A Python Algorithm for Identification of Organic Functional Groups in SMILES codes

William Riddle, Computer Engineering
Dr. Bohdan Schatschneider, Chemistry and Biochemistry

Abstract

Organic functional groups (FGs) are a promising structural descriptor for molecules that can develop parsimoniuos, quantitative structure properties relationships (QSPRs) aimed at the prediction of specific properties in organic materials. The property of focus in this project is on the molecular optical gap energy (E_g^{H-L}). FGs are defined by their unique pattern of bonded atoms which appear as small side chains of connected atoms in molecules. Their bonded makeup can be identified using the Simplified Molecular Input Line Entry System (SMILES). This project developed a python algorithm which extracts the number of functional groups which appear in a molecule based on the SMILES code and used this output to draw functional group QSPR trends with the E_g^{H-L} for a set of 831 organic molecules.

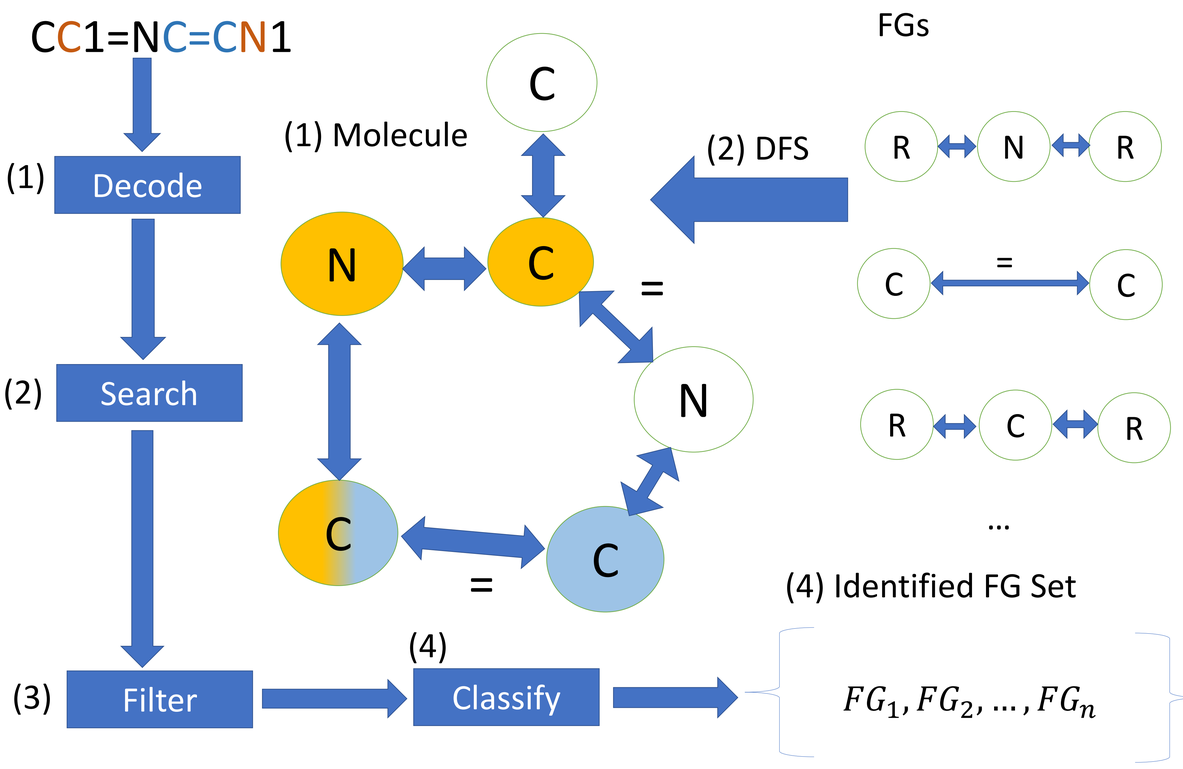
SMILES & Functional Groups

- A SMILES code is an alphanumeric string code of symbols which describes the bonded makeup of a molecule in space, including its rings and charges
- FGs appear as sub patterns in the SMILES string
- FGs are defined using Psuedo-SMILES templates
- Example templates are shown below:

[R]C(=O)O[R]	Ester
[R]N([R])C=O	Amide
C#C	Alkyne

- FGs and molecules defined by the SMILES can be decoded for their bonded makeup using the endowed properties of the SMILES
- A universal decoder can be developed for SMILES

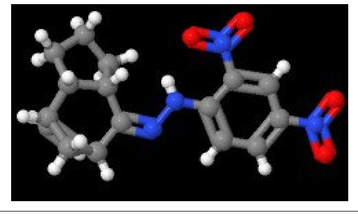
Implementation



(1) SMILES code decoded into a molecule of atom nodes and bond paths
(2) Depth First Search (DFS) algorithm identifies FGs in molecule
(3) Data correction filters are applied to correct search algorithm inaccuracy
(4) Functional groups are classified by their cyclic properties for final output

Visual Snapshot

3D Molecule



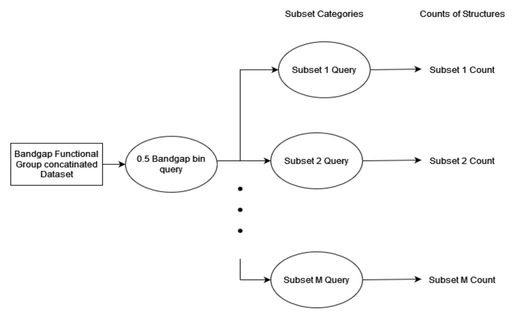
Algorithm Output

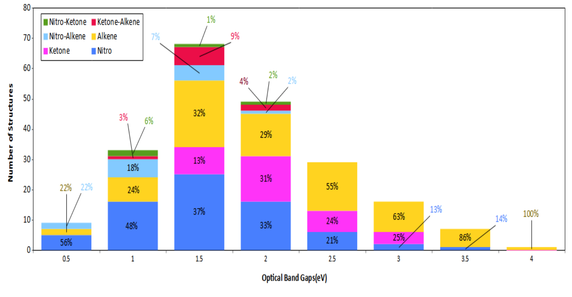
Name	Template	Count
Nitro	RN(=O)=O	2
SecondaryAmine	RNR	1
CyclicSecondaryKetimine	RC(R)=NR	1
CyclicAlkene	C=C	1

O=N(=O)c1ccc(NN=C2C3CC(C=C3)C3CCCC23)c(c1)N(=O)=O
Example SMILES code connected in space with string level FGs highlighted and algorithm output of FGs counted from algorithm

E_g^{H-L} Trend Analysis

The E_g^{H-L} for the 831 molecules were plotted as a function of functional groups present in their structure. Histogram plots visually reveal the QSPR correlations between FGs present and the E_g^{H-L}s of molecules. Ketones, Nitros, and Alkenes are notable FGs which lower the E_g^{H-L}





Conclusion

- Open-Source python module for SMILES codes and FGs at www.github.com/wtriddle/IFG
- Statistical programs for QSPRs
- Extensions to research can apply machine learning for prediction of E_g^{H-L} or other properties based on FGs