Generation of isomers for fast development of molecular datasets

Anna Catton, Swansea University, 1902838@swansea.ac.uk
Supervised by Dr Francisco Martin-Martinez, f.j.martin-Martinez@swansea.ac.uk

Skills 4 Scientists

Sponsored by: optibrium, dotmatics knowledge solutions

1 – Introduction

- Biocrude oils from hydrothermal liquefaction (HTL) of biomass, and other complex mixtures of organic molecules, are challenging to characterize. [1] Only molecular types are described
- Molecular models can be developed by automating the generation of molecular isomers of those molecular types and optimizing their geometry with Density Functional Theory (DFT) methods

2 – Methods

- Python scripts with Rdkit [2] and Atomic Simulation Environment [3] (ASE) packages
- DFT calculations B3LYP functional and 6-31G basis set
- ORCA software used for DFT calculations

3 – Discussion (on going work)

- Initial experimental data on molecular composition of two different types of biocrude oils: fast HTL and Isothermal HTL
- Concatenation of different python scripts following Figure 1
- High throughput generation of structural and stereoisomers
- DFT optimization of structures to generate a data base that could train machine learning (ML) algorithms

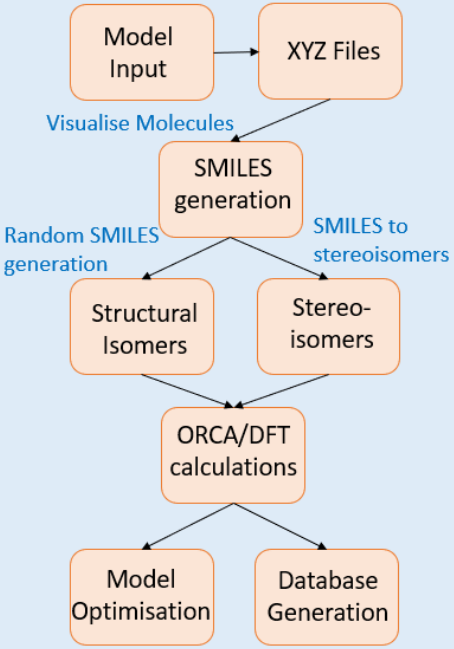

Figure 1: Flowchart to show the process of developing the molecular dataset

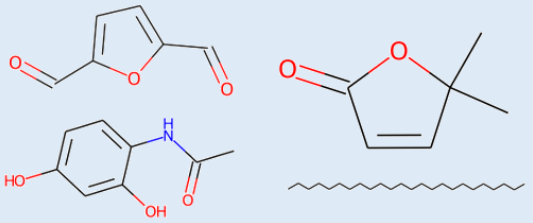

Figure 2: A sample of molecules visualised from the fast HTL dataset

4 - Future Work

- Integrate all scripts into one and debug.
- DFT calculations to be performed
- Model optimisation
- Data set generation for training ML algorithms

References

(1) D. Barreiro, F. Martin-Martinez, C. Torri, W. Prins, M. Buehler, *Algal Research*, 2018, **35**, 262-273
(2) RDKit, https://www.rdkit.org/, Accessed 19th July 2021
(3) Atomic Simulation Environment, https://wiki.fysik.dtu.dk/ase/ , Accessed 19th July 2021

AI³ Science Network+

PSDS PHYSICAL SCIENCES DATA-SCIENCE SERVICE

RSC INTEREST GROUP CHEMICAL INFORMATION AND COMPUTER APPLICATIONS GROUP