Olfaction properties & molecules-odor analysis

Tanmay Shirish Shukla
001340336

A Project Report submitted to
University at Albany
In partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE
in
Computer Science

Under the supervision of

Prof. Abram Magner

05-15-2020

Authorization for Reproduction of Project

I grant permission for the reproduction of this project in its entirety, without further authorization from me, on the condition that the person or agency requesting reproduction, absorb the cost and provide proper acknowledgment of authorship.

Tanmay Shirish Shukla
University at Albany, SUNY 1400 Washington Ave, NY 12222

Abstract

The world of machine learning and artificial intelligence has been dominating in all the fields. Predicting and building models for automation has eased the way we live now-a-days. Lots of research has been carried out in varied sense of scientific knowledge like defense services, surveillance, entertainment, sports, biology as well as chemistry. Understanding molecular structure and extracting information from those, has been one of the primary goals of chemical scientists. The level of study and work required requires a lot of time along with close monitoring, because slight errors in manual study can change the course of outcomes entirely. With this aspect in mind, machine learners around the globe have started making contributions towards this domain. This project focuses on making useful predictions based on the relationship between a structure of molecule and kind of odor associated with it. Technical wording for this problem is as Quantitative Structure Odor Relationship modelling. The proposed solution to this problem is use of graph convolution network, where structural information of compounds is provided as input and useful vector information is generated. This vector information comprises of atoms and their bonding relationship, which proves to be a useful detail for getting the odor information of any molecule. The vector form is represented as in 2d matrix where each compound is listed along with a specific id. This vector form shall then be inputted to a classifier model, where prediction logic would help to provide outcome. There is a specific list of odors extracted from the dataset. The outcome is obtained in binary format, like a type of odor present in compound shall have a value 1 and 0 otherwise. Thereby, this project aims to predict the odor of any compound information provided as input. It is an ongoing research carried out worldwide and is technically a large project; the prediction results require some extent of time for completion. This project towards the end, has all the information in the form of vectors ready for rf classifier.

Table of Contents

Chapter 1

Introduction

Making predictions from molecules and their structure has been the most sought out research topic in scientific domain. This topic has applications to a varied number of commercial products and markets like the perfume industry as well as for those who have defect in identifying odor. Such molecular properties are more popularly known as Olfaction properties of compounds. Many of these olfaction properties depends on structure-odor relationship. Instead having such importance, the property prediction of molecules has made less progress in forming QSOR (Quantitative Structure-Odor Relationship) models [5]. This has been valid problem for about more than 50 years in the research areas.

The odor receptors in human beings play vital role in perceiving smells, where such receptors are a combination of 300 to 400 multiple types. These are also known as Olfactory Sense Neurons. The way these sensors work is, signals are transmitted to olfactory bulb and later to different brain structural components [6]. Improvements in visual and audio domains has advanced the chances of predicting sensory outcomes. This indirectly helps in making way for using odorant values as well. So, this project first takes a dataset of compound names totaling to about 2400 number of molecules. Along with their name, information about their smell is also provided. The smell can be listed for illustration as citrusy, limy, sweet, and so on. Thereby, we can say the problem to be multi-class problem where one compound can have a multiple number of odors.

Odorant	Odor
trimethylamine	fish
ethanal	pungent, ether
methanethiol	sulfur, gasoline, garlic
propanal	solvent, pungent
pentane	alkane
propanol	alcohol, pungent

Figure 1: Snapshot of compounds along with their odor list

There are various previous assumptions also made to achieve the results. One such assumption involved were where odor of compounds will be similar of those which have similar structures. However, there is one problem that can lead to the above-mentioned assumption, if a simple

change happens to the structure then that can cause the molecule to be odorless. For the dataset used in this project, it is been curated very smartly and carefully, and has been reviewed from trusted database website.

These molecules cannot be easily used for extracting information when entered within any kind of data structure, as there is not much resource available for the same. However, when these molecules are converted into a sort of eligible format, then it becomes easy for getting ahead with the further vector formation stages. Those special format of molecules is known as SMILES string. A SMILES string is nothing, but string of compounds presents in the structure, but having the data of bonds and internal connections as well. SMILES (Simplified Molecular Input Line Entry System) is a simple notation where molecules are easily represented [1]. The advantage of using these types of strings is that it consumes way less storage space, almost about 50 to 70 percent less storage space [1]. Later, getting them, adjacency matrices are generated by importing RdKit libraries. Those matrices are then converted to vectors, which needs the machine to be trained. Trained models can thereby, easily read the matrix data and spit out the required vector data with accurate precision. This method consumes up as the most important part of project. This is because, if vector representation of the molecules is not as required, then the data that shall be fed to the classifier or prediction models would be emit wrong predictions with respect to compound odor. The report talks about different properties of molecules that are being extracted in-details. The properties like fingerprints of the structure, MACCS keys provide cheminformatics, more specifically about their similarities. However, problems with the MACCS key implementing is that they are not fully agreeing within themselves. This project has successfully extracted the properties mentioned above and needs to be carried forward towards inputting into classifiers. As we know that there are various number of odors that are possible for a single compound. For the very reason, the prediction should be listed in a way that it is easy to classify the type of odor emitted by that compound. For this purpose, binary classification for each molecule's odor is done. That means that the odor shall be listed as columns in oppose to the compounds listed as rows. And the matrix format shall have the values 0 or 1, depending on whether that compound has that specific odor or not. If any compound suppose has a smell listed, then the value for that cell shall be 1, and if not then 0. This project is partially successful in ways that, further prediction techniques can be applied. Future scope for this project involves testing with different kinds of prediction models like either rf classifiers or clustering methods as well, depending on which one gives the maximum accuracy in predicting.

Chapter 2

Review of literature

As mentioned above, even though cheminformatics is the most values research program in the combination of chemistry and machine learning, but there isn't much successful work or research ideology present in today's world. Even now, there are several ongoing experiments carried out, with results being updated with each iteration. There were two research papers which proved to be very useful for my project as their work inclined with the kind of work was required for me.

The Machine learning for Scent [2] was the most useful research paper that heled me grasp the importance of the project and its applications. This approach has addressed method to identify quantitative model and structural relationship for reducing the impact caused on ecosystem on use of natural products for generating fragrance. Their contribution has eased the learning of human sensory perception done by the brain. The difference they applied than previous work was that they showed results based on perceptual similarities, instead of taking structural similarities. Each molecule from the dataset was transformed into graph structure. This graph structure had neural network layers for each node, representing compound. Finally, in the graph convolution part, the graph vector operation was performed that helped them to build a network for prediction of molecular properties. This was achieved by specifically embedding the graphs and embedding the space. They had a total of 138 odor descriptors to classify upon and used the above built and trained network. They also had an odorless descriptor included within their list as well for those with no olfaction property of smell in it. They visualized the results in the form of co-occurrence structure between the odor descriptors reflecting whether any descriptor was similar to others or not. The classification analysis of their system performance suggested as a multi-class problem. Their graph neural network for compared against random forest models and k-nearest neighbors based on a RdKit fingerprints. Other factors that they compared their models based on was Mordred features and count-based Morgan fingerprints. Their results were compared on mean AUROC. The AUROC feature and GNN embeddings used by their approach has helped their research to be compared and analyzed on examining global structure of learned space of odor. They did a comparison of RF-cFP and graph neural networks divided into description of odor. However, their work just showed learned knowledge of embeddings by extracting useful information from molecules. This could be applied in the field of odorants and perfume industries [2].

Other research work that I reviewed was the Compound-protein interaction prediction done with end-to-end learning of neural networks for graphs and sequences [3]. Their work was highly motivated towards making predictions for drug discovery and analysis. Their study showed the importance and impact of Compound-protein interaction in the field of discovering drugs. They performed end-to-end learning of discrete symbolic data with the help of deep learning

methodology. This proved to be an excellent performance measure on different levels of problem statements. For their problem, the data was inputted as a form of symbolic data, that means compounds were shown as graphs and vertices were represented as atoms. The edges were shown as the chemical bonds within the compound, proteins were the sequences with characters as amino acids. They integrated the representations and developed a whole together new form of compound-protein interaction prediction method after successfully combining the graph neural network for compounds and convolution neural network for proteins [3]. They used three kinds of datasets and proposed an end-to-end method for achieving a competitive result as compared to previous methods. Their approach demonstrated a data-driven ideology for proteins and compounds. It is a common problem that deep learning models are usually harder for analyzing as per their black box property; however, their research overcame this as they made use of a neural attention mechanism. This mechanism proved to be useful for considering only that sequence which would be vital for a drug compound while making predictions. The above-mentioned mechanism also provided an effective way of visualizing any model, no matter that real values were into representations. The part which proved useful for my research was their way of finding relationship between compound and protein. They sorted out two kinds of vectors from SMILES string of molecules, one was the compound vector, and other was the protein vector. Both these vectors were later concatenated and fed to a prediction model for making interaction predictions. Their research provided baseline for deciding the importance of graph neural network in terms of molecules and their vector format. They made use of two types of neural networks: convolution networks and graph convolutions. The earlier one was used primarily just for protein amino sequences provided as input [3]. Their performance evaluation measure was in-depth with analysis using AUROC factor, effects of hyperparameters done on compound-protein interaction [3].

The on-going research work done by Google on “Learning to Smell: Using Deep learning to predict Olfactory Properties of Molecules” is aimed at building a AI machine that can directly provide a predicted value for odor descriptors [4]. Their work, when complete, shall be able to achieve the result without following the traditional rules, like need for some format of vector. They, thereby, claim to achieve a highly improved odor prediction performance when compared to already placed traditional methods [4].

There were some other methods also done which involved demonstrating the results in just binary format like sweet or not, bitter or not, and so on. The bitter-sweet forest [7] was another such binary classifier for predicting bitterness and sweetness property of compounds. This was the first KNIME workflow mechanism that provided a base for predicting bitter and sweet taste of any chemical molecule making use of their fingerprints and random forest classifiers. This model was supposed to have receive an accuracy of 95% [7], however due to its limitation to only two kinds of taste or odor descriptors, simply put a big constraint on its usage.

Chapter 3

Proposed work and Tool Implementation

Based on all the above-mentioned reviewed literature, I proposed a novel approach of using Graph Neural Networks for generating vectors, which further can be applied to any rf classifier models or other prediction strategies.

3.1 Reading the dataset

Collecting a dataset proves to the foundation of a machine learning project. If the data is appropriate and well apt with the project requirements, then heading to further processing steps becomes smooth. With this point of view, I extracted data from FlavorDB online repository, curated by IIT-Delhi [8]. They have developed our Flavor Network which visualizes the range of Flavors in which compounds might fall. Network uses its own repository of compounds connected to different flavors. For my work, I have used 2,372 compounds in all from FlavorDB; and read all the compounds using Pandas. The list of compounds and their odor was stored in csv (Comma delimited) format. This was done because it becomes easy to load csv files using pandas package.

The dataset was read and printed as follows:

```
>>> df = pd.DataFrame(data = dataset)
>>> print df
                                  compounds                                  flavor
0                          1-Aminopropan-2-Ol                                   fishy
1                        2-Deoxyhexopyranose                                   sweet
2                 3-Methyl-2-Oxovaleric Acid                                     NaN
3                3-Methyl-2-Oxobutanoic Acid                                  fruity
4                        2-Ketoglutaric Acid                                odorless
5                         2-Oxobutanoic Acid  brown, caramel, lactonic, sweet, creamy
6                 4-Methyl-2-Oxovaleric Acid                                  fruity
7                  3,4-Dihydroxybenzoic Acid              phenolic, balsamic, mild
8                                   AC1L18F4                             sweet-like
9                     3-Phenylpropanoic Acid  balsamic, rose, fatty, sweet, musk, cinnamon
10                       4-Aminobutyric Acid                          savory, meaty
11                   4-Hydroxybenzyl Alcohol  coconut, bitter, sweet, fruity, almond
12                    4-Hydroxybenzaldehyde     woody, sweet, balsam, nutty, almond
13                    4-Hydroxybenzoic Acid                        nutty, phenolic
14                               Acetic Acid          sour, pungent, sharp, vinegar
```

Figure 2: FlavorDB containing molecule name and their odor

It is evident that the above format of input won't work with graph neural networks as input. This is because, getting data about the molecule just from its name is really tricky and difficult to code with. However, if a certain format of string related to those molecules is used, then it would prove

useful. For the very same reason, I first extracted just the molecule names, and then converted into SMILES format of strings.

SMILES string: Simplified Molecular Input Line Entry System is a notation in chemical terms, that is used for representing any chemical structure in an eligible way that can be understood by machines. There are some rules associated with them [9]. All the elements that are present in periodic element are supported by SMILES. Atomic symbol is used for determining that particular atoms. If there is a case where any atomic symbol shall have multiple letters then the second letter of that symbol shall be a lowercase letter. Different types of bonds supported by SMILES string are single bond, double bond, triple bond, aromatic and disconnected structures. A simple illustration of chemical representation is as follows: a compound ethane shall have the form "CC"; where it denotes that one carbon is connected to other non-aromatic carbon compound with a single bond. In a similar way, there are other rules for representing all different types of structures and their connections with the help of bonds.

3.2 Converting Molecular names to SMILES string

Given the advantages of SMILES string, I converted all the molecular names within my dataset into that format. For this purpose, I used **"cirpy package"** which is one of a python interfaces for showing Chemical Identifier Resolver determined by CADD group.

```
>>> import cirpy
>>>
>>> for i in range(738):
...     cirpy.resolve(df['odorant'][i], 'smiles')
...
'CN(C)C'
'CC=O'
'CS'
'CCC=O'
'CCCCC'
'CCCO'
'CSC'
'CCOC=O'
```

Figure 3: Conversion of molecular name to SMILES string

The dataset now had a list of compounds, its odor and respective smiles strings as well. However, the data still needed to be checked for completeness.

3.3 Handling the Blank values

There were some of the compounds which had no flavors defined for them; instead those values were just left blank. Such compounds were straightaway removed from the list as handling that kind of data would have affected the accuracy of model. Filling up the blank values with Nan or any such substitutes can lead to adding complications in effectiveness.

The preprocessing step now has only one part left in it. The data is being handled with discrepancies; now further, the SMILES string needs to be converted to a graph input.

3.4 Converting SMILES to Graph and Adjacency matrix:

The models working for machine learning purposes tend to require a fixed-shape or regular-shaped data to be inputted, such as vectors represented as numbers and so on. However, this is where the graph neural networks prove to be advantageous, as they allow us for usage of irregular shaped data input. Thereby, for my project, I can use graphs as an input for application. The way I started to visualize graphs is converting into adjacency matrix, these adjacency matrices has chemical atoms as nodes and bonds were showed in the form of numbers [2].

The "pysmiles" package was used for creation of adjacency matrix graph by giving input as SMILES string for representing as node graph.

```
from pysmiles import read_smiles
import networkx as nx
import pandas as pd
smiles = pd.read_csv("FlavorDB_manual_none.csv")
print smiles["smiles"][0]
df = pd.DataFrame(data = smiles)
print df
for i in range(2371):
    mol = read_smiles(df['smiles'][i])
    print (mol.nodes(data='element'))
    print (nx.to_numpy_matrix(mol))
```

Figure 4: Working of conversion of strings to adjacency matrix graph

Adjacency matrix is a representation of number of vertices within a graph where it is shown as a 2-dimensional array. It becomes easy to represent the connection between nodes and their connected edges. Another benefit of representing in this form is that implementation takes very less time.

```
[(0, 'S'), (1, 'C'), (2, 'C'), (3, 'C'), (4, 'C'), (5, 'C'), (6, 'C'), (7, 'S')]
[[0. 1. 0. 0. 0. 0. 0. 0.]
 [1. 0. 1. 0. 0. 0. 0. 0.]
 [0. 1. 0. 1. 0. 0. 0. 0.]
 [0. 0. 1. 0. 1. 0. 0. 0.]
 [0. 0. 0. 1. 0. 1. 0. 0.]
 [0. 0. 0. 0. 1. 0. 1. 0.]
 [0. 0. 0. 0. 0. 1. 0. 1.]
 [0. 0. 0. 0. 0. 0. 1. 0.]]
```

Figure 5: Adjacency matrix of each corresponding compound

By representing atoms in the form of nodes, bonds in the form of edges; implementation of molecules is then done by adjacency graph. The graph neural networks are known for their learning skills of features [2]. This project explored the feature extraction like MACCS key, molecular fingerprints with the help of adjacency matrix generated. The working model is completed until this part of feature extraction. More amount of time frame is required for building a complete prediction model for accomplishing results.

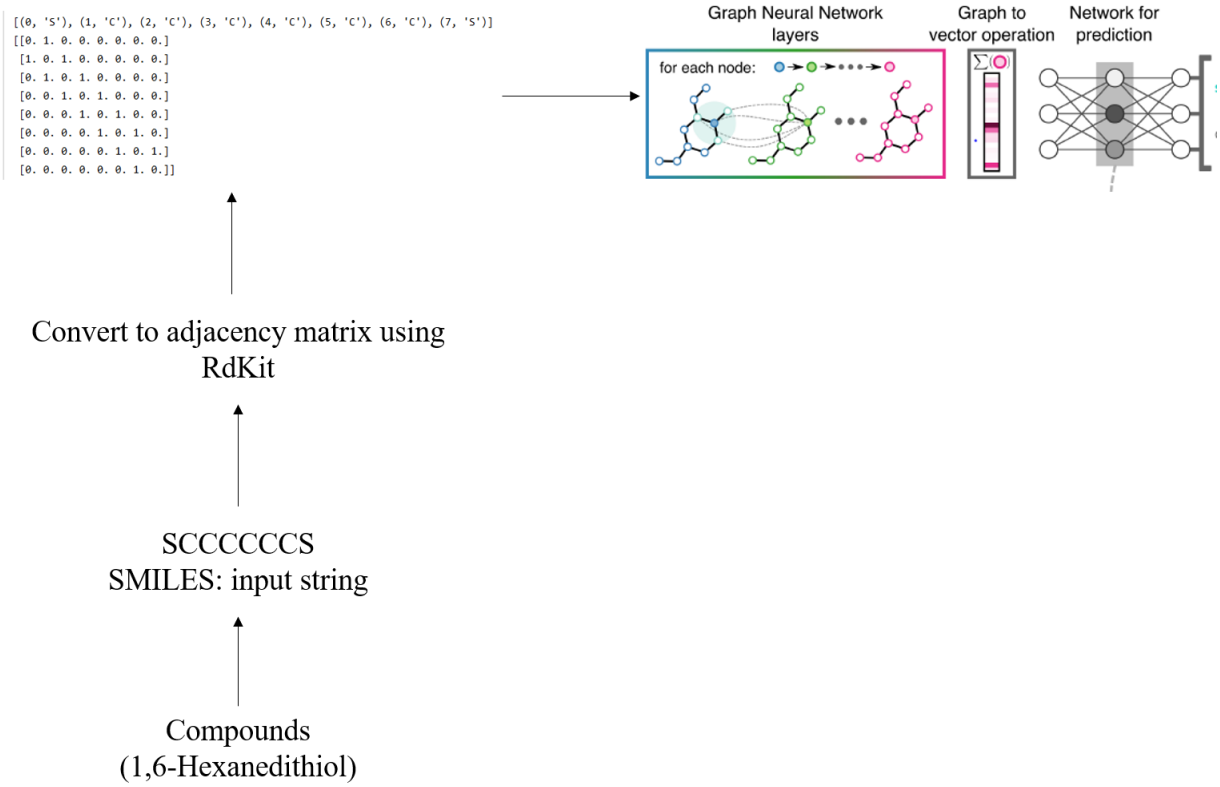

Figure 6: Working model of Olfaction Prediction

The above figure represents the current working model of this project. The adjacency graph generated is used for extracting the molecular properties that are fed to Graph Neural Network layers, where vectors are formed. These vectors can further be utilized for prediction models.

Chapter 5

Future Work

Given the complexities and data availability of the project domain, the time frame for completion becomes larger. Currently, companies like Google and Kaggle have their researchers working hard around similar project ideas. Most challenging factor towards accomplishment of work is the availability of dataset. Choosing the right format of data, and getting reliable dataset is very important. Fortunately, online repositories have made some amount of data available publicly, thereby I was able to complete about 70% of work. This project when divided into four phases; I completed with three of them, with the last phase of selecting an appropriate classifier model left. The right choice of classifier model would further accelerate prediction methodologies, giving us a binarily classified result of which odors are present in the compounds present in that dataset. Binary classification shall easily help classify multiple odors when present in a single molecule. For example, a compound having odor citrusy, fruity and sweet shall have values as 1 for those cells under them, and the rest shall have value 0. Now, what kind of prediction model shall prove to be most efficient one for the purpose. The benchmark performance of classification problems for the number of odor descriptors can be easily compared among k-nearest neighbors as well as random forest classifying trees based on fingerprints extracted using RdKit [2]. AUROC performance results also can also prove helpful for data visualization.

Chapter 6

Conclusion

I successfully assembled a large set of data labelled with compound names, and their respective odors. Further, built a working model that can convert those molecules into a graph acceptable SMILES string. Properties of compounds were extracted accordingly, with respect to the adjacency matrix generated by using python and RdKit. Packages embedded in python programming language including Pytorch, cirpy and scikit-learning turned handy for the purpose of graph matrix generation and gathering fingerprints information of those molecules. The current completed work of the project took about three months of working. However, the complete project requires a little more timeframe than passed; thereby building and training a graph neural network for further prediction models need to be worked on. Progress done until now, clearly provides an insight into overall project description, so that further process could be easily carry forwarded. For predicting odors accurately, a random forest classifier model is to be built. This shall lead towards correct binary classification of each compound odor. The models previously implemented have been kind of an analysis reports, suggesting the types of models and the number of accuracies they have potential to achieve with. So, on successful completion, this project shall prove to be a benchmark on this domain for its performance and results.

References

[1] *Daylight Theory: SMILES*. [Online]. Available: https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html.

[2] Benjamin, Wei, L. Jennifer N., B. K., Gerkin, A.-G. Richard C., Wiltschko, and A. B., "Machine Learning for Scent: Learning Generalizable Perceptual Representations of Small Molecules," *arXiv.org*, 25-Oct-2019. [Online]. Available: https://arxiv.org/abs/1910.10685.

[3] M. Tsubaki, K. Tomii, and J. Sese, "Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences," *Bioinformatics (Oxford, England)*, 15-Jan-2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29982330.

[4] "Learning to Smell: Using Deep Learning to Predict the Olfactory Properties of Molecules," *Google AI Blog*, 24-Oct-2019. [Online]. Available: https://ai.googleblog.com/2019/10/learning-to-smell-using-deep-learning.html.

[5] "Structure−Odor Relationships," *Chemical Reviews*. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/cr950068a?src=recsys.

[6] C.-Y. Su, K. Menuz, and J. R. Carlson, "Olfactory perception: receptors, cells, and circuits," *Cell*, 02-Oct-2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19804753.

[7] P. Banerjee and R. Preissner, "BitterSweetForest: A Random Forest Based Binary Classifier to Predict Bitterness and Sweetness of Chemical Compounds," *Frontiers in chemistry*, 11-Apr-2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5905275/.

[8] *FlavorDB*. [Online]. Available: https://cosylab.iiitd.edu.in/flavordb/molecules?common_name=&functional_group=&flavor_profile=&fema_flavor=&molecular_weight_from=&h_bond_donors=&h_bond_acceptors=&type=&smile=&page=50.

[9] Mid-Continent Ecology DIvision, "SMILES tutorial", [Online]. Available: https://archive.epa.gov/med/med_archive_03/web/html/smiles.html