Bittersweet

Manveet Singh Harsh Hingorani Karan Yadav

Ganesh Bagler

April 11, 2025

1 Data Compilation and Curation

In this study, we compiled and curated data from previous scientific study, BitterSweet which carefully compiled dataset from variety of resources and excluded molecules with unverified or conflicting taste information.The following resources were used to compile data: (a) Biochemical Targets of Plant Bioactive Compounds by Gideon Polya; (b) BitterDB; (c) Fenaroli's Handbook of Flavor Ingredients (5th Edition); (d) Rodgers et al.; (e) Rojas et al.; (f) SuperSweet; (g) TOXNET; (h) The Good Scents Company Database (www.thegoodscentscompany.com); and (i) BitterPredict.

The data set consists of 918 bitter and 1510 non-bitter molecules and 1247 sweet and 1119 non-sweet molecules. For the creation of machine learning models, the data set was split into training and testing. The data set consists of the name and structure of molecules stored in string format using SMILES (Simplified Molecular Input Line-Entry System) along with bitter/non-bitter and sweet/non-sweet classification. To visulalize the high dimensional feature space training dataset, 2D t-SNE plot was generated using all set of molecular descriptors (Fig 2). The complete schema describing the process of data collection, feature selection, and model building is depicted in Fig 1

2 Materials and Methodology

2.1 Molecular Descriptors

Molecule descriptors define the physical, chemical, and structural properties of a molecule. Online Chemical Modeling Environment (see- https://ochem.eu/login/show.do) was used to calculate 15046 molecular descriptors, namely Dragon 2D/3D (5270), ECFP (1024), JPlogP (115), MOLD2 (777), MORDRED (1613), PaDEL2 (2756), and RDKit (3941).

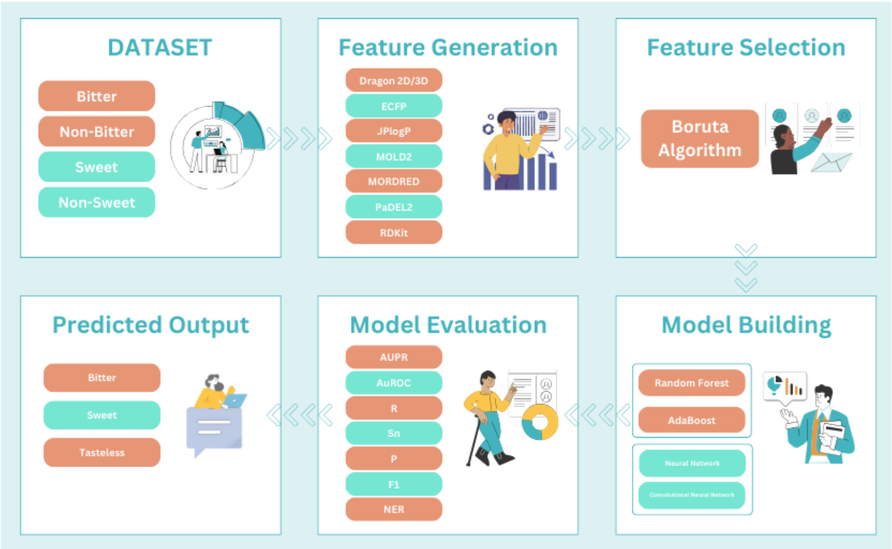

Figure 1: Schema for building bitter/non-bitter and sweet/non-sweet prediction models

2.2 Data Preprocessing and Feature Selection

Since lot of descriptors were common among different classes of molecular descriptors, we removed duplicates and imputed null/missing values with mean of the concerned feature.

Stadard-Scaler was used to normalize each descriptor. The outliers were removed using a Quantile transformer. Features having low variance and highcorrelation were removed using Variance-Threshold method and Correlation-Based method to avoid redundancy and obtain relevant descriptors.

Further Boruta algorithm was applied to remove irrelevant features. A significant proportion of features (65%) in the remaing descriptor were identified to be irrelevant by the Boruta algorithm .

3 Model development & interpretability

In order to classify a molecule into bitter, sweet and tasteless, two different approaches were implemented

Firstly, Random Forest (RF) and Adaboost (AB) machine learning models were implemented to classify a molecule as bitter/non-bitter and sweet/non-sweet, using molecular descriptors chosen from the Boruta algorithm.

Secondly, Neural Network (NN) and Convolutional Neural Network (CNN)

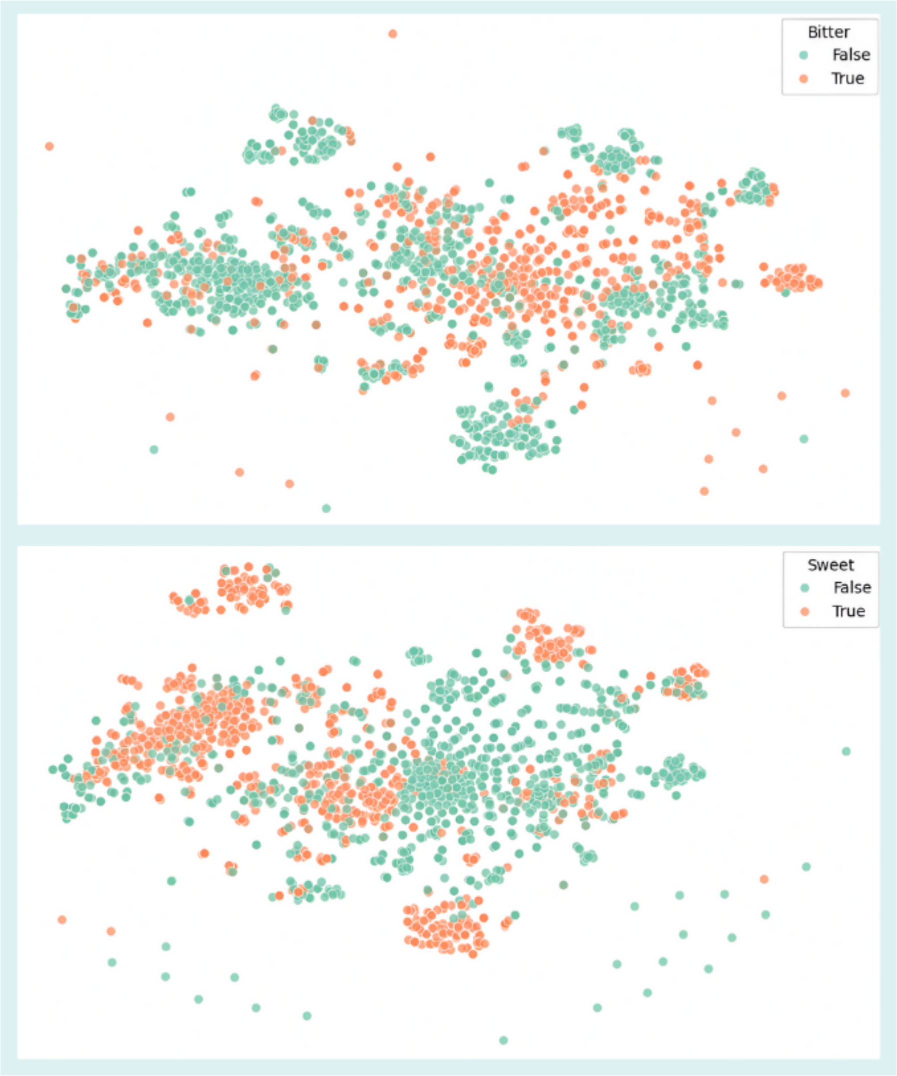

Figure 2: 2D t-SNE scatterplot generated using all set of molecular descriptors for bitter/non-bitter and sweet/non-sweet.

deep learning models were implemented to classify a molecule as bitter, sweet or tasteless. Bitter and sweet data was integrated into single dataset. NN with four hidden layers of 512, 256, 128, and 64 neurons was used. CNN with two Conv1D layers (64 & 128 filters, kernel size 3) was used. For regularization and to avoid overfitting, dropout method and early stopping were implemented, respectively.

For measuring model performance, threshold-independent measures were employed to prevent classification inconsistencies: Area Under the Precision-Recall Curve (AUPR), Area Under the Receiver Operating Characteristic Curve (AUROC) F1-score, Specificity, Sensitivity. Further to find features which contributed most to the prediction of molecule into bitter/non-sweet and sweet/non-sweet SHAP (SHapley Additive exPlanations) Analysis was used.

Random forest (RF). Random Forest algorithm is a type of ensemble learning method that utilizes bagged decision trees. They are quite versatile and can be used for both classification as well as regression. RF works by building a number of decision trees (usually greater than 100) at training time, each utilizing a subset of features and data points. At the time of prediction, the predictions made by its constituent decision trees are aggregated

Adaboost (AB). AdaBoost is an ensemble method that operates by iteratively training weak classifiers in sequence, each of which puts more emphasis on the instances previously misclassified by the earlier ones. It assigns weights to the training examples, raising them for the misclassified samples so that the next classifier puts more emphasis on them. It then pools all weak classifiers by a weighted majority vote (in classification) or weighted sum (in regression).

Neural Network (NN). A Neural Network is a machine learning model inspired by the human brain. It is made up of layers of nodes (neurons) that are interconnected and feed data. NNs learn patterns by adjusting the weights with the assistance of backpropagation. NNs are applied in classification, regression, and pattern recognition tasks.

Convolutional Neural Network (CNN). A Convolutional Neural Network (CNN) is a form of deep learning architectures that are particularly designed to process and understand grid-structured data, e.g., images. CNNs can automatically and flexibly learn spatial hierarchies of features and are able to do so through the utilization of convolutional layers that perform filtering on the input data, thereby extracting meaningful patterns such as edges, textures, or objects.

SHAP (SHapley Additive exPlanations) is one of the most widely used explainable AI (XAI) methods, which provides a contribution value (a SHAP value) to every feature of a machine learning model to interpret the prediction for an instance. The concept relies on cooperative game theory's Shapley values, in which the prediction is considered a ”payout” shared equally among all the features according to their individual contribution.

4 Results

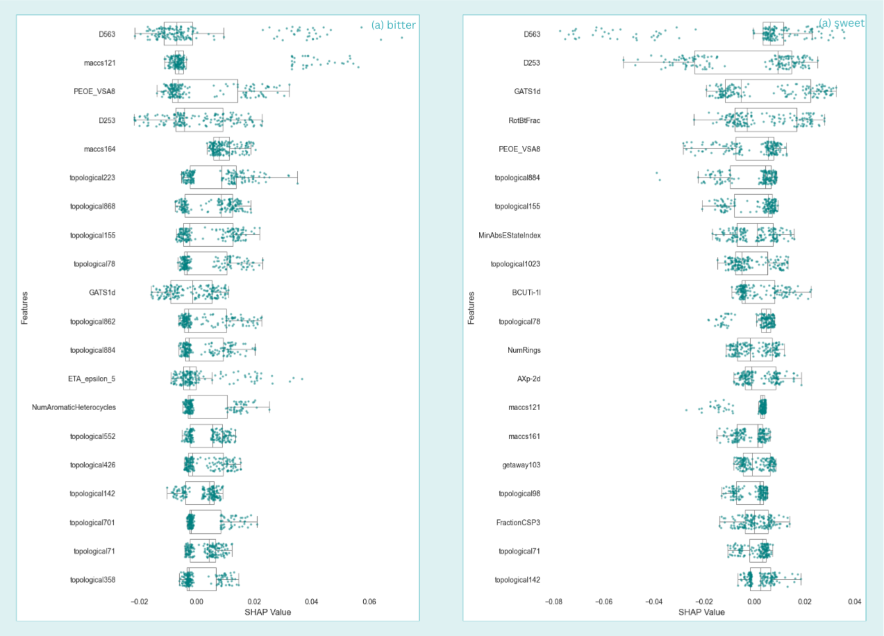

Figure 3: Box plot of importance scores of molecular descriptors for (a) bitter/non-bitter prediction and (b)sweet/non-sweet prediction.

Model	**AUPR**	**AUROC**	**F1**	**NER**	**Sn**	**Sp**
RF Model	**0.9175**	**0.9686**	**0.8631**	**0.9006**	**0.9002**	**0.9394**
CNN Model	0.9162	0.9357	0.8515	0.9048	0.8571	0.8821
NN Model	0.9014	0.9326	0.8616	0.9089	0.8634	0.9034
AB Model	0.5841	0.7630	0.6249	0.7453	0.6883	0.6667

Table 1: Comparison of performance on the test sets of Random Forest (RF) model, AdaBoost (AB) Model, Convolutional Neural Network (CNN) and Neural Network (NN) model