IIITD

Compilation of flavor molecules and their taste-based classification

-Avneet Kaur , under the supervision of Dr. Ganesh Bagler

INTRODUCTION

In this project, an attempt at taste based classification of molecules was made to understand and analyze the flavor profiles of various molecules. 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, one of them being synthesis of artificial flavors, enhancing the flavor in food, checking of adulteration in food. Flavor analysis is also used as part of quality control, and can inform quality specifications for a product. This may be particularly important when considering how a product's flavor changes as it ages. In this project, an attempt to understand flavor in terms of various properties was made.

AIM

The aim of this project was two fold:

1. To develop a binary classifier which can predict whether a compound tastes bitter or not based on certain physiochemical and ADMET properties.Then,To identify chemical descriptors important for classification of bitter compounds.
2. 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. Then, to find if these molecules could be clustered together based on their flavor terms.

METHODOLOGY & PIPELINE

I. CLASSIFICATION OF BITTER VS NON BITTER

Dataset

Positive Set:

The positive set comprises of bitter compounds collected from BitterDB [1] (A database of bitter compounds.) and from the study of Rojas *et al*[2].

Negative Set:

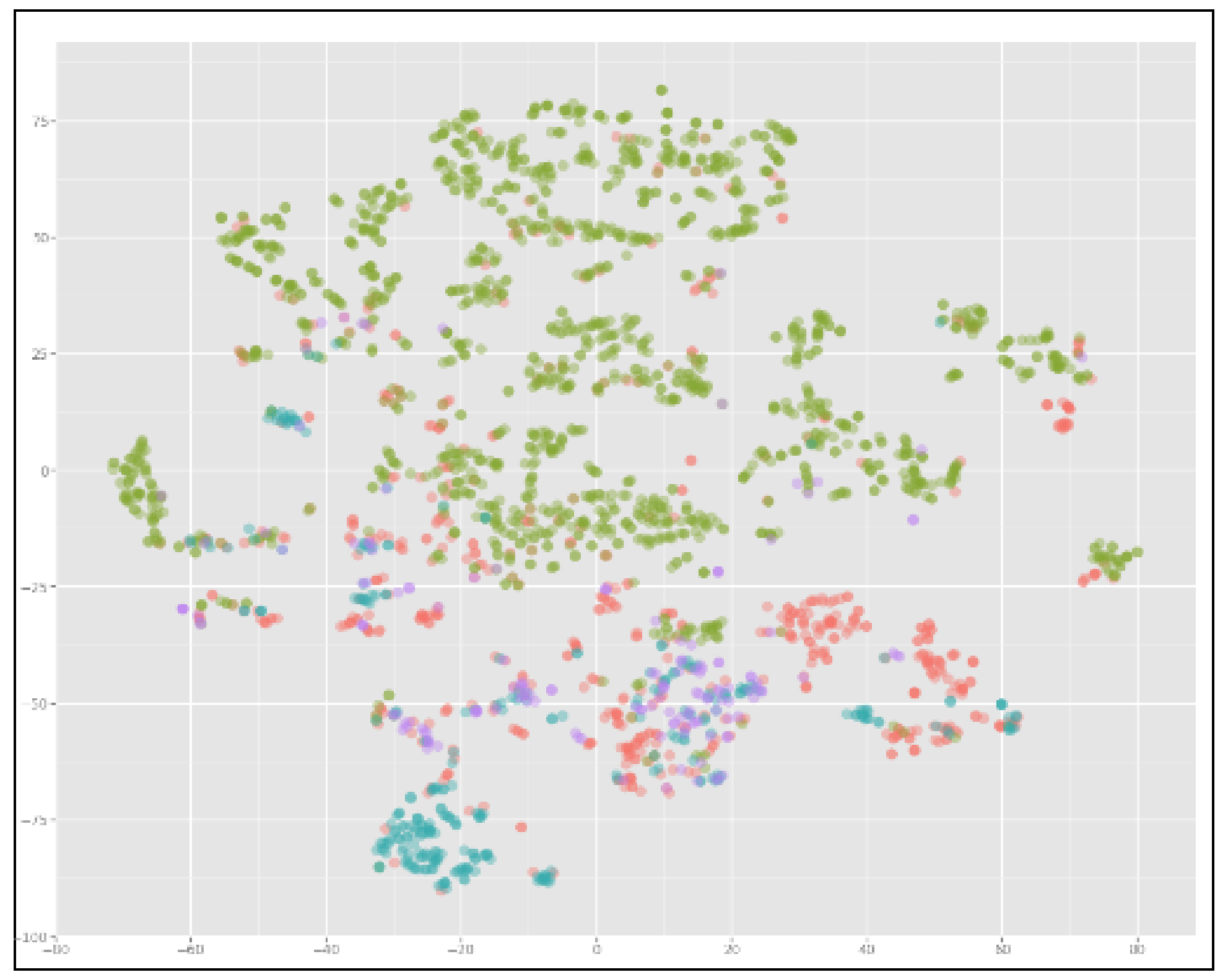
The negative set comprises of non bitter molecules, tasteless molecules, sweet molecules compiled from various sources which include Fenaroli's handbook of flavour molecules(Burdock, 2010) [3] , Rojas at *el*[2].

Type	Source	Number of molecules after database preparation procedures
Positive set	Bitter DB	686
Positive set	Additional bitter : (Rojas et al)(Rojas et al., 2016)	81
Negative Set	Non-bitter: Fenaroli's handbook of flavour molecules(Burdock, 2010)	1753
Negative Set	Sweet : (Rojas et al)	435
Negative Set	Tasteless: (Rojas et al)(Rojas et al., 2016)	133

- The analysis was done in two parts. First, only the physiochemical descriptors were used. Then the physiochemical properties along with the ADMET properties were used. There were in all 3087 molecules, 12 physiochemical properties and 48 ADMET descriptors that were used. Taste information was avalable for all the molecules.

Exploratory Data Analysis

- PCA and TSNE plots were first made for only the 12 physiochemical properties and then, with 12 physiochemical + 47 ADMET descriptors. Below is the Tsne plot for the second case.



Bitter:Red, Non-Bitter:Green, Sweet:Blue, Tasteless: Violet

Classification

Using 12 physiochemical properties, a random forest classifier was implemented. The ratio of bitter and various non bitter molecules including sweet and tasteless was preserved by doing stratified split. A five fold cross validation was performed and the following ressults were obtained.

Set	Positive	Negative	TP	FN	TN	FN	Senstivity	Specificity	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	0.79	
Tasteless				10	26		0.72	0.72	
NonBitter				57	461		0.88	0.11	

Analysis for 12 physiochemical properties.

A similar classifier was built using the 12 physiochemical + 48 ADMET descriptors. The following results were obtained

Set	Pos	Neg	TP	FP	TN	FN	Specificity (TN/TN + FP)	Sensitivity TP/TP + FN	Accuracy TP + TN/Pos + Neg
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		

Analysis for 47 ADMET + 12 physiochemical properties

II. CLUSTERING OF FLAVOUR TERMS

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-occurrence and clustering the molecules based on flavor and structural (2D, 3D) properties. The data was extracted from Flavor DB[5] version 2.0. 4 clustering algorithms and various similarity measures were used for clustering. However, no particular clustering measure could segregate the flavor terms into well formed clusters.

A measure for the assessment of the quality of clusters both quantitative and qualitative, could not be developed.

Also, 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 and them belonging to the same class . The notion of a good cluster could not be developed.

CONCLUSION

The 12 properties weren't sufficient for classification (first model.) However when the 47 chemical descriptors were used in addition to the 12 physiochemical properties, the predictions were better. The overall accuracy was 86%. However, it is able to classify the non bitter molecules better than the bitter ones as indicated by the specificity and sensitivity values. This analysis shows that the second model offers a better classification as compared to the first model.

REFERENCES

Burdock, G. A. (2010). Fenaroli's handbook of flavor ingredients. Taylor & Francis Group.

Dagan-Wiener, A., Nissim, I., Ben Abu, N., Borgonovo, G., Bassoli, A., & Niv, M. Y. (2017). Bitter or not? BitterPredict, a tool for predicting taste from chemical structure. Scientific Reports, 7(1),

Garg, N., Sethupathy, A., Tuwani, R., NK, R., Dokania, S., Iyer, A., … Bagler, G. (2017). FlavorDB: a database of flavor molecules. Nucleic Acids Research, (November), 1–7.

Rojas, C., Ballabio, D., Consonni, V., Tripaldi, P., Mauri, A., & Todeschini, R. (2016). Quantitative structure–activity relationships to predict sweet and non-sweet tastes.

Wiener, A., Shudler, M., Levit, A., & Niv, M. Y. (2012). BitterDB: A database of bitter compounds. Nucleic Acids Research, 40(D1).