№15.2: Предсказание частот поглощения и испускания молекул с использованием графовых нейросетей

Введение:

Оптические и фотофизические свойства хромофоров и флуорофоров важны для промышленного и академического применения в различных областях исследований, таких как органические солнечные элементы, светодиоды, датчики и органические красители. [1]

Методы:

Подготовительные работы: Исходные данные были представляли из себя таблицу в .csv формате. Удаление выбросов, NaN для целевых колонок, молекул слишком высокой молекулярной массы и содержащих йод, и последующая обработка были проделаны при помощи pandas. Визуализация и построение графиков выполнено с помощью matplotlib.pyplot.

Выбросы были обнаружены в большом количестве в колонке с молекулярной массой молекул. Удаление производилось, если значение выходило за пределы 1.5*IQR от 0.25 и 0.75 квартилей.

1. **Поиск стабильных конформаций**

Формулы SMILES были преобразованы в 3D-конформации при помощи пакета RDkit. Было сгенерировано по 10 конформаций для каждой формулы (EmbedMultipleConfs), после чего выполнялась их оптимизация с использованием силовых полей (при помощи UFFOptimizeMoleculeConfs) и выбор наилучшей конформации, обладающей минимальной энергией. Результаты были сохранены в формате ASE Database [7].

Расчёт векторных представлений молекул (эмбеддингов)

Расчет векторных представлений молекул проводился при помощи графовых нейронных сетей из пакета nablaDFT [2], а также MACE. Для этого были взяты предобученные модели:

PaiNN: "PaiNN_train_large_traj_medium": "https://a002dlils-kadurin-nabladft.obs.ru-moscow-1.hc.sbercloud.ru/data/nablaDFTv2/models_checkpoints/PaiNN/painn_100k_traj_10k.ckpt" [3]

SchNet: "SchNet_train_large": "https://a002dlils-kadurin-nabladft.obs.ru-moscow-1.hc.sbercloud.ru/data/nablaDFTv2/models_checkpoints/SchNet/schnet_100k.ckpt" [4]

DimeNet++: "DimeNet++_train_large":

"https://a002dlils-kadurin-nabladft.obs.ru-moscow-1.hc.sbercloud.ru/data/nablaDFTv2/models_checkpoints/DimeNet%2b%2b/DimeNet%2b%2b_dataset_train_100k_epoch=0258.ckpt" [5]

MACE: “MACE-OFF23: Transferable Organic Force Fields”: MACE-OFF23_large.model https://github.com/ACEsuit/mace-off/blob/main/mace_off23/MACE-OFF23_large.model [6]

Для учета того, что описываемые среды представляют собой системы из хромофора и растворителя, агрегированный эмбеддинг системы был получен путем конкатенации эмбеддингов хромофора и растворителя.

Результаты:

В сыром датасете представлено 18386 уникальных сочетаний хромофор-растворитель для тренировочных данных. После очистки и подготовки количество данных в обучающем датасете сократилось до 10337 сочетаний хромофор-растворитель. Для тестовых данных после подготовки данных количество сочетаний хромофор-растворитель сократилось до 1115 с 1850.

Перевели уникальные молекулы хромофоров и растворителей из SMILES в формат Ase Database для работы с конформациями. Для ~30 конформаций не получилось перевести структуры SMILES в 3д, также не для всех атомов подбирались корректные состояния гибридизации, заряды, кроме того RDkit не определял атомы Se2+. Таким образом, было получено 4384 уникальных оптимизированных конформаций для хромофоров и 263 для растворителей в данных для тренировочного сета и 898 для хромофоров и 48 для растворителей.

Предсказание частот испускания, поглощения и квантового сдвига

Для предсказания таргетных свойств был написан MLP, в который подавались эмбеддинги и таргетные свойства.

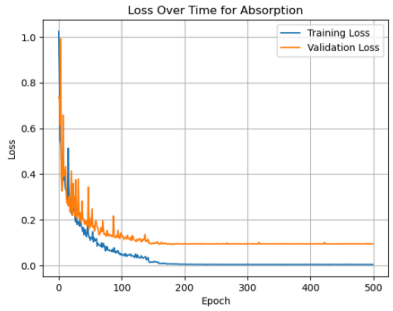
```
MLPModel(
  (mlp): Sequential(
    (0): Dense(in_features=256, out_features=512, bias=True)
    (1): Dense(in_features=512, out_features=256, bias=True)
    (2): Dense(in_features=256, out_features=128, bias=True)
    (3): Dense(
      in_features=128, out_features=1, bias=True
      (activation): Identity()
    )
  )
)
```

Для изменения скорости обучения использовался ReduceLROnPlateau, который заключался в снижении lr в 2 раза каждые 10 эпох, в которые не

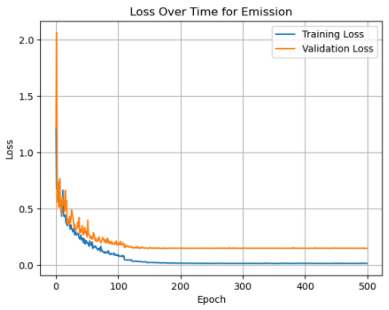
происходило снижения лосса.

Далее приведено сравнение результатов обучения при использовании эмбеддингов, полученных с помощью различных графовых нейронных сетей.

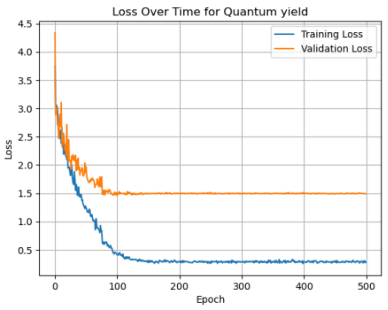
Schnet:



R^2: 0.937
RMSE: 24.226

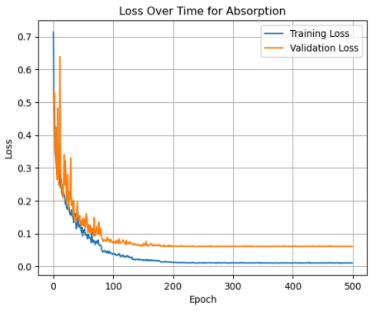


R^2: 0.848
RMSE: 36.761

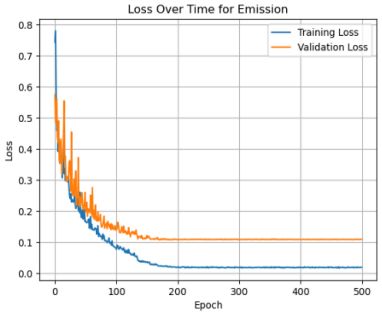


R^2: 0.551
RMSE: 1.223

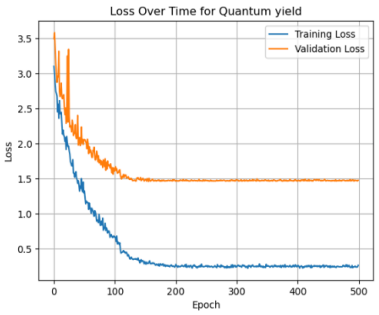
PaiNN:



R^2: 0.941
RMSE: 23.480



R^2: 0.889
RMSE: 31.558



R^2: 0.578
RMSE: 1.185

Dimnet++:

При обучении возникли ошибки (Loss = NaN), при этом проверка эмбеддингов и таргетов на наличие NaN не помогла решить ошибку.

MACE:

Предобученная модель, использованная для расчета эмбеддингов, была обучена на датасете, состоящем из малых молекул, имеющих в своём составе атомы 10 типов: H, C, N, O, P, S, F, Cl, Br, I. В предложенном датасете имелись молекулы, содержащие в том числе атомы других типов, поэтому при расчете mace-эмбеддингов данные были дополнительно отфильтрованы, что привело к потере некоторой части данных. Обучение на полученных данных дает плохие метрики, возможная причина – некорректное составление глобального эмбеддинга каждой молекулы. Возможно, необходима свёртка тензора локальных эмбеддингов, а не stack.

Выводы:

Запустив несколько моделей, заметили, что предсказания напрямую зависят от описания систем, т.е. эмбеддингов. Лучший результат для всех 3 таргетных величин (абсорбции, эмиссии, квантового выхода) показала модель PaiNN.

Источники литературы:

1. Joung, J. F. et al. Deep Learning Optical Spectroscopy Based on Experimental Database: Potential Applications to Molecular Design. JACS Au 1, 427–438 (2021).
2. https://github.com/AIRI-Institute/nablaDFT
3. http://proceedings.mlr.press/v139/schutt21a/schutt21a.pdf
4. https://pubs.aip.org/aip/jcp/article-abstract/148/24/241722/962591/SchNet-A-deep-learning-architecture-for-molecules?redirectedFrom=fulltext
5. https://www.cs.cit.tum.de/daml/dimenet/
6. nhdPreprint: MACE-OFF23: Transferable Machine Learning Force Fields for Organic Molecules: https://arxiv.org/abs/2312.15211
7. https://docs.datamol.io/0.9.1/tutorials/Conformers.html