BTP Report: Compilation of flavor molecules and their taste-based classification

Avneet Kaur, under the supervision of Dr. Ganesh Bagler,IIIT D

I. INTRODUCTION

Flavor of a molecule is defined in terms of the gustatory and olfactory sensation during tasting. The analysis of flavor profiles has various applications in the food industry, such as synthesis of artificial flavors, enhancing flavors in food and checking of adulteration in food. In this project, an attempt at taste based classification, to understand taste, an aspect of flavor, was made, to understand and analyze the taste and flavor profiles of various compounds based on physico-chemical and ADMET properties.

II. OBJECTIVE

A. Bitter v/s Non-Bitter classification

Here the aim was to develop a binary classifier which can classify whether a compound tastes bitter or not based on certain physico-chemical and ADMET properties and identifying the chemical descriptors important for classification of bitter compounds.

B. Clustering of FEMA Flavor terms

Given a set of molecules along with their FEMA Flavor terms and various 2D/3D properties, to find if theses flavor terms can be clustered together based on their co-occurrence in these molecules and to find if these molecules could be clustered together based on their flavor terms.

III. BITTER V/S NON-BITTER CLASSIFICATION

A. Dataset Curation

The dataset for this problem was mainly collected from the following sources:

1) Positive Set:

- *BitterDB*[1]: It is database of compounds that have bitter taste or are known to activate at least one human bitter receptor. It includes structurally diverse compounds such as ions, peptides, alkaloids, polyphenols, glucosinolates and more. The bitter molecules were mainly collected from this database.
- *Rojas et al.*[2]: Additional 81 bitter molecules were collected from this study.

2) Negative Set: The negative set comprises of non bitter molecules like tasteless molecules, sweet molecules and others compiled from the sources below:

- *Fenarolis handbook of flavour molecules*[3]: A total of 1753 molecules were taken from Fenarolis handbook of flavor molecules here. Compounds were considered as non-bitter if the word bitter did not appear in its description.
- *Rojas et al.*[2]: Sweet and tasteless subsets were compiled from sweet (435) and tasteless (133) compounds recently reported by this study.

Type	Source	Number
Positive set	Bitter DB[1]	686
Positive set	Additional bitter :(Rojas et al. [2])	81
Negative Set	Non-bitter: Fenarolis handbook of flavour molecules[3]	1753
Negative Set	Sweet : (Rojas et al. [2])	435
Negative Set	Tasteless:(Rojas et al. [2])	133

TABLE I
PERFORMANCE WHEN 12 PHYSICO-CHEMICAL PROPERTIES WERE USED

B. Properties used for analysis

- *Physicochemical properties*: The following 12 physicochemical basic properties were used : Molecular weight (MW), lipophilicity (ALogP, the atomic LogP), rotatable bonds count (RB), polar surface area (PSA), electrotopological states (estate), molecular refractivity (MR), molecular polarizability (Polar), hydrogen bond acceptor (HBA), hydrogen bond donor (HBD), rings count (ring), chiral centers count (chiral) and heavy atoms count (HA).
- *Admet Descriptors*: Due to the widely assumed connection between bitterness and toxicity, ADMET (absorption, distribution, metabolism, excretion and toxicity) descriptors from the QikProp[4], Ligprep[5] and Canvas[6] packages were used. The QikProp[4] package predicts physically and pharmaceutically significant properties of organic molecules based on the full 3D

molecular structure. In all 12 physico-chemical properties + 47 additional molecular descriptors extracted from Qikprop[4].

C. Exploratory Data Analysis

The exploratory data analysis was performed for first, the 12 physiochemical properties and then these 12 properties + 47 ADMET descriptors. BitterPredict:Bitter or not[7] was referred to for the methods and data.

1) Visualization: A variety of techniques were used for visualizing the higher dimensional data in a lower dimensional space. The results have been summarized below along with the plots.

- *PCA*
  Principal component analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. It does so, by finding the principal component of maximum variance in the data, and then representing the data along those components. The plots for explained variance and pca for both cases are given below. It can be seen that 55 percent and 18 percent variance is explained by the first two principal components.(Fig. 1)

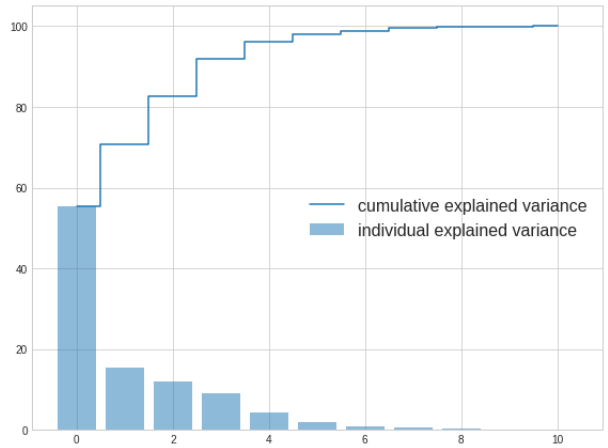

Fig. 1. PCA for 12 physico-chemical properties.

It can be seen in the Fig 2. that certain points superimpose on each other (Bitter, and Non Bitter molecules). Non Bitter molecules in Red are scattered around. Also, the data doesnt appear to be well separable.
These plots were also plotted for 12 + 47 ADMET descriptors.Variance explained by first and second principal component are 36 percent and 18 percent.
Since PCA is a linear algorithm, it will not be able to interpret complex polynomial relationship between the various features. Therefore, some other visualization techniques like TSNE(t-distributed stochastic neighbour embedding), ICA(Independent component analysis) and LDA(linear discriminant analysis) were used to embed the high dimensional data in lower dimensional space for visualization purposes and comparison.

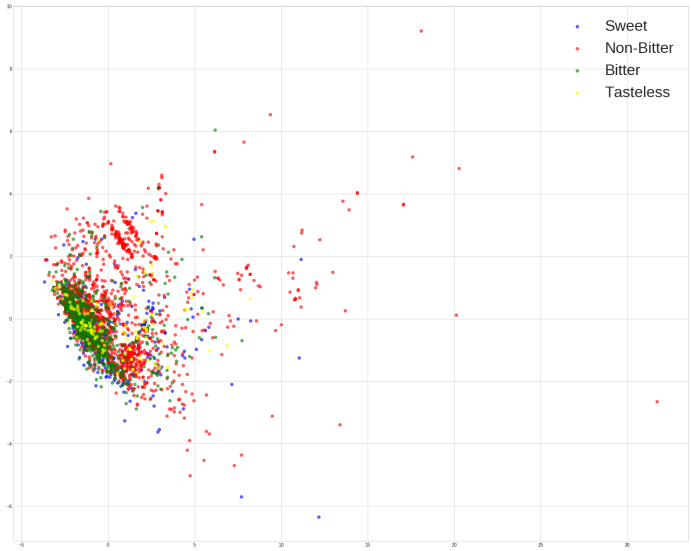

Fig. 2. PCA for 12 physico-chemical properties

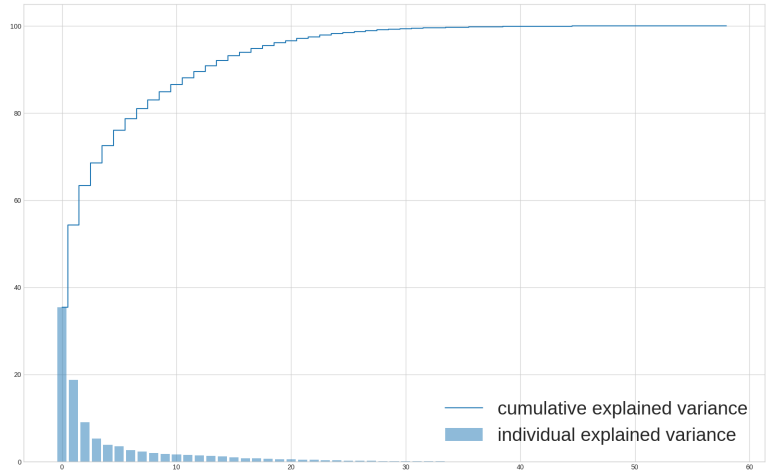

Fig. 3. Explained variance for 12 physico-chemical properties + 47 ADMET descriptors

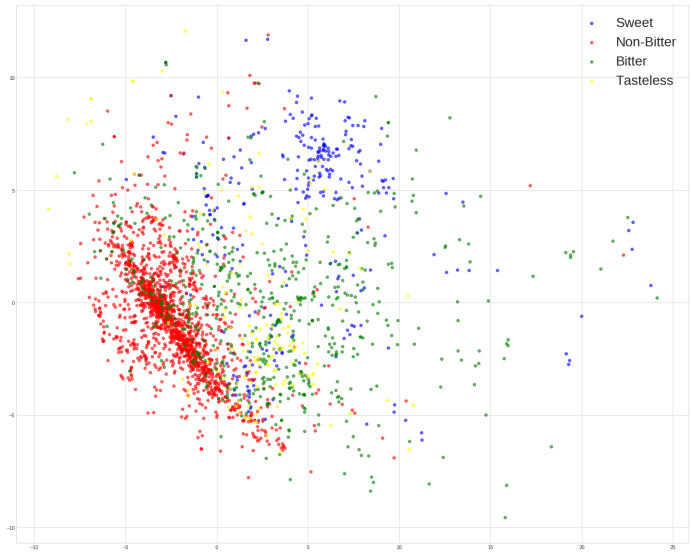

Fig. 4. PCA for 12 physico-chemical properties + 47 ADMET descriptors

- *LDA*
  The goal of Linear Discriminant analysis(LDA) is to project a dataset onto a lower-dimensional space with good class-separability in order avoid over fitting and also reduce computational costs. Both PCA and LDA are commonly used dimensionality reduction techniques, however PCA is an unsupervised algorithm and it ignores the class labels and aims at finding the direction of the maximum variance. LDA on the other hand is supervised and computes the directions (linear discriminants) that will represent the axes that that maximize the separation between multiple classes.

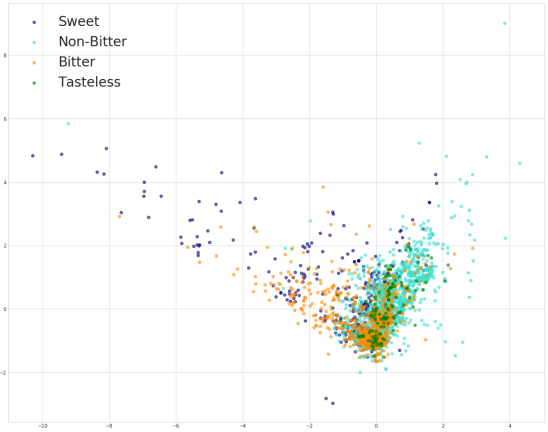

Fig. 5. LDA for 12 physico-chemical properties

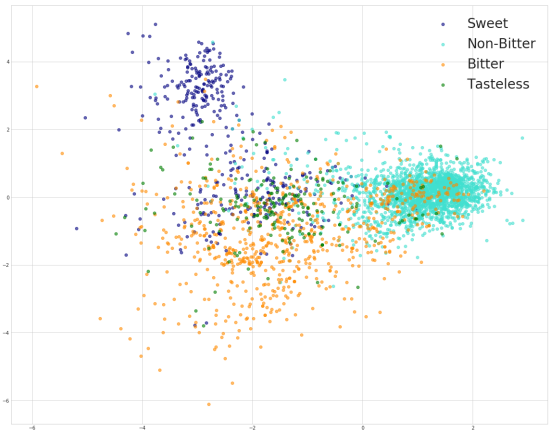

Fig. 6. LDA for 12 physico-chemical properties + 47 ADMET descriptors

- *ICA* Independent component analysis (ICA) is used to estimate sources given noisy measurements. For example: If we have 3 instruments playing simultaneously and 3 microphones recording the mixed signals , ICA is used to recover the sources i.e. what is played by each instrument. ICA was used here to see if a better visualization could be obtained.

- *TSNE*
  T-SNE (t-distributed stochastic neighbour embedding) is based on probability distributions with random walk on neighborhood graphs to find the structure within the data, which enables it to recognize complex non linear structures. It is based on a local approach, which tries to place similar points in high dimensions closer to each other in lower dimensions, i.e., it tries to preserve local structure. Perplexity value indicates how many neighbours of a point the algorithm takes into account while performing the dimensionality reduction for visualization. It can be seen that TSNE in case of 12 + 47 ADMET descriptors provides a clear separation between different classes (Fig 10 and Fig 11) , since there is no superimposition as opposed to PCA.

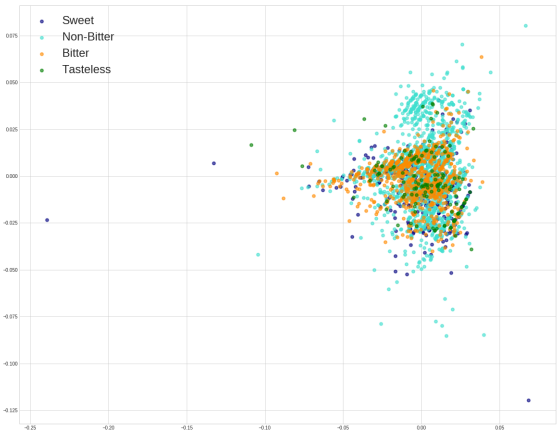

Fig. 7. ICA for 12 physico-chemical properties

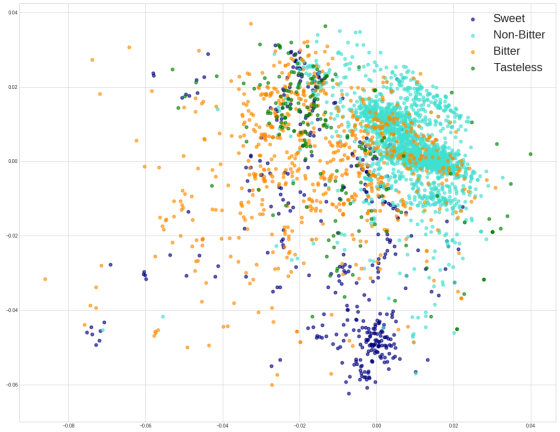

Fig. 8. ICA for 12 physico-chemical properties + 47 ADMET descriptors

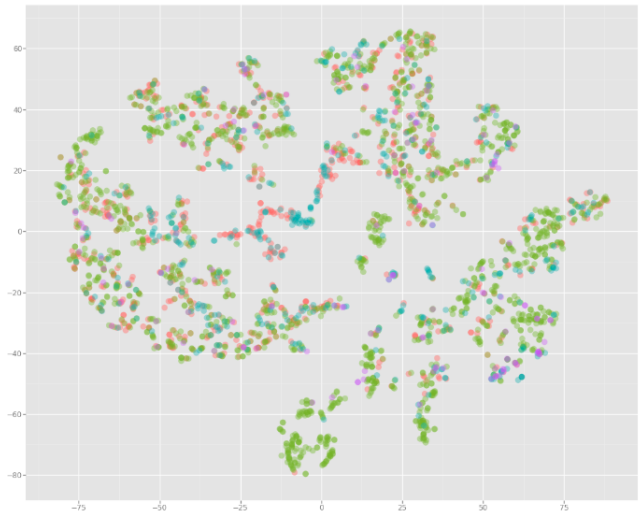

Fig. 9. TSNE for 12 physico-chemical properties. *(Bitter:Green , Non-Bitter:Pink, Tasteless:Violet , Sweet:Blue)*

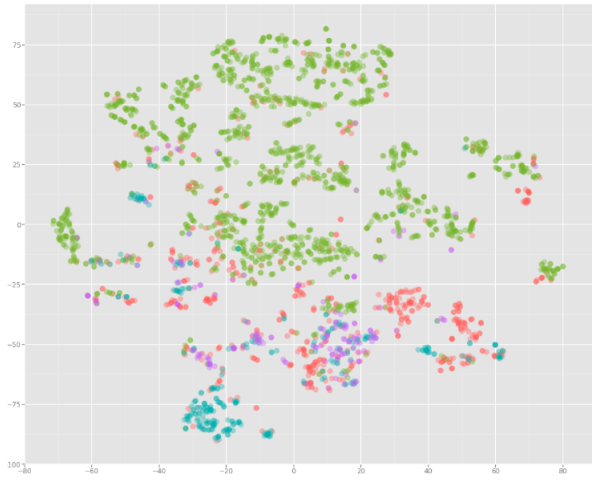

Fig. 10. TSNE for 12 physico-chemical properties + 47 ADMET descriptors *(Bitter:Green , Non-Bitter:Pink, Tasteless:Violet , Sweet:Blue)*

D. Model for classification of bitter vs non bitter

The data was split into a stratified 70 percent training set and 30 percent test set. The stratified split preserves the proportion of different flavors (bitter,sweet,tasteless,non-bitter) in the train and test data in comparison to the original data. Then an ensemble learning method namely, a Random Forest classifier was used for classification of molecules into bitter vs non bitter. Random forests,combine several models to make one single prediction model. They use an ensemble of decision trees, for prediction. The model was trained using the 70 percent training set and a 10 fold cross validation was performed, after which the performance as tested, which is indicated below.

The following measures were calculated: Specificity was calculated as $TN/(TN + FP)$. Sensitivity was calculated as $TP/(TP + FN)$. Accuracy was calculated as $(TP + TN)/(Pos + Neg)$. The 12 physico-chemical properties are themselves not sufficient for the bitter vs non bitter classification problem, since the accuracy reported was very low. The physicochemical properties do not give us much insight into predicting the flavor of a particular compound. Also, this model works poorly on classification of bitter molecules since the sensitivity reported was very low.Therefore, now the analysis was redone using these 12 + 47 ADMET properties and the performance can be seen in Table 2.

Set	Pos	Neg	TP	FP	TN	FN	Specificity	Sensitivity	Accuracy
Training	517	1633	491	10	1633	26	0.966	0.94	0.97
Test	249	678	69	93	585	180	0.86	0.27	0.70
Sweet				26	98		0.79		
Tasteless				18	50		0.73		
NonBitter				7	600		0.98		

TABLE II

PERFORMANCE WHEN 12 PHYSICO-CHEMICAL PROPERTIES WERE USED

Set	Pos	Neg	TP	FP	TN	FN	Specificity	Sensitivity	Accuracy
Training	508	1876	350	65	1811	158	0.96	0.68	0.906
Test	241	781	143	38	743	98	0.95	0.59	0.86
Sweet				13	93		0.87		
Tasteless				18	50		0.73		
NonBitter				7	600		0.98		

TABLE III

PERFORMANCE WHEN 12 PHYSICO-CHEMICAL + 47 ADMET PROPERTIES WERE USED

The second model (Table 2) performs better than the earlier one in classifying bitter vs non-bitter molecules giving 10 fold cross validation accuracy of 86 percent. The 12 properties werent sufficient for classification but some of the properties from the molecular descriptors were successful in this classification problem. The specificity on train and test data was respectively 0.96 and 0.95 and sensitivity was 0.68 and 0.59 respectively. Specificity indicates the negative samples classified correctly whereas sensitivity indicates the positive samples classified correctly.The overall accuracy was 86 percent. This analysis shows that the model offers a better classification as compared to the previous model. However, it is able to classify the non bitter molecules better than the bitter ones as indicated by the specificity and sensitivity values.

E. Parameter Tuning for Random Forest Classifier

In order to find the optimal number of trees to be used for classification model a curve was plotted between error rate of the prediction model and the number of trees used. This is represented in Figure 12 and Figure 13. The number of features considered for the split at each step in the decision trees were also considered (log (number of features) in the

first case(orange curve), square root(number of features) in the second case(blue line) and all features represented by the green curve).Trees are added gradually at each iteration, till the error rate stops to fall. Around 250 trees seemed to be optimal as inferred from the graph when all the features were used.

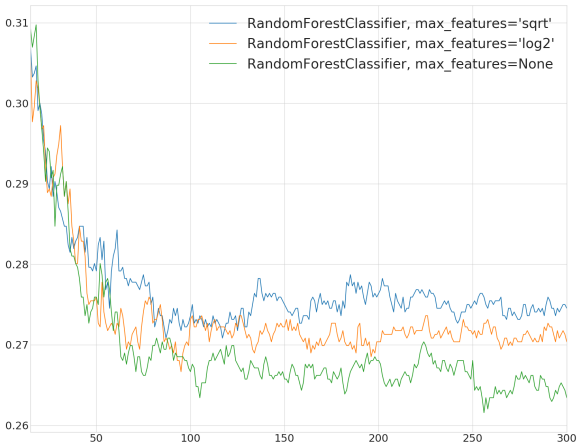

Fig. 11. Error rate(y-axis) v/s Number of trees (x-axis) when 12 physico-chemical properties were used.

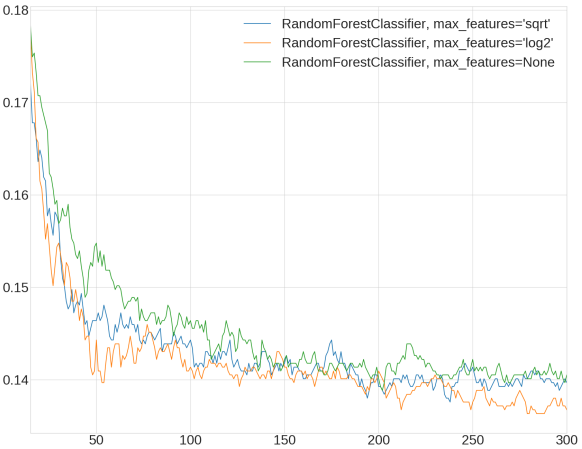

Fig. 12. Error rate(y-axis) v/s Number of trees (x-axis) when 12 physico-chemical + 47 ADMET properties were used.

F. Feature Importance

The contribution of features towards the classification model is indicated by the graph in Fig 16.

Some of the important contributors are: FOSA : 0.06 , PISA: 0.04 , dip ^{2}/V : 0.04, ACxDN $^{.}$5/ SA : 0.04, QPlogPw: 0.04 , QPlogBB : 0.06 , QPlogKp:0.04 , IP(eV): 0.08 .

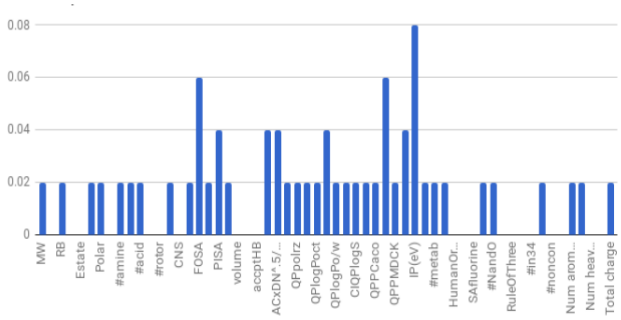

Fig. 13. Importance(y-axis) v/s features (x-axis) when 12 physico-chemical properties + 47 ADMET properties were used.

G. Discussion and Inferences

There was an improve in accuracy and sensitivity from the first to the second model, however the sensitivity in the second model is still quite low (60 percent). So, this model can be improved by using some other classification techniques. The classification of non-bitter molecules, is relatively better, and it can be inferred, that this model may be useful in classification of non bitter molecules but not the bitter ones.

IV. Clustering of FEMA Flavor terms

A. Aim

The motivation was to find whether FEMA flavor terms of given set of molecules can be clustered into different groups based on their frequency of co-occurrance and basic structural (2D, 3D) properties. The aim was also to find essential molecular features for major clusters or groups of flavor terms so formed.

B. Dataset

The data was extracted from Flavour DB[8] version 2.0. For each compound, the following information was available- pubchem id, iupac name, common name, smile, molecular weight,num H donors,number of H acceptors,number of rotatable bonds, complexity, topological polar surface area,monoisotopic mass, exact mass, xlogp3, formal charge, heavy atom count, defined atom stereocenter count, undefined atom stereocenter count, defined bond stereocenter count,undefined bond stereocenter count, isotope atom count,covalently bonded unit count, cas id , fema number, fema flavor profile , odor. There are a total of 2229 compounds out of which 1629 of them have one or more FEMA flavor terms available. These molecules for which flavor information is available were used for experimental analysis.

C. Preprocessing

The FEMA flavour terms were checked manually, spelling errors removed. (eg: Fragrant changed to fragrant).Then, all the @ symbols were replaced with (space) so that later the data in the rows could

be split on the basis of space for processing and separated into different flavour terms. After all the preprocessing steps, 248 unique flavour terms were there.

D. Exploratory data analysis

A graph between the unique flavor terms and the frequency of occurrence was plotted.

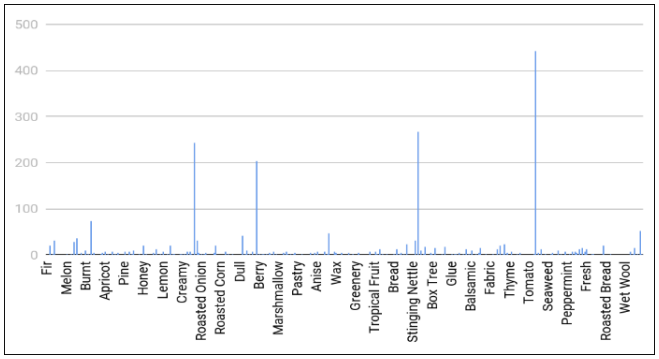

Fig. 14. Frequency of FEMA terms.

A co occurrence matrix was created where each cell indicated the frequency of co occurrence of two FEMA terms together normalized y some measure. Measures from different groups of the paper titled A Survey of Binary Similarity and Distance Measures [9] were implemented for normalization purposes. Experiments ere done using the following measures: A= Number of compounds having flavor FEMA term 1 (FT1) B= Number of compounds having flavor FEMA term 2 (FT2) a= A.B (both present) (intersection) b= (A complement)B (A absent B present) c= A(B complement) (A present B absent) d= (A complement) (B complement) Jacquard Similarity : $(a)/(b + c + d)$ Manhattan distance : $(b + c)$ Maximum of the frequency occurrence of two flavour terms.

E. Clustering

The aim of these experiments was to find meaningful clustering for the various flavour terms and group them 4 types of clustering methods of used:

- *KMeans*
  This is a standard clustering algorithm. The data is in the form of a vector matrix with rows and columns as FEMA terms with each cell denoting the normalized number of molecules in which both those FEMA terms co-occur.Euclidean distance is used as a metric.We seek to minimize a Euclidean distance between the cluster center and the members of the cluster. The intuition behind this is that the radial distance from the cluster-center to the element location should ”be similar” for all elements of that cluster. Convergence criteria is sum of squared errors, when the sum of the squared distances between clusters stops changing or is minimal, we say it has converged. When this algorithm was used to perform clustering with the different normalization metrics, one giant cluster was obtained and the rest of the data was broken into smaller clusters of size 1 or 2 each. The algorithm was run for cluster sizes starting from 1,2 and so on upto 30. However, no meaningful clusters could be obtained.

- *Spherical KMeans*
  This algorithm is similar to k-means except that it uses cosine similarity as a measure rather than the euclidean distance, because due to the curse of dimensionality in higher dimensions, the euclidean distance measure fails to work. However using this method too meaningful clustering , qualitatively and quantitatively was not obtained.

- *Spectral clustering*
  This algorithm tries to find non linear patterns in data and uses affinity matrix (similarity matrix). Using jaccard similarity and spectral clustering, qualitatively few FEMA terms originally having a higher score in the normalized matrix were occurring in the same cluster. However, looking at the clusters formed, meaningful information could not be derived.

- *Density Based Clustering*
  The intuition behind this clustering algorithm is that clusters are dense regions in the data space,separated by regions of lower object density. A cluster is defined as a maximal set of density connected points. When used for clustering no matter what measure was used this algorithm wasnt able to form any clusters which indicated that it couldnt find significant dense regions to form a cluster and hence formed one giant cluster using all the molecules.

F. Discussion and Inferences

No particular clustering measure could segregate the FEMA flavor terms into well formed clusters with minimized within cluster and maximized intra-cluster similarity. The four measures used gave different results with the four clustering algorithms and it was concluded that co-occurrence flavor terms may not necessarily signify any correlation between the flavor terms, that is, we cannot assume linear relationships between co occurrence FEMA terms and them belonging to the same class flavor . Moreover, a good measure of quantifying the quality of clusters that were obtained could not be found and the notion of a good clustering could not be defined.

ACKNOWLEDGMENT

I would like to thank my supervisor, Dr. Ganesh Bagler, for his continuous guidance and support. I am grateful for his constant encouragement, which helped me progress and

make my best attempt at achieving the desired goals. Apart from Dr. Ganesh, I would also like to thank Mr. Rudraksh Tuwani(Research Assistant, Centre for Computational Biology, IIITD), for his continued guidance through the course of this project.

REFERENCES

[1] A. Wiener, M. Shudler, A. Levit, and M. Y. Niv, "BitterDB: A database of bitter compounds," *Nucleic Acids Research*, vol. 40, no. D1, 2012.

[2] C. Rojas, D. Ballabio, V. Consonni, P. Tripaldi, A. Mauri, and R. Todeschini, "Quantitative structureactivity relationships to predict sweet and non-sweet tastes," *Theoretical Chemistry Accounts*, vol. 135, no. 3, pp. 1–13, 2016.

[3] G. A. Burdock, *Fenaroli's handbook of flavor ingredients*, 2010.

[4] Schrodinger, "Qikprop, schrdinger, llc, new york, ny, 2017." [Online]. Available: https://www.schrodinger.com

[5] ——, "Ligprep, schrdinger, llc, new york, ny, 2017." [Online]. Available: https://www.schrodinger.com

[6] ——, "Canvas, schrdinger, llc, new york, ny, 2017." [Online]. Available: https://www.schrodinger.com

[7] A. Dagan-Wiener, I. Nissim, N. Ben Abu, G. Borgonovo, A. Bassoli, and M. Y. Niv, "Bitter or not? BitterPredict, a tool for predicting taste from chemical structure," *Scientific Reports*, vol. 7, no. 1, p. 12074, 2017. [Online]. Available: http://www.nature.com/articles/s41598-017-12359-7

[8] N. Garg, A. Sethupathy, R. Tuwani, R. NK, S. Dokania, A. Iyer, A. Gupta, S. Agrawal, N. Singh, S. Shukla, K. Kathuria, R. Badhwar, R. Kanji, A. Jain, A. Kaur, R. Nagpal, and G. Bagler, "FlavorDB: a database of flavor molecules," *Nucleic Acids Research*, no. November, pp. 1–7, 2017. [Online]. Available: http://academic.oup.com/nar/article/doi/10.1093/nar/gkx957/4559748

[9] C. Seung-Seok, C. Sung-Hyuk, and C. C. Tappert, "A Survey of Binary Similarity and Distance Measures." *Journal of Systemics, Cybernetics & Informatics*, vol. 8, no. 1, pp. 43–48, 2010.