Abstract

This project uses machine learning to accelerate the discovery of new organic materials for solar energy applications by focusing on singlet fission. The core dataset (SMILES.csv), which is a set of known singlet fission materials (265 SMILES strings) is obtain from research article (A. Singlet fission molecules among known compounds: finding a few needles in a haystack).

To expand training dataset we perform data augmentation method which is fragment shuffling using RECAP (Retrosynthetic Combinatorial Analysis Procedure) algorithm on core dataset (SMILES.csv. This fragment shuffling method generated 400K augmented SMILES (augmented.csv) using RECAP algorithm. Additionally, we used GDB17 database which contains 166.4 billion small organic molecules (up to 17 atoms of C, N, O, S, and halogens) that follow basic chemical rules, from this large database, we randomly selected 100K SMILES (GDB17.csv).

To generate new SMILES which undergo singlet fission we employed 2 deep learning models which is pre-train model and fine tune model. First, we pre-trained a character-level LSTM model on the combined dataset of 400K augmented SMILE & 100K GDB17 molecules (augmented.csv & GDB17.csv). This pre-training phase allowed the model to learn fundamental chemical patterns of SMILE strings and molecular structures. After that, we fine-tuned this pre-trained model using only our original, singlet fission dataset (SMILES.csv). This fine-tuned training helped the model focus on the specific structural features that enable singlet fission. The fine-tuned model can then generate novel, chemically valid SMILES strings that are likely to exhibit singlet fission properties.

The newly generated SMILES/molecules is validated using RDKit package to make sure chemical feasibility, followed by filtering to eliminate duplicates and select the most candidates for DFT analysis.

GitHub Link: https://github.com/christy726/SingletFission_ML