Spectra prediction models

Training data

- 989 compounds with experimentally obtained UV Vis spectra
  - From SRI’s internal collection
  - And collection purchased from OTAVA chemicals
- Split
  - 90% training and 10% holdout validation set
- All models trained using same train + test sets, then compared independently on the validation set

Models

4 ML models they tested

- UVvis-Schnet
- UVvis-DTNN
- UVvis-MPPN
- UVvis-Transformer

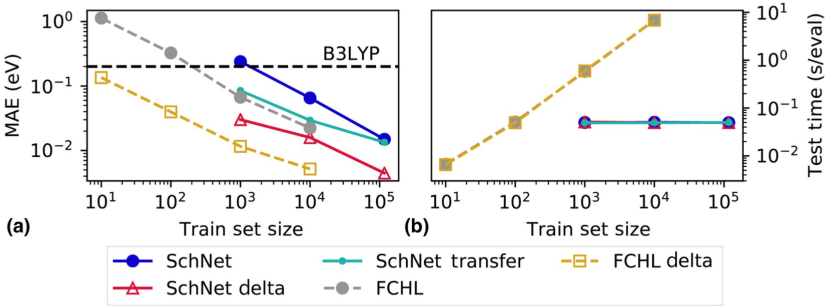
All models are adapted from previously created models. None of the adapted models are publicly available, however, most of the original models are.

Original models were modified to predict UVvis spectral features from a collection of molecular inputs (SMILES, 3D coordinates, QM-UV-vis predicted spectra)

UVvis Schnet

- Adapts the physics-informed method outlined in this paper
    - Uses Schnetpack with Δ-machine learning approaches
    - Trained with:
        - QM optimized 3D coordinates (their github provides workflow for physics calculations)
        - QM predicted UV-vis spectra
        - Experimentally obtained UV-vis spectra
- Once trained, model only requires 3D coordinate files of the desired molecules to predict UV vis spectra

Original Schnet Model



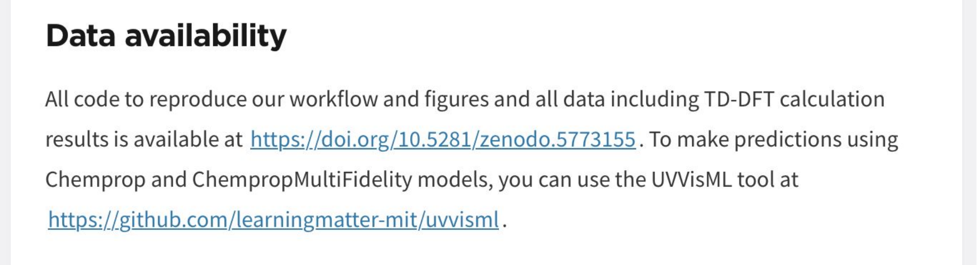
- Used QM9-G4MP2 dataset
  - The identities, molec coordinates, and computed properties of molecules in G4MP2 and the exact train/test splits used for training are available on this github
- “All source code needed to create models using [Schnet] approaches is available in [the] GitHub repository that includes scripts with the hyperparameter choices for our models, results showing that we replicate previous literature, and the exact versions of SchNetPack used in our study”
- Created multiple Schnet-based models. Schnet delta performed the best (lowest MAE)
- Schnet delta
  - train a model that learns the difference between B3LYP and G4MP2 energies.
  - use molecular/atomic properties computed with B3LYP as model inputs

UVvis DTNN

- Deep tensor neural network utilizes QM optimized structures for the input molecules to predict the UV vis absorption spectra
- Adapts model developed in this paper
    - Embeds atoms in each molecule and then calculates the interactions between them
    - Represents it as an interaction tensor
    - Each atom has its own interactions, which helps describe interactions in terms of
        - Interatomic distances
        - Dihedral angles
        - Higher-order interatomic relations
- Adapted above method to use 3D coordinates and predict UV-vis absorption spectra

UVvis MPPN

- Model adapted from Chemprop-IR, IR spectra prediction model
  - Relies on 2D information in the form of molecular graphs derived from SMILES to create features and produce predictions
- Incorporated functionality for UV-vis spectra prediction to consider 3D features
- UV-vis peak prediction model using Chemprop was recently developed and published (this paper)
  - Contains list of existing datasets of experimental UV vis spectroscopic properties and computed excitation energies
  - github



Data availability

All code to reproduce our workflow and figures and all data including TD-DFT calculation results is available at https://doi.org/10.5281/zenodo.5773155. To make predictions using Chemprop and ChempropMultiFidelity models, you can use the UVVisML tool at https://github.com/learningmatter-mit/uvvisml.

UVvis Transformer

- Adapted from existing SolTranNet code, which predicts aqueous solubility values from SMILES
  - SolTranNet is distributed via pip
  - SolTranNet source code
- Datasets and scripts used in SolTranNet paper are available in this github

Determining optimal model

- Used RMSE and R2
- Determined that there was no truly optimal model for predicting UV-vis spectra predications, but relative to the models used UVvis-MPPN is most promising
    - RMSE of 0.06 (UV vis Schnet also had a comparable RMSE of 0.07)
- Best performance with features of experimental spectra, QM predicted spectra, and 3D QM mechanically optimized coordinates