ML-based analysis of Chemical Compounds and their Spectrograms

Anshika Jhamb[1], Niranjan Rajesh[1], Riyosha Sharma[1], Sristi Bafna[1], Debayan Gupta[1], and Subhajyoti Chaudhuri[2]

[1]**Ashoka University**
[2]**Northwestern University**

ABSTRACT

In this project, we aim to employ machine learning techniques to infer chemical properties of compounds based on Infrared Spectroscopy (IR) and UV-vis Spectrophotometry (UV-vis). Inferring features like the existence of functional groups in a compound, recognition of all functional groups in a compound and identification of the compound based on corresponding Spectrograms are the main concerns of the project.

Keywords: Machine Learning, Spectrograms, Chemical Functional Groups

INTRODUCTION

The identification of functional groups in chemistry is crucial, as these atomic combinations within molecules define their unique characteristics, affecting attributes like boiling point and solubility. In addition to playing a vital role in organic synthesis planning, these groups are pivotal in structure elucidation.

Ultraviolet-Visible (UV-Vis) spectrophotometry is a valuable tool for understanding compound properties. By measuring the absorption of UV and visible light, it helps identify specific functional groups. The interpretation of UV-Vis spectra is vital for comprehending electronic transitions in compounds, with applications ranging from quantifying substance concentrations to studying chemical reaction kinetics. Effective spectral analysis methods are essential for precise insights into electronic properties and functional groups in chemical compounds using UV-Vis spectroscopy.

A SMILES (Simplified Molecular Input Line Entry System) string is a concise, interpretable representation of a chemical compound's structure. SMILES strings can encode chemical information for ML models to predict functional groups and aromaticity from spectroscopy data, bridging the gap between chemical expertise and data-driven analysis in infrared and UV-Visible spectroscopy. While highly versatile, SMILES strings have a fixed-length format, which can be inadequate for compounds of varying complexity. Additionally, they may not directly represent 3D structural details, making them less suitable for applications requiring spatial information. Parsing complex or ambiguous compounds may be challenging, and handling tautomeric forms can be cumbersome. Some of these challenges can be overcome by working with molecular fingerprints. Their variable-length format accommodates compounds of diverse complexities and makes them well-suited for handling structurally diverse molecules. They capture structural features more comprehensively, including 3D information, and can represent compounds with multiple tautomeric forms accurately.

This study aims to investigate novel approaches for identifying chemical compound properties based on UV–Visible spectrophotometry (UV-Vis). The objectives include detecting the presence of functional groups within a compound, recognizing all functional groups, as well as identifying the compound itself, in terms of its SMILES representation or molecular fingerprint. Additionally, we explore the prospect of reconstructing spectra from these attributes. In essence, our project leverages machine learning to analyze the link between chemical compounds and their respective spectra.

DATASET

We scraped respective data of around 1500 compounds from the NIST Webbook's UVVis Spectra collection. This data included the molecule name, formula, and corresponding SMILES string representation as well as UV-Vis spectral data (wavelengths and logarithmic intensities) of the compounds. The input for all the CNN models were normalised logarithmic intensities for compounds represented as vectors of dimensions 831x1. For the output, we further extracted the functional groups in the compound molecule from the SMILES strings as a one-hot encoded vector representation.

METHODS AND RESULTS

In order to analyse chemical compound spectrograms, we employed two algorithms: Decision Tree and Convolutional Neural Networks. Beyond classification and recognition of functional groups, emphasis was also placed on how viable the algorithms are as more functional groups are introduced - especially in the Decision Tree experiment.

Decision Tree Experiment

Identification of functional groups from spectrograms is a task that can be done manually. In fact, high-school and college students are successfully taught to recognise the compound's functional groups (to some reasonable limit) by identifying peaks and patterns in their corresponding spectrograms. Due to the inherent decision-based logic of the problem, we trained a simple Decision Tree classifier to classify functional groups present in a chemical compound with input as the normalised log intensities in the UV-vis spectra of chemical compounds.

As mentioned before, we also utilised the model to find the threshold number of functional groups for classifier accuracy. In other words, we wanted to repeat the classification experiment with multiple classes (functional groups) to identify the number of classes at which the classifier breaks down in terms of accuracy.

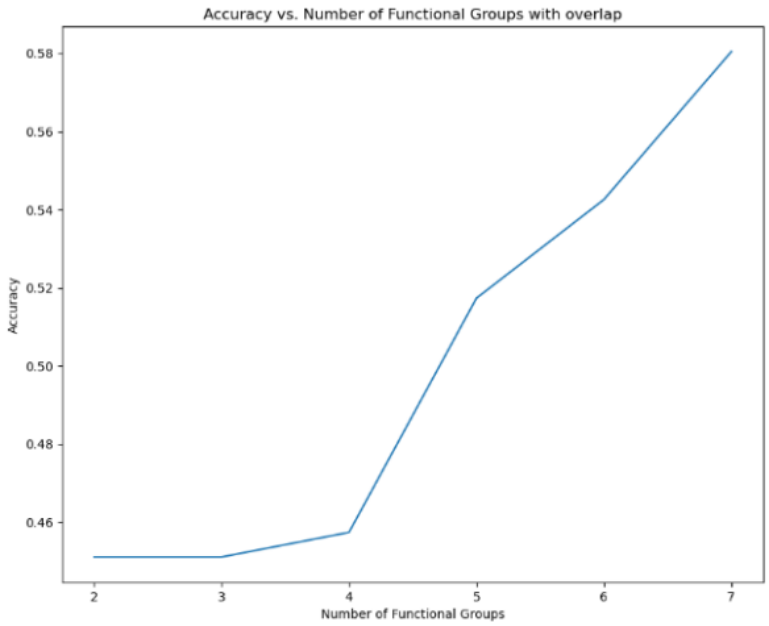


Figure 1. Accuracy change with functional groups

The above is an unexpected output as accuracy of the classifier seems to be improving the more classes you add. An important feature of this experiment was how the classes were generated. For the experiment with n functional groups, the n most common functional groups present in the dataset were taken. This introduced a $n+1^{th}$, 'complex' class which consisted of compounds in which all n groups occurred

together. The final, $n+2^{th}$ class was the 'other' class in which all other compounds that didn't consist the n functional groups. An inherent problem with this design was that as the number of functional groups grew in size, so did the 'complex' class as the possibility of the 'common groups' occurring together increased. This caused a class imbalance to grow with the number of functional groups which led to the classifier learning to over-predict the majority - 'complex' class which explains the increase in accuracy. This was subsequently verified through inspection of confusion matrices.

In order to find the functional group threshold for the classifier, we decided to remove the overlap 'complex' class completely. Any data points that contained multiple functional groups from the other classes were discarded.

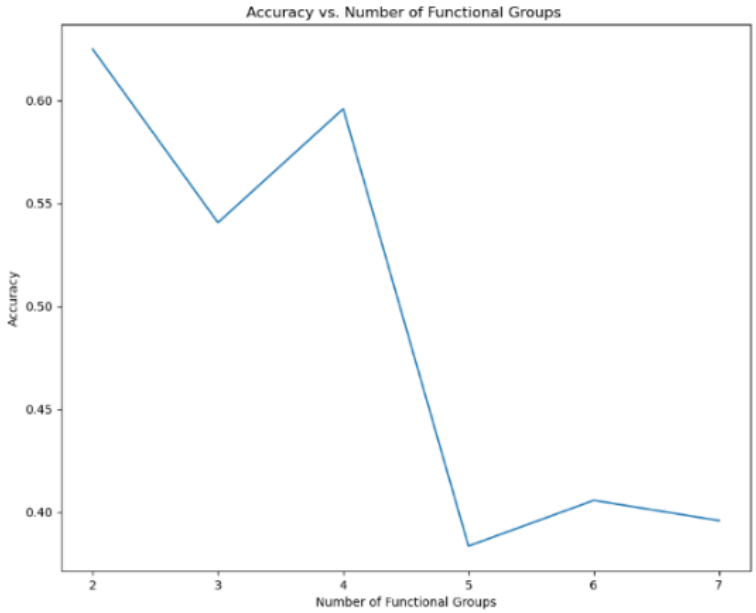


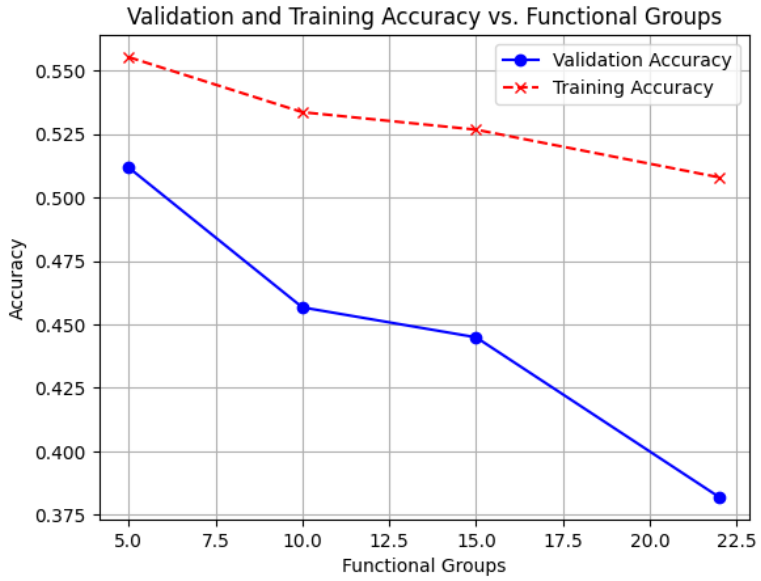
Figure 2. Accuracy change with functional groups without overlap class

The above is a more predictable outcome of the decreasing accuracy over the number of classes. The uncharacteristic spike in between 3 and 4 groups was caused by the model starting to learn the strategy of over-predicting the majority class but the strategy does not fare well when additional groups are added. The above experiment tells us that the functional group threshold of the decision tree classifier is around 3-4 groups.

CNN Experiment

We also tried to adapt similar existing spectra models particularly to UV-vis dataset. A sequential CNN model with convolutional, batch normalization, max pooling, flattening, dense and dropout layers fine-tuned using Bayesian optimization was used to predict the one-hot encoding of functional groups present in a compound from its normalised log intensity data from its UV-vis spectrum.

The model was modified to identify the 5, 10, 15, and 22 most common functional groups. While the training accuracy evenly decreases from 0.56 to 0.51, the validation accuracy drops drastically from 0.51 to 0.38 as the functional groups are increased from 5 to 22. This is in contrast to a similar model employed for IR-spectra, which achieved a much higher accuracy (0.96).



We also employed a similar sequential CNN model to predict the entire Morgan fingerprint from the normalised log intensities of the compound, which yielded very low and unstable accuracy.

DISCUSSION

Over the course of the project we have applied machine learning techniques to gain insights into chemical compounds through their spectra. We have laid the groundwork for further analysis by identifying the optimal number of functional groups for classification using decision trees. The experiment has also brought into light the difficulty in functional group identification due to the extent of overlap in these compounds. Most compounds in our dataset contain multiple functional groups so additional detail needs to be attributed to this problem. A potential way to address this problem could be multi-label based classification algorithms. Additionally, we observed a significant discrepancy in the data available for different spectra. Models that worked well with a considerably larger IR spectra (30,000+ data points) dataset performed poorly with the UV-vis spectra data, indicating the need for differentmodels altogether or a larger training set to adapt IR spectra models to UV-vis models efficiently. We believe there is a lot more to be done in this area. Further research could look into better representation of the input spectra - such as intensity values for a compound, log intensity, normalised log intensity, or only the peaks in the spectra. Different representations of the compounds besides SMILES strings and molecular fingerprints could be explored. Functional groups could be classified using meta-learning algorithms and their recognition task could also be looked at as a multi-class problem. Finally, research could look into the possibility of using the presence of functional groups to determine, partially or wholly, the characteristics of the IR and UV-Vis spectra of compounds, as well as going from the details of the IR spectrum of a compound to predicting its UV-Vis spectra details and vice versa.

REFERENCES

Jung, G., Jung, S. G., & Cole, J. M. (2023). Automatic materials characterization from infrared spectra using convolutional neural networks. *Chemical science*, 14(13), 3600–3609. https://doi.org/10.1039/d2sc05892h