should combine ingredients with a strong link in the flavour network, we need data on ingredient combinations that are commonly accepted and liked by people, which can be found in the form of culinary recipes. Flavour pairing was applied as a fundamental algorithm to identify new ingredient pairings in recipes, develop a recipe recommendation system, and generate new recipes, contributing to the creative domain of culinary arts [5].

3.3.2 Quantifying similarity between cuisines

3.3.2.1 Clustering of regional cuisine

We use t-Distributed Stochastic Neighbor Embedding (t-SNE) for clustering the regional cuisine in the ingredient space and flavour space. This method reduces the dimensionality of a dataset by converting it to a low-dimensional two- or three-dimensional space, which can then be visualized. With t-SNE, we can calculate the probability distribution of the high-dimensional data, in such a way that similar points have a higher likelihood of being selected, while dissimilar points have a lower likelihood of being selected [23].

3.3.2.2 Cosine similarity

Analysis of flavour similarity in regional cuisine helps in finding regional cuisine that shares similar flavour preferences with the selected cuisine [32]. The similarity between the regional cuisine was quantified using cosine similarity. It measures the similarity of different regional cuisine based on flavour preferences. Cosine similarity is projected as a network based on ingredient scores to calculate the similarity of ingredient and flavour preferences between the regional cuisine (Fig. 3.21) [50]. For regional cuisines *a* and *b*, the cosine similarity is represented using a dot product and magnitude as given in Eq. 3.7. It is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them [32]. The value of cosine similarity is bound to a range of 0 to 1. Where 0 means no similarity whereas 1 means both the items are 100% similar.

$$\cos(p^a, p^b) = \frac{\sum_i P_i^a P_i^b}{\sqrt{\sum_i (P_i^a)^2}\sqrt{\sum_i (P_i^b)^2}} \tag{3.7}$$

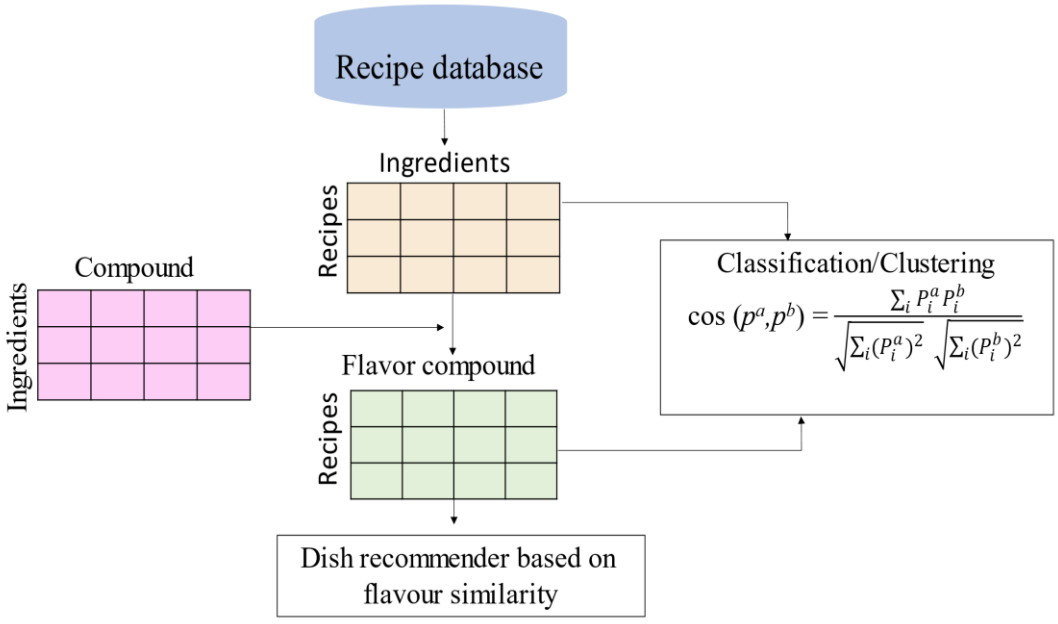

Fig. 3.21 Illustration of analysis based on flavour similarity

3.4 Methodology for generation of recipe compositions based on identified consumer preference for ingredient pairing

3.4.1 Data-driven models

3.4.1.1 Theory

The food pairing theory fails to explain which ingredient makes the best combination with all ingredients present. And it is important to consider the flavour in recipe composition as most of the authentic ingredients in Northeast cuisine are from the spice category where flavours are not so focused on (ingredients do not share much flavour compounds). A data-driven model is built to solve such problems. These models give, for a given set of ingredients, those ingredients that can best be combined with all of the given ingredients. We created two binary matrices consisting of a data file for each regional cuisine. The first matrix consists of a data file containing recipes and their corresponding ingredients. The second matrix consists of the data file containing ingredients and their category, flavour compounds and ingredients with their flavour compounds.

3.4.1.2 Approach

Two models have been applied non-negative matrix factorization (NMF) and the two-step recursive least squares method (RLS) (Fig.3.22). NMF is a decomposition technique that approximates a matrix by a product of two low-rank matrices resulting in the elimination of noise in data. NMF assumes that the data is non-negative and only allows additive combinations to represent the data. This facilitates the interpretation of the representation of data. While RLS bases its suggestion not only on existing profiles of the ingredients but it makes it more convenient to find possible new ingredient combinations. The model was used to complete three-ingredient sets into a recipe.

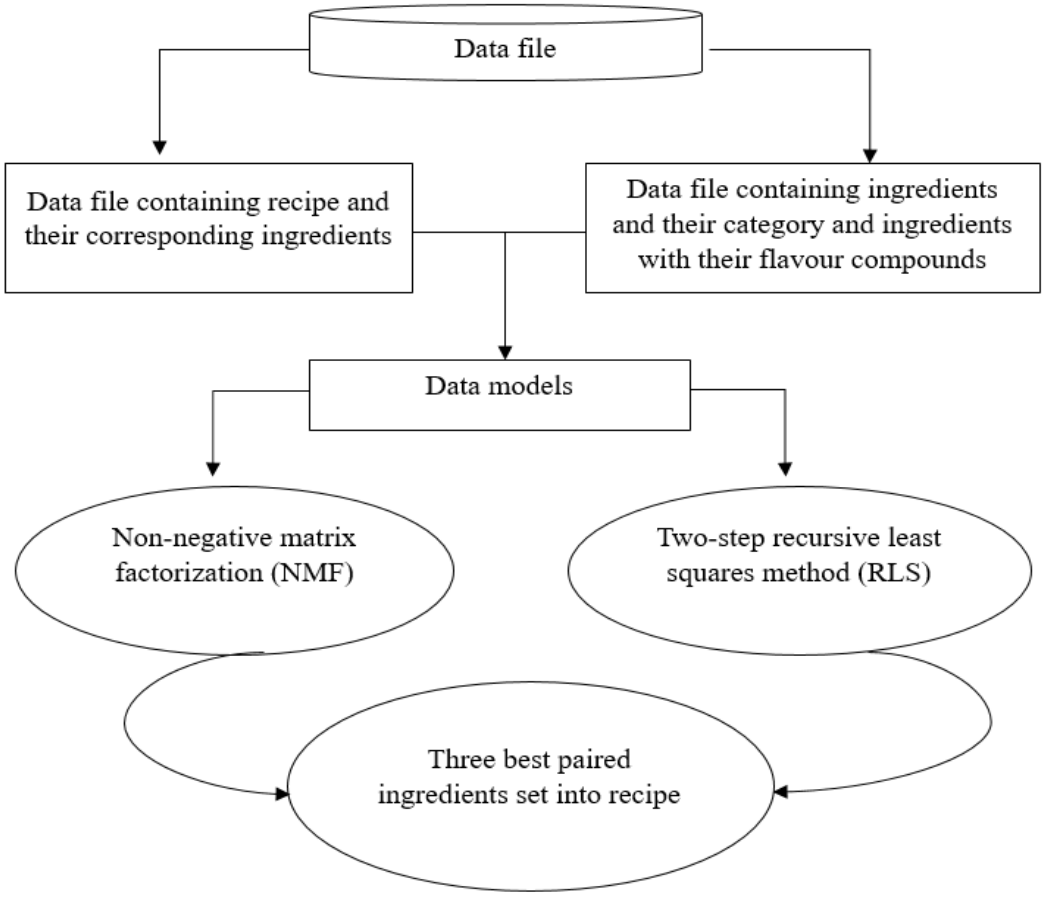

Figure 3.22 Flow chart of the implementation process of the data-driven models for recipe completion

3.4.1.2.1 NMF: This decomposition technique was chosen because the ingredient list was enormous, so we selected the best three ingredients by correlation without a random way. Since it is an unsupervised learning, the ingredient list was taken as input (X). The matrix size was *359*106.* It was then decomposed into two matrices *356*k* and *k*106* using NMF, where k is no. of features. NMF model with 3 components was chosen, and the X data set was transformed. Feature set was then extracted.

3.4.1.2.2 RLS: The dataset was arranged according to the required form. The list of ingredients were the features (X), and their respective recipe IDs (y) were considered as labels. The first dataset consists of 359 recipes from Assamese cuisine. The accompanying ingredients were enumerated while training for the model. The data was in string/character format, so it was scaled in binary format by arranging all of the ingredients (106) as column labels and the recipe IDs as row labels. The recursive model was applied over the binary data. The significance value of $p < 0.05$ was checked. The coefficients associated with ingredients were calculated and using these values, the RLS

model was used for prediction against the triplets chosen using NMF factorization. Each algorithm iteration in Recursive Least Squares analyses a single new data point in order to enhance the estimation of model parameters rather than minimise the least squared error.

3.5 Methodology for developing ingredient combinations for a food product with customized specifications with the application of the flavour network theory

3.5.1 Theory

People with dietary limitations are frequently required to follow dietary restrictions and limit the amount of food they eat as a part of their therapy. Given the range of meals that may be ingested, it is important to take into account that the substances used to make alternatives or supplements have a balance of flavour, taste, and nutritional qualities by creating a flavour profile that is comparable to that of a chosen item.

Using statistical techniques, new ingredient combinations were created that closely resemble the flavour attributes of the original item being replaced. An alternative ingredient recommendation method is presented that takes into account the similarity and compatibility of different ingredients. Using alternative ingredients, a new delicious dish can be created that is similar to an original recipe. Considering two factors, we proposed recommending new additional ingredients when considering whether food can be used as an alternative ingredient or not? Ingredients to be exchanged are checked for similarity and compatibility to be used as alternatives to the items to be replaced. An illustration is shown in Fig. 3.23.

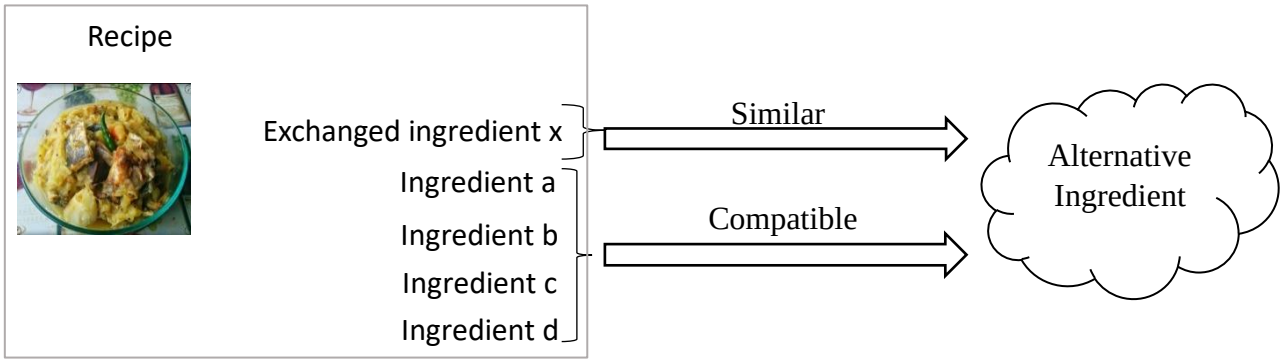

Figure 3.23 General concept for the alternative ingredient selection based on the similarity and compatibility of ingredients

3.5.2 Approach

a) Similarities with the recipe ingredient: A variety of characteristics are found in ingredients, including taste, texture, flavor, nutrition, and color. According to the study, we consider ingredients in the same category share more similarities than those in other categories. It refers to the idea that ingredients of the same category have similar textures and nutrition when compared to those of other categories [70].

b) Compatibilities with other ingredients in the recipe: This study investigates the compatibility of combinations of ingredients with high co-occurrence frequencies. Using the recipe database, we analyze the relative frequencies of two ingredients that are concurrently used in a given recipe to estimate the co-occurrence relation and to determine the principle whether an ingredient can be the alternative-ingredient or not. Equation 1 calculated the compatibility S_{k,R_i} score between the ith ingredient in recipe *R* for all *k* kinds of ingredients.

$$S_{k,R_i} = \left(\prod_{n=0}^{r} \frac{CoOc(k, F_{i,n})}{N_{F_{i,n}}} \right) \times N_k$$

$CoOc(k, F_{i,n})$ shows the frequency of ingredient *k* co-occurred with each ingredient $F_{i,0}F_{i,r}$ in given recipe R_i. Where, $N_{F_{i,n}}$ is the frequency of single ingredient $F_{i,n}$ and N_k is the frequency of the focused ingredient *k*. The co-occurrence frequency is calculated using,

$$\frac{CoOc(k, F_{i,n})}{N_{F_{i,n}}} = \begin{cases} \dfrac{CoOc(k, F_{i,n})}{N_{F_{i,n}}}, & (N_{F_{i,n}} > 0), \\ 1, & (N_{F_{i,n}} = 0), \end{cases}$$

If an ingredient $F_{i,n}$ is not available in the recipe database, then, $N_{F_{i,n}} = 0$ whereas if an ingredient does not co-occur in the recipe database with the focused ingredient, then $S_{k,R_i} = 0$ as $CoOc(k, F_{i,n}) = 0$. However, in such case we represent the data $CoOc(k, F_{i,n}) = 0.0001$ as the infinitesimal score represented as,

$$CoOc(k, F_{i,n}) = \begin{cases} CoOc(k, F_{i,n}), & (CoOc(k, F_{i,n}) > 0), \\ 0.0001, & (CoOc(k, F_{i,n}) = 0), \end{cases}$$

Finally, we calculate $GM_{x\in k,R_i}$ based on compatibility scores $S_{x\in k,R_i}$ of all ingredients within the recipe database considering that an ingredient *x* occurs with ingredient $F_{i,n}$.

$$GM_{x\in k,R_i} = \frac{S_{x\in k,R_i}}{\sum_{k=0}^{k} S_{k,R_i}}$$

3.6 Chapter summary

The recipe data information was gathered from a variety of sources, including the internet and cookbooks, which are frequently used in commercial food service establishments. The statistics of the shared compound hypothesis were estimated to determine the food pairing behaviour and whether foods that share flavour compounds appear more frequently in the cuisine. The analysis of ingredient contribution was also carried out to determine the degree of contribution of the ingredient towards the food pairing effect and to offer explanations of the preference based on flavour pairing theory. The flavour network was used to project recipes to the flavour space and to find similar dishes based on similar flavour profiles. Similarities among the regional cuisines were explored using a clustering algorithm in the ingredient space and flavour space. t-SNE clustering of the Northeast regional cuisine for comparison with Indian cuisines and cuisines from other countries (*American, Italian and Chinese*) and to visually examine whether there is any overlap in the choice of ingredients and the flavour profile. The cosine similarity analysis, which is based on determining the similarities and compatibility of ingredients categories and the relative frequency of ingredients, is used for the quantification of the similarity between the two recipes. It estimated the difference in food choice across the regional cuisine which is the result of the differences in flavour preferences. A few authentic dishes of each regional cuisine were selected and analysed for similarity across the other regional cuisines in terms of similarity in the pattern of ingredient usage. For the generation of new recipes from a given set of ingredients, a data-driven approach was applied. The search was based on three sets of the best-paired ingredient for each regional cuisine, selected based on co-occurrence and the shared flavour compounds. RLS models recommend ingredients which can be paired with the three sets of ingredients forming a recipe. The RLS model was found to be consistent in recommended ingredients from the same categories itself.

For the recommendation of alternative ingredients for an existing recipe, a statistical technique, that considers the similarity and compatibility of different ingredients, was used. Considering these two factors i.e., the similarity with the recipe ingredient and compatibilities with other ingredients, new additional ingredients/ alternative ingredients were recommended considering whether food can be used as an alternative ingredient or not. One of the criteria incorporated is that ingredients of the same category have similar textures and nutrition when compared to those of other categories. The relative frequencies of two ingredients that are concurrently used in the given recipe were analysed to estimate the co-occurrence relation and to determine the principle of whether an ingredient can be an alternative ingredient or not. The modified recipes were validated through a comparison with pre-existing recipes considering the flavour properties, signifying the utility of the approach.