Graph Neural Networks for Predicting Absorption Spectra: Unraveling Molecular Insights through advanced Message Passing Framework

Shubham Kumar Pandey
Department of Chemical and Biological Engineering
University of Idaho
Moscow, Idaho, United States 83843
pand2623@vandals.uidaho.edu

***Abstract*— UV-visible spectroscopy is a technique used to identify unknown chemicals. It works by measuring the light absorbed by a sample in a range from 190 to 900 nanometers. Knowing the unique wavelength at which each chemical absorbs the maximum light, called maximum absorption (λ_{max}), is crucial. To predict this absorption accurately, a deep learning (DL) model is designed and trained on experimental data. The model uses a message passing neural network (MPNN), a type of graph neural network (GNN), which learns from molecular graph features drawn using RDKit library. With dense neural network layers, the model achieves a good agreement, with a root mean square error (RMSE) of 21 and 30 nm for the training and test sets, respectively. This predictive model proves effective in guessing the absorption of unknown compounds and may be helpful in identifying and discovering new chemical entities.**

Keywords—SMILES, GNN, Transformer, MPNN, Absorption (λ_{max}).

INTRODUCTION

Ultraviolet and visible (UV-Vis) absorption spectroscopy is the measurement of the attenuation of a beam of light after it passes through a sample or after reflection from a sample surface. This article uses the term UV-Vis spectroscopy to encompass a variety of absorption, transmittance, and reflectance measurements in the ultraviolet (UV), visible, and near-infrared (NIR) spectral regions. These measurements can occur at a single wavelength or over an extended spectral range.[10]

Ultraviolet-visible spectroscopy is one of the most ubiquitous analytical and characterization techniques in the field of science and discovery (fields – plant toxins, phytomedicines, cosmetics, agrochemicals, food additives, dyes, paints, drugs, irritants, etc.). There is a linear relationship between absorbance and the concentration of the sample, which allows this spectroscopy technique to analyze the sample both qualitatively and quantitatively. The principle behind UV-vis spectroscopy is that it promotes the excitation of electrons from the ground state to a higher energy state when a certain wavelength of light is absorbed by the compound. Therefore, it is important to note the maximum absorption, which is a characteristic feature of a chemical or biological compound.[11]

There are limited ways we could replace the use of this sophisticated equipment and use some theoretical concept to determine the chemical composition and strength of the compound. One of them is using quantum mechanics approach called time dependent Density Functional Theory (TDDFT) calculation that trace the excitation of electron in the compounds at certain frequency of light (inverse of wavelength). Regardless of its accuracy this method cannot be applied more often due to its high computational demand and the tedious nature of the process.[12]

Machine learning (ML) techniques now have been reported for the prediction of several physicochemical and analytical properties of the chemical compounds that includes predicting UV-vis absorption values as well. These models are trained with experimental values obtained from the literature but also with theoretically calculated data (TDDFT) for faster and precise estimation of molecular properties. The closet prediction them have recoded is root mean square error (RMSE) of 26 nm for the absorption.[13] In the same context, this work uses deep learning technique called message passing neural network (MPNN) which is an undirected learning and training architecture that learn to encode the key chemical features from the molecular graph representation of the molecule.[14] This architecture extract information in the form of molecular graphs derived from SMIELS to create atom and bond (vertex and edge) features and make prediction (see in Figure 1).

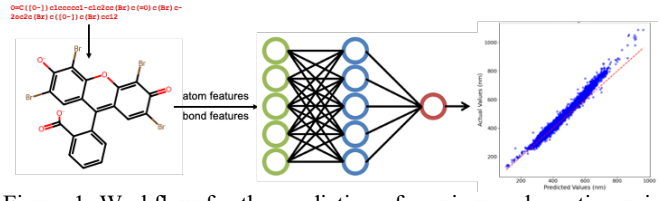

Figure 1. Workflow for the prediction of maximum absorption using MPNN. Workflow depicts four steps, conversion of SMILES into molecular structures, feature acquisition, model development and prediction.

METHODS

Message Passing Neural Networks (MPNNs)

In the MPNN model, which is a subset of Graph Neural Network (GNN), molecules are represented as an undirected graph with node/vertices that correspond to atom features, and the edge represents the bond features.[1] They operate in two phases: a message passing phase, which transmits information across the molecules to build a neural representation of the molecule, and a readout phase, which is the final representation of the molecule to make predictions about the properties of interest which is our case maximum absorption (λ_{max}).

For each vertex *v,* there is an associated hidden state and their message at every time step t.

$$m_v^{t+1} = \sum_{w \in N(v)} M_t(h_v^t, h_w^t, e_{vw})$$

$$h_v^{t+1} = U_t\,(h_v^t, m_v^{t+1})$$

Where, e_{vm} *is* the edge feature, M_t is the message function, U_t is the vertex update function, and *N(v)* is the neighbors of v in the graph.

In the readout phase, a readout function is applied to the final state h_v^T to make the prediction.

$$\hat{y} = R(h_v^T | v \in G)$$

Here G represents molecular graph. The above equation implies to extract and process node features only however in this study both node and edge (bond) features $h_{e_{vw}}^t$ was used in MPNN by introducing them into the hidden states and updating them as mentioned in defined equations.

Features of a Chemical Component

Two main classes of features have been chosen that can extract maximum chemical information using the proposed node and edge features. These features will provide sufficient details about the atoms (vertices) using list of **atom features,** present in the chemical structure and their relationship with neighboring elements through bonds (edges) through **bond features**.[7]

The atom and bond properties used in this study are listed in the tables below:

Table 1. List of atom features to be extracted and its description.

Features	***Description***
symbol	Type of atom present in the molecule ('C', 'N', 'O', 'S', others)
n_valence	Number of electrons in the outermost orbital of atom and can participate in bond formation
n_hydrogen	Number of bonded hydrogen atoms
hybridization	Mixing of atomic orbitals to form new hybrid orbitals suitable for pairing of electron to form chemical bond

Table 2. List of bond feature and its description.

Features	***Description***
bond_type	Type of bond (single, double, triple, aromatic)
conjugated	Whether bond is conjugated (repeated double and single bonds) or not

The third set of features used pair indices to keep track of pairs of atoms involved in interactions within a molecule. This information was crucial for the MPNN, where information is propagated based on how the nodes/atoms are interacting within a molecule.

Graph from SMILES

Functions are defined to work together in converting the SMILES representation of molecules into features suitable for input to an MPNN model. The first function, ***'molecule_from_smiles,'*** is used to return the molecular structure of the corresponding SMILES of the molecule. This ensures that the molecular structure generated using this function follows chemical conventions, and with the help of the sanitization feature of this function. This process of checking and adjusting a molecular structure eliminates the chances of having absurd molecular structures in the dataset.

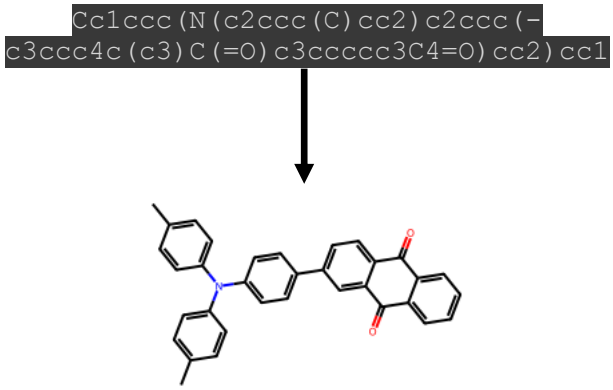

Figure 2. Schematic representation of conversion of 1D SMILES into molecular structure using defined function that uses rdkit library.

Following this, the second function, ***'graph_from_molecules,'*** is used to take the generated molecular structure from the first function and return atom features, bond features, and pair indices for the interacting atoms that count bonds between atoms and their neighbors which are used to generate list of features mentioned in Table 1 and 2.

The last function, '***graph_from_smiles,***' is used to append the respective lists of features generated for each of the molecules into one list and then convert it into a ragged tensor, which is helpful in processing variable-sized inputs.

MPNN Architecture

The defined MPNN comprises several parts, including a transformer encoder layer that operates on an attention mechanism to capture dependencies and relationships between different elements in the input feature list generated from SMILES. It serves as a feature selection tool in the architecture, identifying both low-level features and high-level abstractions.

Other components include a series of dense layers with many nodes for intense training and maximum learning. The strategy of arranging the nodes in descending order is employed to help the model capture both high- and low-level features. The architecture also incorporates a dropout layer, serving as a

regularization tool to introduce non-linearity and assist the model in avoiding overfitting problems.

Implementation

This model was implemented using several useful Python libraries at different stages. RDKit was employed to handle chemical structures, SMILES, convert chemical information from one format to another, and extract features from them.[4] Pandas and NumPy were utilized for cleaning, exploring, and analyzing the dataset, as well as handling multi-dimensional arrays and matrices. TensorFlow and Keras were employed for implementing neural network models (MPNN). Using these libraries, numerical computation and training neural networks were facilitated, offering high flexibility to change hyperparameters. Matplotlib was used to generate interactive plots, which were helpful for visualizing data and drawing any strong inferences.

EXPERIMENTS

Data Gathering

To train the model, a dataset was gathered from two different sources [2, 3]. A dataset containing 20,236 instances with experimental absorption values was obtained as the primary source. However, not all compounds listed in that dataset have absorption values; it was observed that a few instances have "NaN," which underwent removal in the data preprocessing steps. Another set of data with 1,181 entries was collected, and the majority of these entries exhibit absorption in the visible region ($\lambda_{max} > 500$). This helps maintain variability in the datapoints collected and aids the model in making predictions for a wide range of absorption spectra.

Figure 3. Plot represents distribution of absorption values in two different datasets collected and used combinedly in this project.

By combining the above-mentioned two datasets, the final dataset comprises a total of 21,417 SMILES of different compounds and their corresponding absorption (nm) values.

Experimental Procedures

- **Data pre-processing**
  When the final dataset is allowed to pass through the steps of data preprocessing, the treatment involved dropping any NA values from the dataset using the *dropna* function. This resulted in a reduction in the number of rows from 21,417 to 18,476.

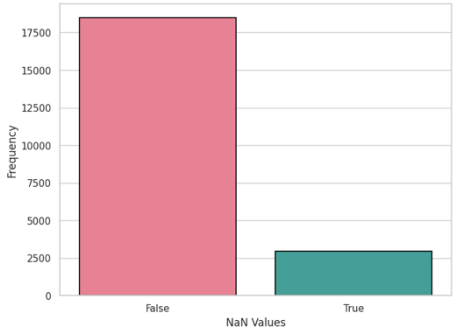

Figure 4. Bar chart represents number of missing values (NaN) in the dataset, that are removed in the data pre-processing step.

In the next stage, an attempt was made to cross-check for duplicate entries in the dataset, as they may be introducing bias to the model, along with other issues such as data leakage and misleading evaluation metrics. Hence, it was deemed mandatory to mitigate such anomalies from the dataset. After analysis and the removal of duplicate entries, it was determined that only 7,173 unique entries are available. This is attributed to the fact that the absorption values for most compounds were recorded in the presence of different solvents, causing changes in absorption within ± 5 nm. For this study, this information is not considered important and does not align with the objective of the research. Therefore, only one absorption value for each compound was retained.

- **Data Splitting**
  While there are no hard-set rules defining the optimal ratio for data splitting into training and test sets, the 80-20 Rule (also known as the Pareto Principle) is renowned in the field of machine learning, asserting that 80% of consequences come from 20% of the causes.[5, 6] Keeping this in mind, we employed the 80/20 ratio, along with 70/30 and 90/10, to analyze the variations in results. The model was constrained to run for only five epochs, and the optimal ratio was determined based on the root mean square error (RMSE – explained in the next section) difference within the three ratios. In my case, the 80/20 ratio performed optimally and was selected for the study.

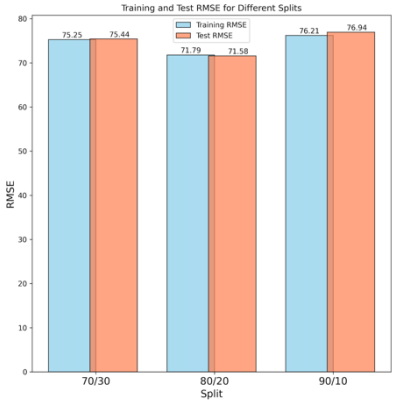

Figure 5. Bar chart represents model performance on training and test set at different data split proportion to find the optimal ratio.

Statistical Measure

To evaluate the performance of the prediction model, four different statistical measures were selected, as mentioned below:

- **Mean Absolute Error (MAE)**
  It calculates the mean absolute difference between the actual and predicted values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **Mean Squared Error (MSE)**
  It measures the average of the squared difference between predicted and actual values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **Root Mean Squared Error (RMSE)**
  RMSE is the square root of the calculated MSE is often used to express the error in the same unit as the target variable.

$$RMSE = \sqrt{MSE}$$

- **R-Squared (Coefficient of Determination)**
  It is represented by the proportion of the variance in the dependent variable that can be predicted from the independent variables.

$$R^2 = 1 - \frac{SSR}{SST}$$

Wherein, SSR is the sum of squared residuals, and SST is the total sum of squares. The value of R-Squared ranges from 0 to 1, with values close to 1 indicating a perfectly fitted model.

Model Training and Optimization

- **Hyperparameter Optimization**
  The architecture of this predictive model incorporates a message-passing and transformer-style mechanism for effective information aggregation and accurate prediction. During model training, several parameters were adjusted to achieve good agreement. In the architecture, padding, multi-head attention, two dense layers with ReLU activation function, and two-layer normalization are utilized in the transformer readout layer. For the fully connected dense layer, the model was initially trained with three layers, and introducing two additional layers resulted in comparably good results. The number of nodes is arranged in a descending order [512—256—128—64—32]. Dropout with a cutoff of 0.4 was applied within the first two layers.

  Throughout the training process, the Adam optimizer was used, as RMSprop and SGD required too many epochs to converge. The model was trained for up to 500 epochs with early stopping set to a patience of 5, although the training eventually completed the 500 epochs. This suggests that decreasing the learning rate from 0.0001 to 0.00001 and increasing the number of epochs could potentially yield even better performance for the model.

- **Baseline**
  Previously conducted projects illustrate how a conventional machine learning algorithm can be utilized to predict the maximum absorption value close to the experimental results when trained with the appropriate type of molecular features. A predictive model was developed by *Mai J, et al.,* using composition-based features and topological distance. When the model was trained using XGBoost, SVR, GBR, etc., the best score achieved was with XGBoost, yielding an RMSE of **23.4 nm**, whereas the result obtained from SVR was an RMSE of **63.6 nm**.[8]
  Another study was conducted by *Joung JF, et al.,* using small organic molecules. They employed a graph convolutional network (GCN) with atom-based features only and fed it to a multi-layered perceptron (MLP) to achieve an RMSE of **26–28 nm.**[9]
  Results obtained by *Shao J, et al.,* from intensive training of FCNN and CNN architectures using molecular fingerprints showed an RMSE of **18 nm** when run for 2000 epochs. However, when they attempted MPNN, the best score achieved was an RMSE **of 48 nm**.[3]

Results and Discussion

This study employs the deep learning techniques subset of machine learning, to predict the absorption of molecules based on their SMILES annotations. UV spectra serve as a powerful tool for investigating and identifying unknown compounds. Developing a machine learning/deep learning-based model accelerates the research process and reduces dependence solely on spectroscopy machinery.

The developed deep learning model employs MPNN to predict the maximum absorption of molecules more precisely. The foundation of this model is an experimental dataset comprising 7173 molecules with a diverse range of absorption values (nm).

This model is designed to be trained on self-generated features that encompass crucial information such as the type of atoms, bonding, their hybridization, neighboring atoms, etc. It not only utilizes GNN and MPNN but is also embedded with the Transformer Encoder Readout architecture, which helps the model better understand the context and input graphs coming from the molecules, followed by FCNN layers.

Despite the scarcity of additional experimental data, this model outperforms with metrics of **15.88, 21.87, and 0.97** (MAE, RMSE, and R-Squared) for the **training set** and **19.13, 30.07, and 0.94** for the **test set.** These results are comparable to previously conducted studies on larger datasets using complex architectures and high computational resources.

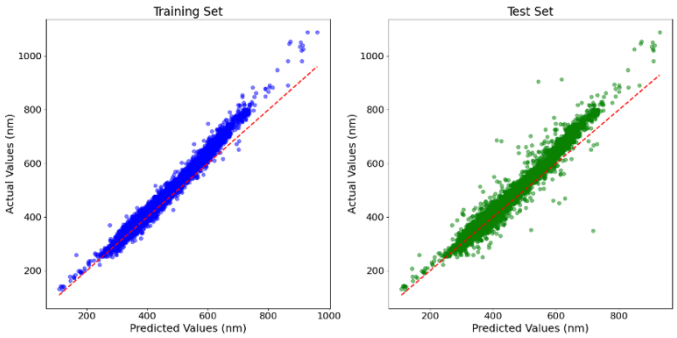

Figure 6. Scatter plot represents the performance of the model on training set (left - blue) and prediction on the test set (right - green).

ACKNOWLEDGMENT

This work was carried out and submitted for partial fulfillment of the course Deep Learning (CS 574), instructed by Dr. Min Xian. This course aimed to provide an understanding of the building blocks of neural networks and assist in implementing Deep Neural Network (DNN) techniques in real-world data. Graph Neural Network (GNN), introduced during Lecture 23 and 24 of this course, was employed in this study. During these sessions, the basics of GNN, the definition of matrices, and applications of GNN were learned. The study also incorporated the use of Message Passing Neural Network (MPNN), proposed herein for predicting the absorption of different chemical compounds.

The entire training process was conducted using Tesla T4 GPUs on the free version of Google Colaboratory for intensive training. Gratitude is extended to Dr. Min Xian and Kyle Lucke (TA) for their contribution to making the entire learning process easy and understandable.

REFERENCES

[1] K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M, Palmer A. Analyzing learned molecular representations for property prediction. Journal of chemical information and modeling. 2019 Jul 30;59(8):3370-88.

[2] Joung JF, Han M, Jeong M, Park S. Experimental database of optical properties of organic compounds. Scientific data. 2020 Sep 8;7(1):295.

[3] Shao J, Liu Y, Yan J, Yan ZY, Wu Y, Ru Z, Liao JY, Miao X, Qian L. Prediction of maximum absorption wavelength using deep neural networks. Journal of Chemical Information and Modeling. 2022 Mar 15;62(6):1368-75.

[4] Landrum G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. Greg Landrum. 2013;8:31.

[5] Nguyen QH, Ly HB, Ho LS, Al-Ansari N, Le HV, Tran VQ, Prakash I, Pham BT. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. Mathematical Problems in Engineering. 2021 Feb 5;2021:1-5.

[6] Sanders R. The Pareto principle: its use and abuse. Journal of Services Marketing. 1987 Feb 1;1(2):37-40.

[7] Soyemi A, Pandey SK, Vaara SA, Szilvási T. Predicting the melting point of liquid crystals with directed message passing neural networks. Liquid Crystals. 2023 Nov 4:1-5.

[8] Mai J, Lu T, Xu P, Lian Z, Li M, Lu W. Predicting the maximum absorption wavelength of azo dyes using an interpretable machine learning strategy. Dyes and Pigments. 2022 Oct 1;206:110647.

[9] Joung JF, Han M, Hwang J, Jeong M, Choi DH, Park S. Deep learning optical spectroscopy based on experimental database: potential applications to molecular design. JACS Au. 2021 Mar 17;1(4):427-38.

[10] Perkampus HH. UV-VIS Spectroscopy and its Applications. Springer Science & Business Media; 2013 Mar 8.

[11] Tissue BM. Ultraviolet and visible absorption spectroscopy. Characterization of Materials. 2002 Oct 15.

[12] Pescitelli G, Bruhn T. Good computational practice in the assignment of absolute configurations by TDDFT calculations of ECD spectra. Chirality. 2016 Jun;28(6):466-74.

[13] Mamede R, Pereira F, Aires-de-Sousa J. Machine learning prediction of UV–Vis spectra features of organic compounds related to photoreactive potential. Scientific Reports. 2021 Dec 9;11(1):23720.

[14] McNaughton AD, Joshi RP, Knutson CR, Fnu A, Luebke KJ, Malerich JP, Madrid PB, Kumar N. Machine Learning Models for Predicting Molecular UV–Vis Spectra with Quantum Mechanical Properties. Journal of Chemical Information and Modeling. 2023 Feb 27;63(5):1462-71.