Testing Graph-Based Decision Tree Models for Phenolic Compound Adhesion Analysis

Racheal Fisher

Department of Chemistry, Purdue University, West Lafayette, IN 47906, United States

Email: fishe290@purdue.edu

ABSTRACT: Ten phenolic compounds containing catechol with known adhesions were trained and tested on various decision tree models to determine how accurately a decision tree could predict the adhesion of the compounds when combined with zein, a non-toxic corn protein, and whether the results could be used to determine what functional group of a phenolic compound would increase adhesivity. Previously determined adhesion values were used. The decision tree was found to make decisions primarily based on resonance and oxygen groups. It was more accurate with larger folds, though its success still varied by compound.

1. Introduction

Many factors are involved in the adhesion levels of a given compound. Surface chemistry, temperature, and intermolecular forces are all important, as well as the method of application. In this study, we investigated the importance of chemical structure on the adhesion of similar catechol-containing phenolic compounds.

To this end, other variables, such as temperature, percent mixture, and surface, were kept as constant as possible.

Zein/phenolic compound mixtures have been found to have comparable adhesion to that of super glue, while also being biodegradable, plant-based, and nontoxic[3].

The investigation into predicting the adhesive properties of the phenolic compounds was performed in Python programming language. Python was chosen due to its large set of open-source libraries and cross platform capabilities. Specifically, sklearn, matplotlib, and rdkit were integral to creating and understanding the decision tree.

The graph-based model used was based off of the graph-based decision tree used to model MOPs[1]. The idea was used to create a decision tree run completely in Python that gives visual result outputs.

The dataset of phenolic compounds provided a large challenge due to its small size. To minimize the effects of this, the decision tree was trained on and tested with different subsets of the dataset to be compared. This way, the performance of the tree could be maximized and compared with different testing/training splits. Figure 1 displays the sample set used.

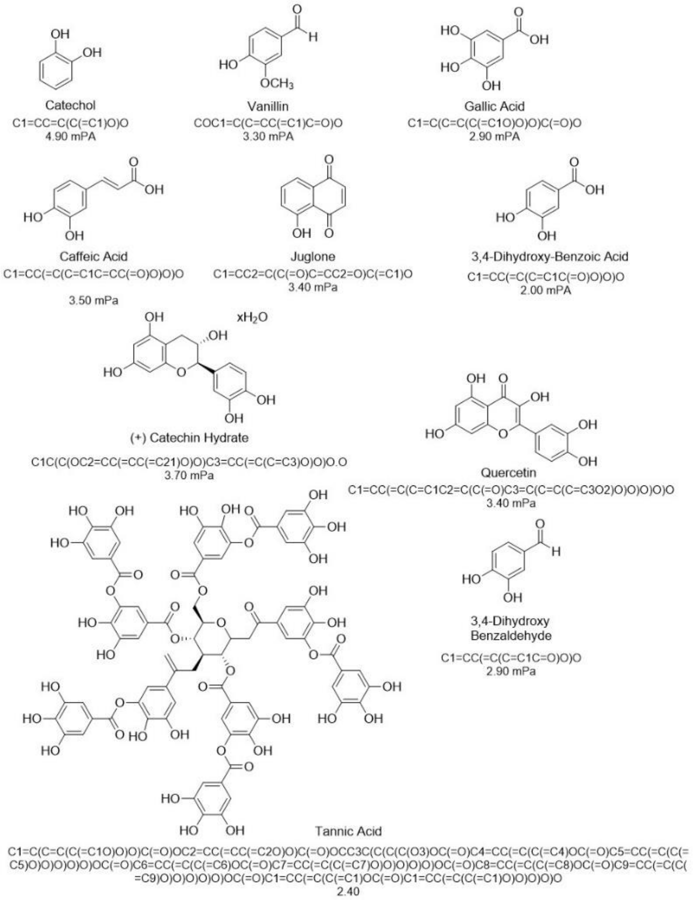
Catechol
C1=CC=C(C(=C1)O)O
4.90 mPA
Vanillin
COC1=C(C=CC(=C1)C=O)O
3.30 mPA
Gallic Acid
C1=C(C=C(C(=C1O)O)O)C(=O)O
2.90 mPA
Caffeic Acid
C1=CC(=C(C=C1C=CC(=O)O)O)O
3.50 mPa
Juglone
C1=CC2=C(C(=O)C=CC2=O)C(=C1)O
3.40 mPa
3,4-Dihydroxy-Benzoic Acid
C1=CC(=C(C=C1C(=O)O)O)O
2.00 mPA
xH2O
(+) Catechin Hydrate
C1C(C(OC2=CC(=CC(=C21)O)O)C3=CC(=C(C=C3)O)O)O.O
3.70 mPa
Quercetin
C1=CC(=C(C=C1C2=C(C(=O)C3=C(C=C(C=C3O2)O)O)O)O)O
3.40 mPa
3,4-Dihydroxy
Benzaldehyde
C1=CC(=C(C=C1C=O)O)O
2.90 mPa
Tannic Acid
C1=C(C=C(C(=C1O)O)O)C(=O)OC2=CC(=CC(=C2O)O)C(=O)OCC3C(C(C(C(O3)OC(=O)C4=CC(=C(C(=C4)OC(=O)C5=CC(=C(C(=C5)O)O)O)O)O)OC(=O)C6=CC(=C(C(=C6)OC(=O)C7=CC(=C(C(=C7)O)O)O)O)O)OC(=O)C8=CC(=C(C(=C8)OC(=O)C9=CC(=C(C(=C9)O)O)O)O)O)OC(=O)C1=CC(=C(C(=C1)OC(=O)C1=CC(=C(C(=C1)O)O)O)O)O
2.40

Figure 1: phenolic compound dataset, labeled with their SMILES strings and adhesion values when combined in a ratio of 94 wt% zein and 6% compound.

2. Results and Discussion

The results discussed pertain both to the performance of the model on the adhesivity of the phenolic compounds, and the information gained from the results of the model.

2.1 Decision Tree Training

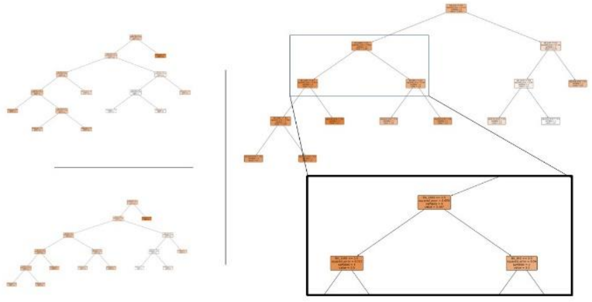

Figure 2: Decision tree output examples for 10 splits

The decision tree was trained on one part of the sample set (the training set), and tested on another (the testing set). The set was rotated so that each compound was tested, and results were displayed for each run. For example, for one run quercetin was tested on the model created by the other 9, while on the next run, catechol was tested on the model created by the other 9 (including quercetin).

Figure 2 displays examples of representations of the trained decision trees produced by sklearn's plot_tree() function. The leaves on the split are labeled with bit numbers derived from the Morgan fingerprints of all molecules in the training set. The representation shown above is how the model related the bits from the training set to adhesion values. As such, there is not always a direct 1:1 translation of each bit to each specific molecule's functional groups. While a majority of the bits are linked to the same functional groups for different smiles strings, occasionally they will have different functional group representations. Because of this, the bits are more an indication of the "importance" of a functional group in adhesion, rather than a direct decision-making factor. In figure 3, the bits used in decision tree model 1 (which was tested on quercetin and trained on all other compounds), were compared with each SMILES string to attempt to find their corresponding functional groups.

The primary structures shown in the tree are either -O or -OH groups, or different types of bonds. Resonance bonds are commonly featured. None of these features are unexpected, as they make up most if not all the features of phenolic compounds. They are also likely to be responsible for adhesion due to their polar properties. It is interesting that ketones, aldehydes, and carboxylic acids are not featured on the tree, when they are on the compounds. They appear to be the only major feature not included. This could be due to several reasons. A potential reason is that they are unimportant in adhesion. This seems unlikely, as carbonyl-containing groups are polar and could lead to dipole-dipole interactions influencing adhesion. They would also lead to hydrogen bonding and increased stability. Another reason could be due to structure redundancy. The model identified oxygen and double bonds as important features, and it is plausible that it did not distinguish between those features.

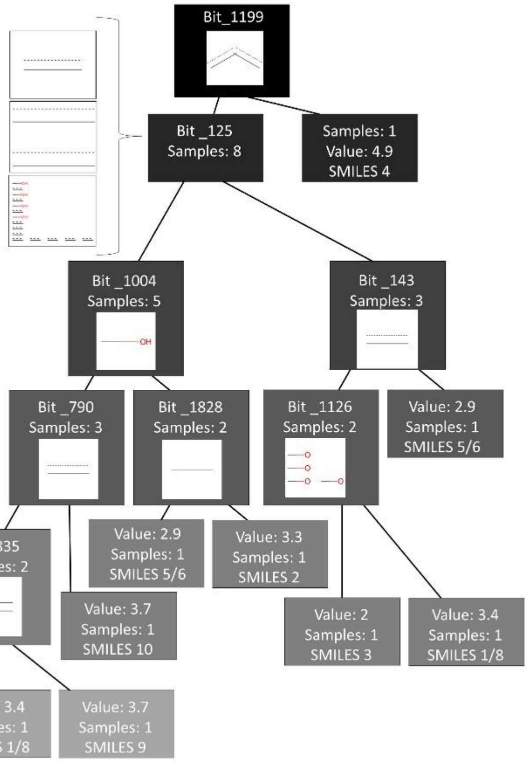

Figure 3: Decision tree trained to be tested on gallic acid. The bits are shown as their corresponding functional groups.

and carbonyls. Other reasons for the lack of carbonyl compounds would be due to the limitations of the model itself. The small sample size or limitations of decision trees could be responsible for the oversight.

2.2 Decision Tree Results

With each run, the decision tree assigned an adhesion value to each compound based on the tree it created training off the other compounds. To analyze how the tree was performing, three different methods were used. The tree was analyzed when using both a 10-fold model (train on 9 compounds, test on 1) and a 5-fold model (train on 8 compounds, test on 2). Later on, a 2-fold model (train on 5 compounds, test on 5 compounds) is used.

First, as seen in Figure 4, each model was run several times. The average and standard deviation of the adhesion value of each compound was then compared to the known values. For some compounds, the model predicted adhesion values very similar to the known values each time. Vanillin and Quercetin were good examples of this. For several of the compounds, the output adhesion values of the 5-fold and 10-fold trials were much closer to one another than they were to the given value. Visually, there is not a strong trend for one model being more accurate than another.

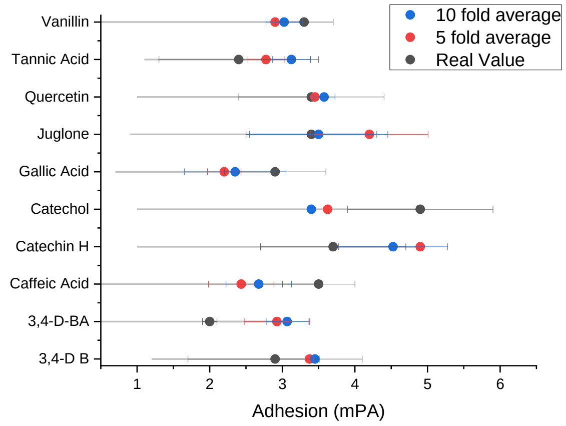

Figure 4: Comparison of average adhesion values

A possible reason for the accurate predictions of vanillin is that three of the other compounds appear very similar to vanillin. Gallic acid, 3,4-D-BA, and 3,4-D B all have structures that share most of the same shape and functional groups as vanillin. This could have given the decision tree a better understanding of the adhesive qualities of vanillin. However, this would make it odd that gallic acid, 3,4-D-Ba, and 3,4 D B were not predicted with similar accuracy. If the same logic is applied to Quercetin, we could say that it shares structural similarities to catechin H and tannic acid. However, it should be stated that these disparities in accuracy could also be related to small sample size.

To gain a better comparison of the two-fold parameters, we compared the average relative error of each of the adhesion outputs using Eqn. 1. The results are shown in figure 5.

$$Eqn.\ 1: Relative\ Error: \frac{Predicted\ Value - Training\ Value}{Training\ Value}$$

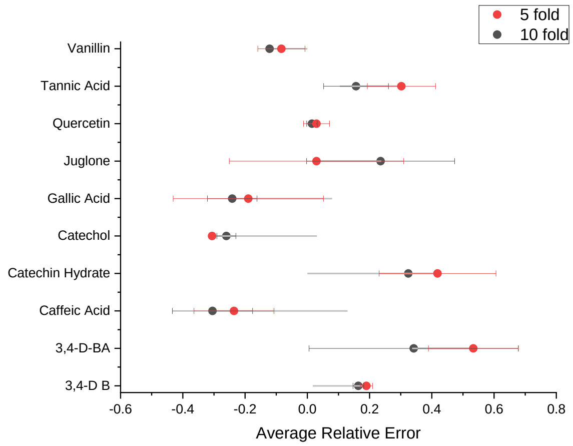

Figure 5: Comparison of relative error by fold number

As expected, vanillin and quercetin have very similar relative error differences between the 5-fold and 10-fold models. Gallic acid, catechol, 3,4 D B, and caffeic acid are also very close. Within the other compounds, the 10-fold model's relative error was lower for tannic acid, caffeic acid, and 3,4 D-BA, while the 5-fold model's relative error was lower for juglone.

The 10-fold model performed slightly better; however, the relative error was much closer than expected. To see if the trend continued, a 2-fold model was tested. The results are below in figure 6.

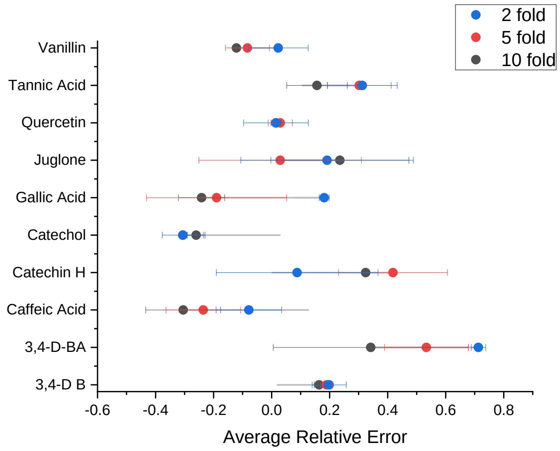

Figure 2: Comparison of relative errors, with a 2-fold model added.

For quercetin and 3, 4-D B, the average relative error for 2-fold matches very closely with the other models. For juglone, the relative error is in between that of the other two models. For catechol and catechin H, the error for 2-fold was lower than the other models. For 3,4-D-BA, gallic acid, tannic acid, caffeic acid, and vanillin, the error for 2-fold was higher. This aligns with the previous trend, but it does not show a strong difference.

Relative error was the best method of analysis for the accuracy of the decision tree. Cohen's kappa was considered but was not used as it is superior for categorical data.

As the final method of analysis, table 1 below shows the MSE values for each fold of the methods along with the tested compounds. The average MSE for the 10-fold model was .569, the average MSE for the 5-fold model was .645, and the average MSE for the 2-fold model was .92. As the number of folds decreases, the MSE value increases, meaning that the model becomes less reliable.

Table 1: Results

Fold	Test Compound	MSE
1	Quercetin	0.01
2	Vanillin	0.16
3	3,4-Dihydroxy-Benzoic Acid	0.81
4	Catechol	1.44
5	3,4-Dihydroxy Benzaldehyde	0.16
6	Gallic Acid	0.81
7	Tannic Acid	0.25
8	Juglone	0.25
9	Caffeic Acid	0.36
10	Catechin Hydrate	1.44
5-Fold Model		
1	Gallic Acid	.585
	3,4-Dihydroxy Benzaldehyde	
2	Catechol	1.53
	3,4-Dihydroxy-Benzoic Acid	
3	Juglone	.725
	Catechin Hydrate	
4	Caffeic Acid	.18
	Quercetin	
5	Vanillin	.948
	Tannic Acid	
2-Fold Model		
1	Catechol	0.948
	Gallic Acid	
	3,4-Dihydroxy-Benzoic Acid	
	3,4-Dihydroxy Benzaldehyde	
	Quercetin	
	Vanillin	0.892
	Caffeic Acid	
	Juglone	
	Catechin Hydrate	
	Tannic Acid	
	Vanillin	

This shows that a larger training size will lead to better predictions.

2.3 Overall Performance

The model ran better with a larger fold size, and better with specific compounds. Quercetin had a .01 MSE in the 10-fold model and very accurate prediction values. A unique feature in quercetin is that there is both a carbonyl and an ether. The decision tree in figure 3 is what quercetin was tested on. In that context, quercetin was SMILES 1, and SMILES 8 was juglone. This means that quercetin was assigned to the 3.4 value dependent on the 4 -O groups, which is something that juglone does not have. Juglone was so accurately classified from its -O groups and resonance bonds.

Tannic acid was expected to be predicted very poorly due to its large size when compared to the other compounds. However, it had a standard relative error. This is likely due to how it shares functional groups with the other compounds regardless of size.

Overall, the decision tree was marginally successful in predicting the adhesion of phenolic compounds when combined with zein. We are interested in seeing how it would change when a larger sample size was available for training and testing. While it had some accurate predictions, it also had some inaccurate ones. Its predicted adhesion trends follow those of the sample, but it is not yet reliable for predicting unknowns.

REFERENCES

[1] Fine, J. A., Liu, J. K.-Y., Beck, A., Alzarieni, K., Ma, X., Boulos, V., Kenttämaa, H., & Chopra, G. (2019). Graph based machine learning interprets diagnostic isomer-selective ion-molecule reactions in tandem mass spectrometry. *Chemical Science*. https://doi.org/10.26434/chemrxiv.11466183

[2] North, M. A., Del Grosso, C. A., & Wilker, J. J. (2017). High strength underwater bonding with polymer mimics of mussel adhesive proteins. *ACS Applied Materials & Interfaces*, *9*(8), 7866–7872. https://doi.org/10.1021/acsami.7b00270

[3] Schmidt, G., Hamaker, B. R., & Wilker, J. J. (2018). High strength adhesives from catechol cross-linking of Zein Protein and plant phenolics. *Advanced Sustainable Systems*, *2*(3), 1700159. https://doi.org/10.1002/adsu.201700159

[4] Schmidt, G., Smith, K. H., Miles, L. J., Gettelfinger, C. K., Hawthorne, J. A., Fruzyna, E. C., & Wilker, J. J. (2022). Tunable tannic acid–zein adhesives for bonding different substrates. *Advanced Sustainable Systems*, *6*(6), 2100392. https://doi.org/10.1002/adsu.202100392

[5] Schmidt, G., Woods, J. T., Fung, L. X. B., Gilpin, C. J., Hamaker, B. R., & Wilker, J. J. (2019). Strong adhesives from corn protein and tannic acid. *Advanced Sustainable Systems*, *3*(12), 1900077. https://doi.org/10.1002/adsu.201900077