Spectra Prediction for the Excitation and Emission of Dyes and other Conjugated Organic Molecules

Joe Abbott [1], Ryan Beck [1], Hang Hu [2], Yang Liu [1], Lixin Lu [1]

[1] Department of Chemistry, University of Washington, Seattle, WA 98195

[2] Molecular Engineering & Sciences Institute, University of Washington, Seattle, WA 98195

Overview

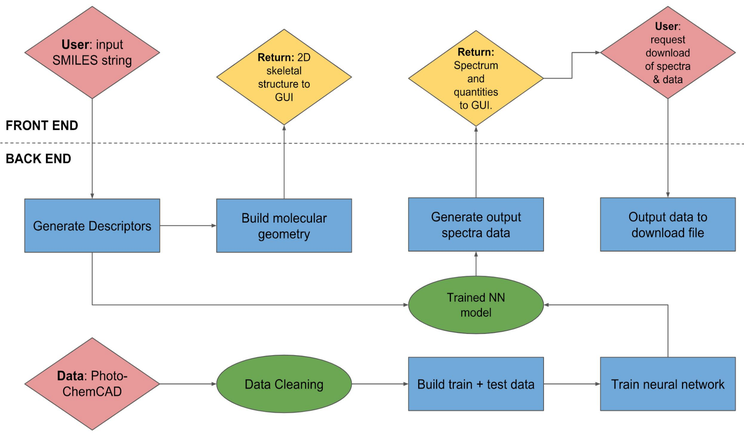
SPEEDCOM is an open-source python package that uses deep learning methods to predict the absorption and emission spectra of small organic molecules.

GitHub: https://github.com/emissible/SPEEDCOM

Motivations

The use of *ab initio* methods to calculate molecular spectra is usually lengthy, expensive, and may even be inaccurate depending on the choices for the level of theory. As such, a fast, experimental, data-derived method for predicting excitation and emission spectra for organic species is proposed to aid in rapid prediction of spectral features. This has potential uses in applications such as fluorophore-design.

Use Cases



***Via* a GUI, users can...**

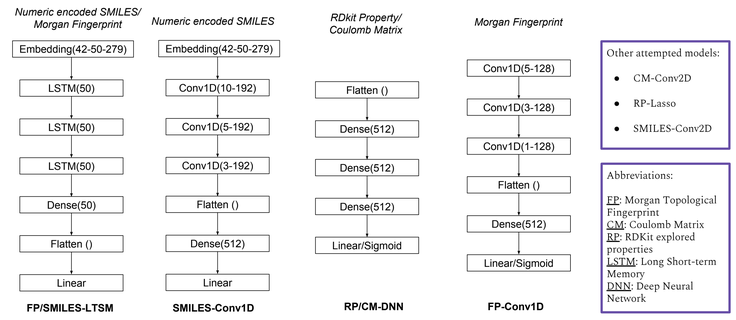
- Input the SMILES string of a given molecule
- Visualize the 2D skeletal structure of this molecule
- Visualize and download predicted spectra and associated characteristics such as the quantum yield and molar extinction coefficient.

Data Cleaning

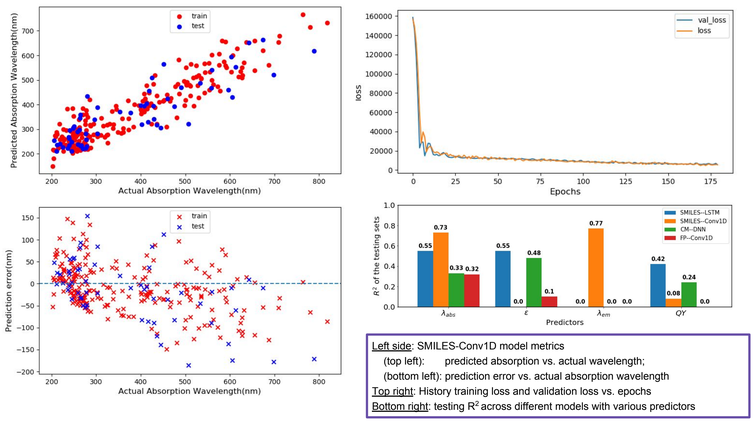
To obtain the dataset used for training, the files from the database were parsed to:

- Obtain the absorption and emission spectra;
- Obtain the smiles strings for molecules using *pubchempy* package;
- Removing extraneous counter ions from generated SMILES strings;
- Generating descriptors using *RDkit* package:
  - Coulomb Matrix of nuclei
  - Morgan Topological Fingerprint
  - Molecular Properties

Model Architectures



Metrics

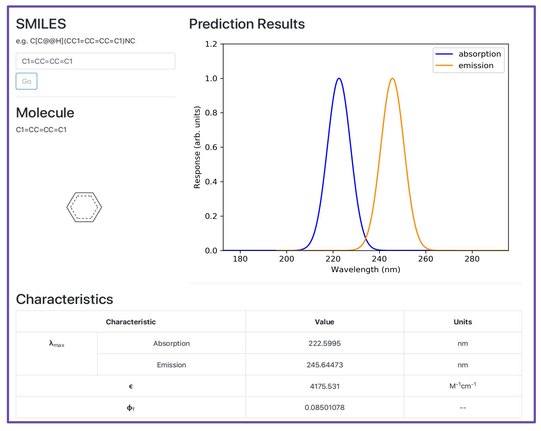


Left side: SMILES-Conv1D model metrics
(top left): predicted absorption vs. actual wavelength;
(bottom left): prediction error vs. actual absorption wavelength
Top right: History training loss and validation loss vs. epochs
Bottom right: testing R^2 across different models with various predictors

Results & Discussion

- The proposed package frameworks were built successfully with intended functionalities and have achieved $R^2 > 0.7$ for wavelength prediction with validation data
- Multidimensional property exploration was performed for the molecules included in the database
- The structural information encoded in SMILES/ connectivity fingerprint/ Coulomb matrix can be used to calculate spectroscopic properties
- The accuracies of our models are largely limited by the small size of the dataset, and the complexity of the problems.
- With the pre-trained models weights, the prediction speed can be guaranteed, while the fine-tuned accurate models still remain as biggest challenges.

Example GUI:



Future Work

- Sanitize SMILES input; add alternative input options
- Expand database and tune parameters for more accurate models
- Include multiple features in predicted absorption/emission spectra
- Allow users to train models with their own data
- Add a feature for pipelining predictions

References

Data: PhotoChemCAD (http://www.photochemcad.com/PhotochemCAD.html)

Dependencies: Keras, TensorFlow, RDKit, Pandas, Numpy, PubChemPy (all open-source).

Publication: Garrett B. Goh et al. 2018. SMILES2vec. In Proceedings of ACM SIGKDD Conference, London, UK, Aug, 2018 (KDD 2018), 8 pages

SPEEDCOM would like the thank UW DIRECT and the Clean Energy Institute at the University of Washington.