2023

Machine learning for ultraviolet spectral prediction

Linh Ho Manh

MACHINE LEARNING FOR ULTRAVIOLET SPECTRAL PREDICTION

by

LINH HO MANH

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

May 2023

To my father Lam Ho and my mother Bac Nguyen Minh

who set the example and made me who I am.

ACKNOWLEDGEMENTS

I would like to thank my supervising professor Dr. Victoria Chen for constantly motivating and encouraging me and for her invaluable advice during the course of my doctoral studies. In particular, Dr. Chen encouraged me to think critically about the practical aspects of my research, including interpretability vs. more popular complex approaches.

I would also like to thank my dissertation committee members, Dr. Kevin Schug, Dr. Jay Rosenberger, and Dr. Bill Corley, for their support of my research and for taking their precious time to review my thesis and provide thoughtful comments. I wish to give a special thanks to Dr. Kevin Schug. He officially served as one of my dissertation committee members, but in many ways, he was a supervising professor. He proposed my dissertation topic and has provided continual guidance on all things chemistry-related.

I would like to acknowledge National Science Foundation grant CHEM-2108767 for providing financial support for my doctoral studies. I especially want to thank Dr. Kevin Schug, his colleagues, and VUV Analytics company for their interest in my research, their helpful discussions, and their patience in explaining chemical intuition to me. Their input allowed me to incorporate those ideas into machine learning models and achieve several interpretable results.

I am thankful to Dr. Shouyi Wang and Dr. Yi Zhang at the University of Texas at Arlington and Dr. Huihui Zhang at the United States Department of Agriculture for their support and encouragement. Through working on projects with them, I have improved my machine learning knowledge and programming ability. I would

like to thank Dr. Bill Corley, Dr. Jay Rosenberger, and Dr. Aera LeBoulluec for their engaging instruction in my courses, and I would like to thank all academic and technical staff in the Department of Industrial, Manufacturing, & Systems Engineering, especially Ms. Ann Hoang, Mr. Richard Zercher, and Ms. Cindy Royster for their support from my first day here.

I would like to thank all the teachers who taught me during the years I spent in school, first in Vietnam, then in Italy, and finally in the United States. I would like to send my special thanks to Dr. Riccardo Zich, Dr. Paola Pirinoli, and Dr. Farhan Qazi for their encouragement and inspiration to pursue a Ph.D. degree in the United States. During my time at the University of Texas at Arlington, I enjoyed discussing FLASK programming with my younger brother Dang Ho Ha, and I thank him for his constant presence in the family while I am not in Vietnam.

Lastly, I want to express my deepest appreciation to my family. To my mother, Bac Nguyen Minh, who has always supported and ensured my financial stability during my studies, I am forever grateful. I am also immensely thankful to my wife, Mai Le Tuyet, and my daughter, Mai Anh Ho, for their sacrifice, encouragement, and unwavering patience. This thesis is dedicated to my father, Lam Ho, who will forever remain in my heart.

May 16, 2023

ABSTRACT

MACHINE LEARNING FOR ULTRAVIOLET SPECTRAL PREDICTION

Linh Ho Manh, Ph.D.

The University of Texas at Arlington, 2023

Supervising Professor: Victoria C.P. Chen

Machine Learning has found wide applications in material science, including dielectric polymers, superconducting materials, and drug property prediction. The use of data analytics and machine learning methods to predict Vacuum Ultraviolet (VUV) spectra by encoding molecular structure is gaining interest because high quality VUV spectral prediction capability would enable the study of new molecules without costly wet-lab measurements. This dissertation aims to study feature representations for molecular structure that enhance the prediction of VUV spectra via machine learning models. Both interpretable machine learning and deep learning are studied.

Chapter 1 provides an overview of VUV/UV spectra retrieval, and Chapter 2 reviews relevant machine learning models and conventional techniques in molecular analysis from the existing literature. Chapter 3 presents the primary contribution of this dissertation, which introduces a new set of features that captures molecular characteristics that are potentially important for accurate VUV spectral prediction. These new features are combined with features derived from the literature and prediction comparisons are studied for a variety of machine learning models. Findings demon-

strate improvements in accuracy, highlight important features, provide comparisons in computational effort for different methods, and identify directions for future work.

Chapter 4 takes a closer look at two of the deep learning methods studied in Chapter 3, namely graph based and molformer methods. Because deep learning embeds feature engineering within the algorithm, the existing form of these methods cannot take advantage of the features studied in Chapter 3. In order to leverage the success of incorporating these features in VUV spectral prediction, a complementary structure is developed with the deep learning architecture. In addition, the graph-based method is improved by introducing a new edge feature that specifically identifies aromatic cycles. Findings show increased prediction accuracy with the complementary structure, which indicates a potentially generalizable benefit for deep learning. Finally, Chapter 5 provides closing remarks on future research.

This dissertation contributes to the application of machine learning in predicting VUV spectra, providing interpretable models, and facilitating molecular analysis in the domain of Cheminformatics.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Introduction to ultraviolet absorption spectra

Gas phase absorption spectroscopy in the vacuum ultraviolet and ultraviolet (VUV/UV) region of the electromagnetic spectrum probes the electronic structure of molecules in the absence of an interacting solvent. Photons in the VUV/UV region promote quantized ground to excited state transitions for valence electrons of a molecule during absorption. The energies (i.e., wavelengths) and probabilities (i.e., magnitude) of molecular absorption are dictated by the chemical structure and atom connectivity of a molecule. Gas phase VUV/UV absorption spectra measured at sufficient resolution are essentially unique and diagnostic for a particular molecule. Gas phase spectra are not subject to deviations in absorption caused by the presence of a solvent and are consequently highly stable and reproducible from one measurement apparatus to another [1, 3, 4, 5, 6].

The relatively recent advancement of coupling gas chromatography with VUV absorption spectroscopic detection (GC-VUV) now offers routine separation of complex mixtures and speciation of separated components, since most the chemical species absorb and have unique gas phase absorption cross sections in the approximately 120−240 nm wavelength range monitored [1, 7]. Hence, VUV/UV has the ability to overcome some limitations of standardized techniques, such as Gas chromatography-mass spectrometry (GC-MS) and liquid chromatography (LC), when compounds with isomeric, isobaric, and multiple isomers are analyzed [8, 7, 9, 10].

Ultraviolet-visible (UV-Vis) spectroscopy is a widely used technique in many areas of science ranging from bacterial culturing, drug identification, and nucleic acid purity checks and quantitation, to quality control in the beverage industry and chemical research [11]. UV-Vis spectroscopy is an analytical technique that measures the number of discrete wavelengths of UV or visible light that are absorbed by or transmitted through a sample in comparison to a reference or blank sample. This property is influenced by the sample composition, potentially providing information on what is in the sample and at what concentration. Since this spectroscopy technique relies on the use of light, the measured wavelength contributes a huge amount of information about the studied material [5, 6]. Light has a certain amount of energy that is inversely proportional to its wavelength. Thus, shorter wavelengths of light carry more energy and longer wavelengths carry less energy. This is why the absorption of light occurs for different wavelengths in different substances. Humans are able to see a spectrum of visible light, from approximately 380 nm, which we see as violet, to 780 nm, which we see as red. UV light has wavelengths shorter than that of visible light to approximately 100 nm. Figure 1.1 depicts the mechanism of how a spectrophotometer operates [11].

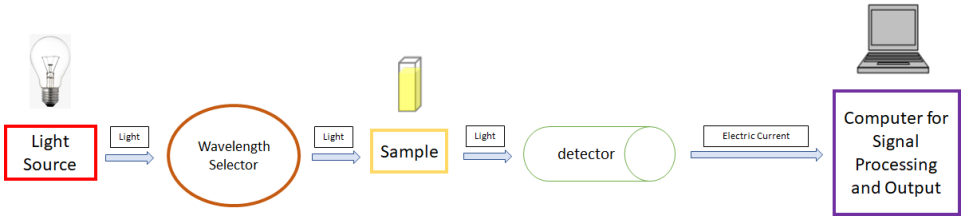

Figure 1.1. A simplified schematic of the main components in a UV-Vis spectrophotometer.

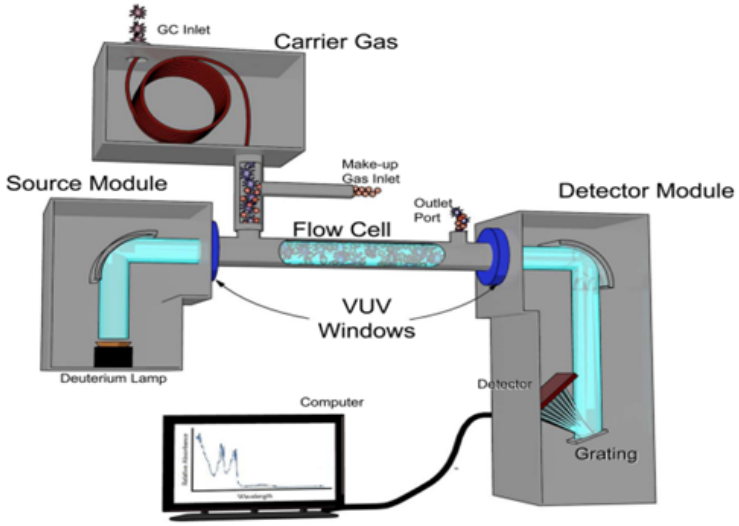

Figure 1.2. Schematic of gas chromatography instrument with a VUV detector (not to scale). Image source[1].

Whichever wavelength selector is used in the spectrophotometer, the light then passes through a sample. The reference sample signal is then later used automatically by the instrument to help obtain the true absorbance values of the analytes. UV-Vis spectroscopy information may be presented as a graph of absorbance, optical density, or transmittance as a function of wavelength. However, the information is more often presented as a graph of absorbance on the vertical y-axis and wavelength on the horizontal x-axis [1, 11]. As can be seen in Figure 1.3, each small molecule holds a distinct VUV spectrum in the studied wavelength from 125 nm to 450 nm.

1.2 Importance of VUV prediction analysis

Forecasting the physicochemical characteristics of compounds plays a crucial role in the exploration of new materials in biotechnology, pharmaceutical studies, energy research, and fuel characterization. It is also significant in drug discovery within pharmaceutical studies. Having a dependable system for predicting VUV/UV

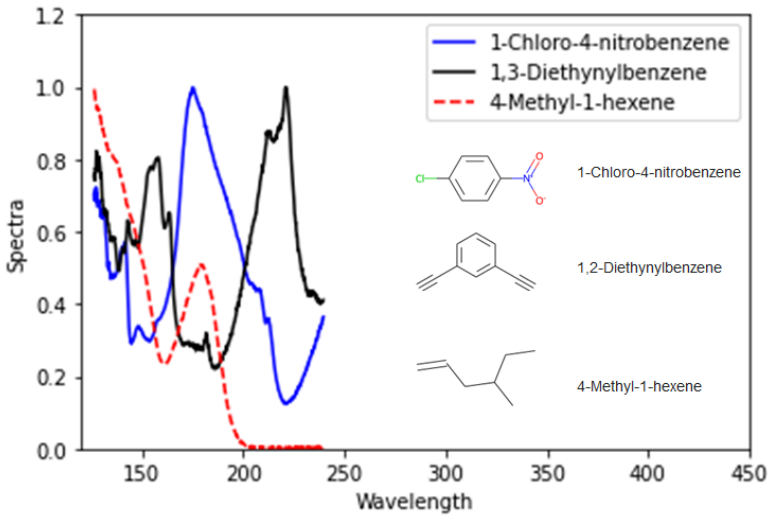

Figure 1.3. Different VUV spectra of different molecules.

spectra would be valuable, as it could expedite the molecular design and analytical measurement. This would enable chemists to identify the unique features of new molecules even before synthesizing them. Since the creation of new compounds and obtaining VUV spectra measurements can be challenging, chemists can benefit from a highly accurate spectral prediction model.

Traditional molecular screening techniques known as Quantitative Structure–Activity Relationships/ Quantitative Structure–Property Relationships (QSAR/QSPR) are based on the correlation between features and functional activities for each compound [12, 13]. However, QSAR/QSPR faces difficulties coping with random and diverse databases with a restricted number of laboratory experiments [14]. Linear solvation energy relationships can provide some useful information on solutes that have detailed descriptors, but the molecules for which these descriptors are known are limited, and the descriptors are not easy to determine for other molecules of interest. On the other hand, properties with straightforward calculations, such as pKa or $\log P$, provide limited useful information when molecules are present in com-

plex systems. QPSK/QSAR also possess limited transferability since their models are typically specific to a particular dataset or chemical domain, which makes it challenging to extrapolate the predictions of a model to new compounds or different chemical classes. Models trained on one dataset may not perform well on a different dataset due to variations in chemical space, molecular diversity, and underlying mechanisms. The interpretability of QPSK/QSAR can be also problematic since QSAR/QSPR models often provide statistical correlations between molecular features and properties but may not offer an understanding of underlying causal relationships. It can also be unclear how to interpret the significance and physical meaning of specific descriptors or their contributions to the predicted property. Since the data for QSPK/QSAR is limited to several domains, QSAR/QSPR models may become outdated as new chemical compounds and structures emerge. Keeping the models up-to-date and relevant requires continual updates and retraining.

Conventional techniques for spectral prediction, such as time-dependent density functional theory (TD-DFT), are often deployed to predict electronic absorption spectra. However, TD-DFT relies on various approximations, such as the adiabatic approximation, linear response approximation, and Tamm-Dancoff approximation. These approximations can introduce errors, particularly for highly excited states or strongly coupled systems. In addition, TD-DFT primarily predicts absorption lines and does not provide detailed information about the shape of the spectrum over a wide bandwidth. To compensate for this limitation, artificial broadening techniques, such as Gaussian functions, are often employed, which may affect the accuracy of the predictions. Another disadvantage of TD-DFT is that the accuracy of TD-DFT calculations is highly sensitive to the choice of basis set and exchange-correlation function used. Different combinations of these parameters can lead to varying results, making it challenging to obtain consistent and reliable predictions.

An alternate solution for predicting measured properties from molecule structure without expensive calculations is machine learning based on already measured data, such as a reference spectra library. Machine learning has shown increasing accuracy in various domains of data science. Prior to the development of deep learning methods, machine learning techniques were already being used for classification and regression tasks in cheminformatics. These methods have been applied to predict various chemical properties, such as energetic properties, logP, atomization energies, and toxicity, often treated as single-output regression problems. In the literature, most of the predicted chemical properties are energetic properties [15], logP [16], atomization energies [17, 18, 19], and toxicity [20, 21]. Prediction of mass spectra, similar to VUV/UV spectra, is considered a multiple-output regression problem [22]. Two prominent approaches for molecular characterization involve feature extraction based on Simplified Molecular-Input Line-Entry System (SMILES) and molecular graphs [15, 16, 18, 23].

The prediction of UV spectra has been limited, due to a lack of high-quality data below 200 nm in wavelength. Previous research on machine learning models for VUV/UV spectral prediction primarily focused on the Long-Short Term Memory (LSTM) model, as discussed by Urbina et al. [24]. Their study utilized an effective wavelength range of 220 nm to 400 nm, with the Extended Connectivity Fingerprint Diameter 6 (ECFP6) as the molecular representation and LSTM as the machine learning model.

This dissertation explores different molecular feature representations and machine learning models using a dataset of 1397 molecules, with an effective wavelength range of 126 nm to 240 nm and a resolution of 0.15 nm. It is important to note that the wavelength scale in our dataset differs from the 1 nm resolution and the range of 220 nm to 400 nm studied by Urbina et al. [24]. The measurements in our study

were conducted in the gas phase, without the presence of a liquid or any interacting medium. This means that the absorption spectra obtained are not affected by the interaction between the molecule of interest and a solvent, as commonly observed in traditional solution-based UV/Vis measurements. The dataset used in our research is derived from measurements performed by the VUV Analytics company and an established commercially available spectral library of VUV/UV. By excluding the influence of solvents, the measured spectra in our dataset exhibit more consistency and are not subject to the variability and potential errors introduced by solvent interactions.

To achieve accurate VUV spectral prediction via machine learning, there is a need to investigate molecular representations and computational techniques to map the relationship between the structural information of molecules and VUV/UV spectra. In Chapter 3, a variety of molecular feature representations and machine learning / deep learning techniques are studied. While Elton's work [15] provides a comprehensive overview of feature extraction for molecules, these lack information on aromaticity and bond properties that are important for predicting VUV spectra. An important contribution in Chapter 3 is the characterization of new feature representations.

In Chapter 4, an examination of deep learning methods studied in Chapter 3 identified two beneficial modifications. The first encodes edge information for molecules with aromatic cycles within graph-based deep learning, and the second develops a complementary structure for deep learning architectures that incorporates features from Chapter 3. These modifications demonstrate how improvements in VUV spectral prediction can be achieved and motivate future feature characterizations and future research utilizing the knowledge gained from VUV spectra prediction.

CHAPTER 2

Literature review

2.1 Conventional spectral prediction techniques

Before the development of machine learning and data science in the field of Cheminformatics, there were two main conventional techniques for spectral prediction, namely time-dependent density functional theory (TD-DFT) and Quantitative Structure-Activity Relationship/ Quantitative Structure-Property relationship. In the first section of this chapter, a brief discussion of these two algorithms and their limitations is presented.

2.1.1 Time-dependent density functional theory (TD-DFT)

TD-DFT is a computational method used in quantum chemistry to study the electronic structure and properties of molecules and materials. It is an extension of the Density Functional Theory (DFT) method, which is a widely used method for calculating the electronic structure of molecules and materials. In TD-DFT, the electronic design of a system is described by the time-dependent density, which is a function of both space and time. The method calculates the time evolution of the density under the influence of an external electric field. This allows for the prediction of properties, such as the absorption spectra of molecules, which are embedded in a wide range of applications, including materials science, chemical synthesis, and pharmaceuticals [25]. TD-DFT has a wide range of applications that study the properties of materials, such as semiconductors and catalysts. The method can be used to predict the optical properties of materials, such as their absorption spectra, which

are important for applications in optoelectronics and photovoltaics [26, 27]. In addition, TD-DFT is also embedded in the biomolecules, such as proteins and DNA, which benefits the prediction of the absorption spectra of these molecules [28]. In pharmaceuticals, this method studies the electronic properties of drug molecules and their interactions with target molecules, such as receptors and enzymes [29, 30].

However, TD-DFT often does not produce the high-resolution spectral structure observed in experimental gas phase VUV spectra [1, 24]. Since TD-DFT primarily predicts absorption lines and does not provide full information about the whole spectrum, artificial broadening techniques, such as Gaussian functions, are often employed. These techniques can have negative effects on the accuracy of TD-DFT method in the prediction of VUV/UV spectra.

2.1.2 Quantitative Structure-Activity Relationships/ Quantitative Structure-Property Relationships (QSAR/QSPR)

Quantitative structure-activity relationship (QSAR) is a computational approach that analyzes the relationship between chemical structures and bioactivity data to predict the biological activity of new chemical compounds. There are various types of molecular descriptors used in QSAR, including constitutional, topological, geometrical, electronic, and quantum-chemical descriptors[31, 32]. These descriptors are used to represent the structural and physicochemical properties of a molecule that are relevant to its biological activity, such as its size, shape, electronegativity, hydrophobicity, and electronic structure. The choice of descriptors and their weighting in the QSAR model affects its accuracy and applicability domain. Despite the advances in QSAR modeling, there are still challenges in the prediction of chemical toxicity and the extrapolation of QSAR models to new chemical structures [33].

To improve the reliability and regulatory acceptance of QSAR models, several guidelines and initiatives have been proposed, such as the Organisation for Economic Co-operation and Development (OECD) principles for QSAR validation and the European Chemicals Agency's (ECHA) QSAR toolbox. These guidelines emphasize the importance of transparency, reproducibility, uncertainty analysis, and external validation in QSAR modeling. QSAR models need to be validated against an independent dataset, and their performance metrics, such as accuracy, precision, and robustness, need to be reported transparently. The use of consensus QSAR models and uncertainty analysis methods can also enhance the reliability and regulatory acceptance of QSAR predictions [34].

Some advantages of QSAR include an ability to rapidly screen a large number of compounds and provide insights into the molecular interactions between a compound and its target, leading to the discovery of new binding sites and targets [35]. However, since the application of QSAR is based on the correlation between features and bioactivities for each compound [12, 13, 36, 37], it faces some difficulties when coping with random and diverse databases with a restricted number of laboratory experiments [14]. Linear solvation energy relationships (LSER) are an example of descriptors used in QSAR. LSER can offer valuable information about solutes with detailed descriptors, but there is limited availability of known molecules with these descriptors, and determining the descriptors for other molecules of interest is not straightforward. On the other hand, properties that can be easily calculated, such as pKa or $\log P$, do not provide significant insights when multiple molecules from different classes exhibit similar values for these properties.

2.2 Machine Learning Methods

Given some of the disadvantages mentioned for TD-DFT and QSAR, there is a need to investigate additional feature representations that can describe molecular structures. Further, these features can be utilized within machine learning models for predictive modeling applications. The scope of this dissertation concentrates on the characterization of molecules into feature representations that can be utilized by machine learning / deep learning models for the prediction of VUV spectra. The standard approach for implementing machine learning in cheminformatics is to represent each molecule as a feature vector in the X domain. Prediction of VUV/UV spectra, similar to mass spectra, is considered a multiple-output regression problem [22]; hence, the spectra output is the Y domain. There are multiple machine learning models that can map the relationship between X: input feature vectors to Y: VUV/UV spectra output [38, 39, 40, 41, 42, 43]. Three machine learning algorithms considered for this molecular representation were a Multi-Layer Perceptron Neural Network (MLP) with two hidden layers of 256 and 128 nodes developed from the Pytorch package, multiple output Random Forest Regressor and Multiple output Gradient Boosting Tree Regressor from sci-kit-learn framework. This section discusses the general overview of these three machine learning methods, which are also implemented in our spectral prediction problem, and the featurization scheme is discussed in the next chapter.

2.2.1 Random Forest Regressor

Random Forest regression is a supervised learning algorithm that uses an ensemble learning mechanism for regression. The ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model [44, 45, 46, 47, 48, 49].

As depicted in Figure 2.1, Random Forest comprises the following steps:

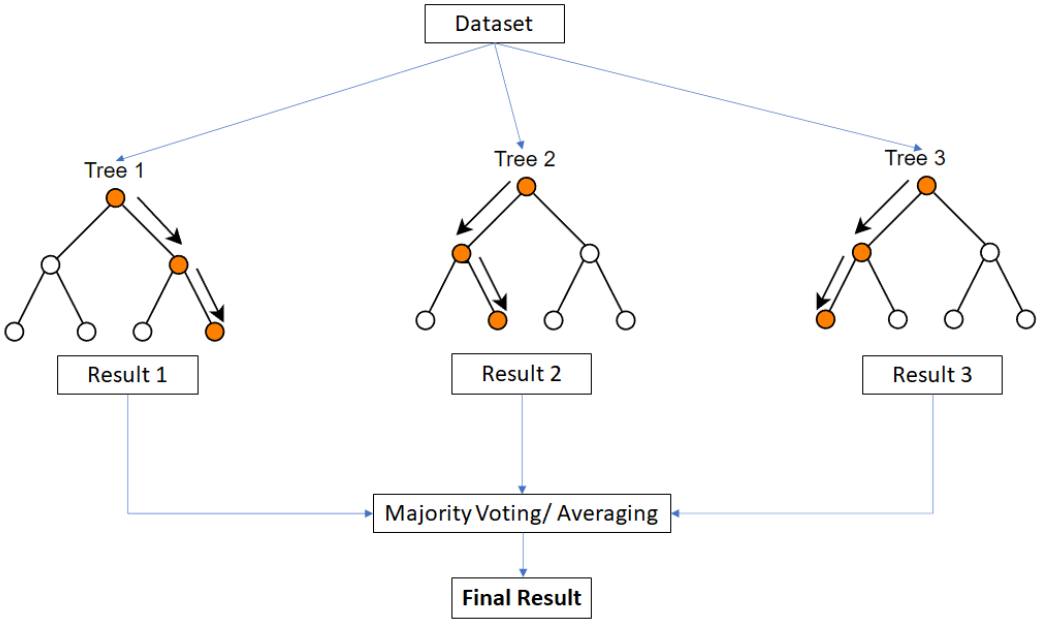

Figure 2.1. Ensemble learning mechanism in Random Forest.

1. Pick at random k data points from the training set.
2. Build a decision tree associated with these k data points.
3. Choose the number N of trees you want to build and repeat steps 1 and 2.
4. For a new data point, make each one of your N-tree trees predict the value of y for the data point in question and assign the new data point to the average across all of the predicted y values [48, 49].

As mentioned above, VUV/UV spectral prediction is a multiple output regression problem, and the details of how modified splitting or stopping rules are modified are discussed in Schmid's work [50].

2.2.2 Gradient Boosting Tree Regression

The Gradient Boosting Tree (GBT) is one of the most powerful techniques for building predictive models for both classification and regression problems. Gradient Boosting is a machine learning algorithm, which works on the ensemble technique

called 'Boosting.' Like other boosting models, Gradient Boosting sequentially combines many weak learners to form a strong learner [51, 52, 53, 54, 55]. Typically, Gradient Boosting uses decision trees as weak learners [56, 51, 57].

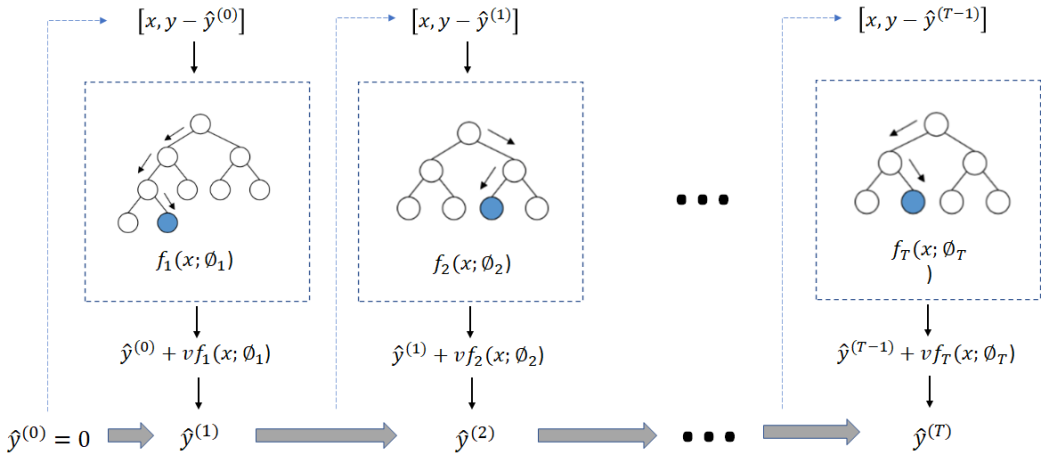

Figure 2.2. Schematic diagram of the gradient boosted regression tree.

As depicted in Figure 2.2, Random Forest comprises of following steps:

1. Construct a base tree with a single root node. It is the initial guess for all the samples.
2. Build a tree from errors of the previous tree.
3. Scale the tree by learning rate (value between 0 and 1). This learning rate determines the contribution of the tree in the prediction
4. Combine the new tree with all the previous trees to predict the result and repeat step 2 until a maximum number of trees is achieved or until the new trees do not improve the fit.
5. The final prediction model is the combination of all the trees.

In the case of multiple output prediction as in the VUV/UV spectra problem, GBT constructs multiple trees corresponding to the output variables [43]. Specifically, in Zhang's work [43], the objective for learning multiple outputs is based on the second-order Taylor expansion of loss. This objective function is approximated and then connected with the objective for a single output. We also formulate the problem of learning a subset of variables and derive its objective. This is achieved by adding an L_0-norm constraint, a regularization technique that imposes a constraint on the number of leaves (terminal nodes) allowed in each tree. By setting an L_0-norm constraint, we limit the complexity of individual trees by restricting the number of possible splits and preventing overfitting.

2.2.3 Multi-layer Perceptron Neural Network

The Multilayer Perceptron (MLP) is commonly used in simple regression problems [58, 59]. However, MLPs are not ideal for processing patterns with sequential and multidimensional data [60]. An MLP Neural Network has input and output layers, and one or more hidden layers with many neurons stacked together[61, 62, 63]. MLP falls under the category of feedforward algorithms because inputs are combined with the initial weights in a weighted sum and subjected to the activation function, such as ReLu or Sigmoid. The internal computation resulting within each layer is input to the next layer, starting from the input layer representing the datam, through the hidden layers, and, finally, to the output layer [60].

Backpropagation is the learning mechanism that allows the Multilayer Perceptron to iteratively adjust the weights in the network, with the goal of minimizing the lack-of-fit to the data. This is equivalent to estimating the model's parameters (i.e., MLP weights) to fit the model to the data. The lack-of-fit is represented by a standard regression loss function that calculates the squared differences between

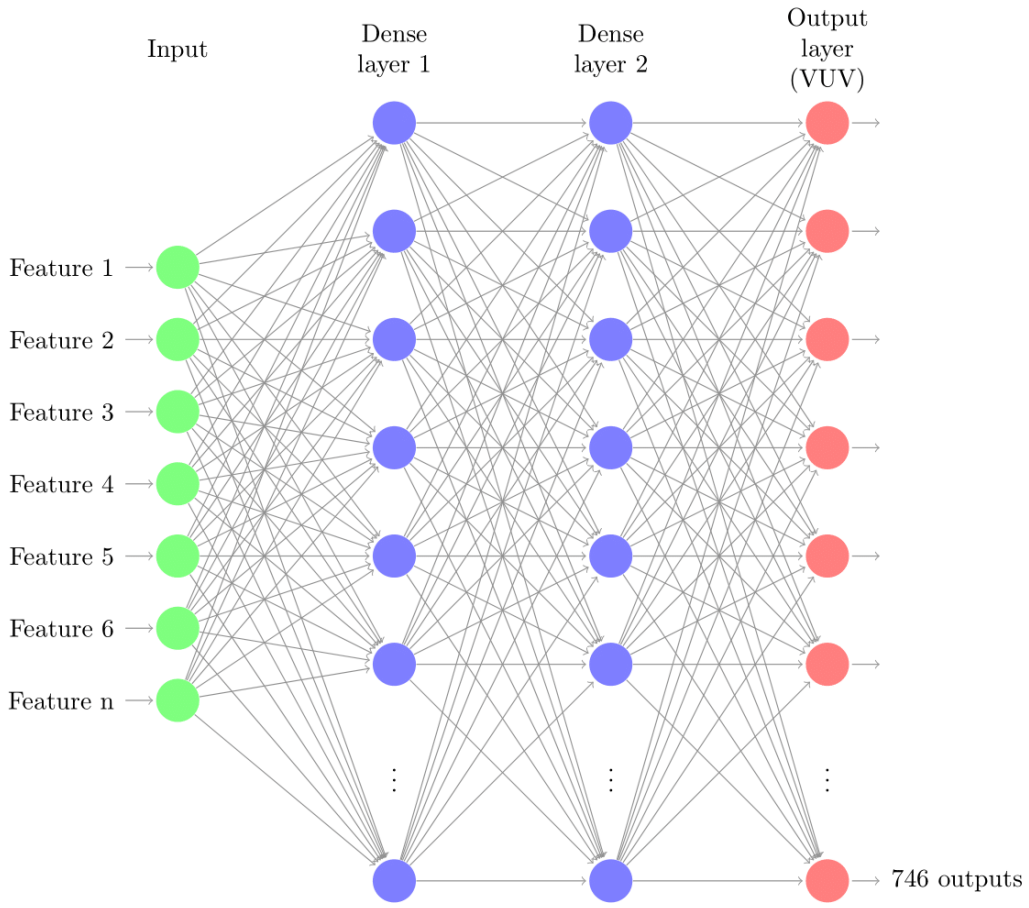

Figure 2.3. Diagram of MLP network with two hidden layers for VUV/UV spectral prediction problem.

Y targets and the MLP predictions, and the associated weights are updated in each iteration to reduce this lack-of-fit. It is clearly stated in Goodfellow's book [64] that there is one hard requirement for backpropagation to work properly. The function that combines inputs and weights in a neuron, for instance, the weighted sum, and the threshold function, for example, ReLU, must be differentiable. These functions must have a bounded derivative because gradient descent is typically the optimization algorithm used in MLP. Figure 2.3 illustrates the implemented neural network for VUV/UV prediction in this research by connecting the feature domain with VUV spectra by an MLP neural network with two hidden layers.

2.3 Deep Learning Architectures

This section explains the overview of several deep learning approaches that are feasible for VUV spectral prediction. The common point of these techniques is that the final feature vector from raw input is extracted by operators and execution units within the algorithm, and they do not extract features as measurable properties.

2.3.1 Computer vision and Convolutional Neural Network

One of the most remarkable differences between convolutional neural networks (CNNs) and traditional MLPs is that CNNs are primarily used in the field of pattern recognition within images [65, 66, 67].

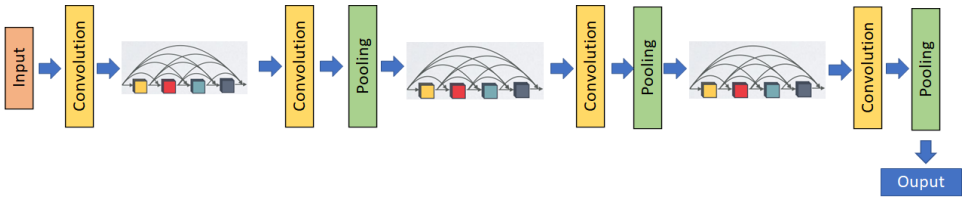

Figure 2.4. The use of CNN in image classification.

Based on the data processing technique presented in [20] and the CNN architecture [68, 69], the functionality of CNN-based architecture can be listed as four key elements as follows:

- **input layer**: Hold the pixel values of three-dimensional images of molecules.
- **Convolutional layer**: Determine the coefficients of neurons which are connected to local regions of the inputs.
- **Polling layer**: Downsample along the spatial dimensionality of the given input.
- **Fully-connected layer**: Produce the final prediction from activation functions and outputs from previous layers.

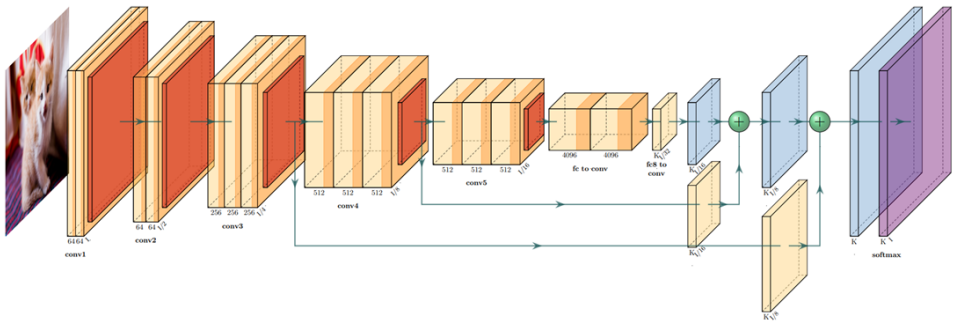

Figure 2.5. Architecture of DenseNet.

Figure 2.7 depicts the architecture of DenseNet, a deep learning framework based on a CNN, which is also implemented in our VUV/UV spectral prediction research. DenseNet was specially developed to improve accuracy by vanishing the gradient in high-level neural networks, due to the long distance between input and output layers, and the information vanishing before reaching its destination [70].

It is worth noting that the convolutional operation in machine learning can be applied to many various data types that are multi-dimensional, making them similar to multi-color images. For example, the transformer-based method, which is discussed in the next section, implements a convolutional unit to study the interaction between atoms in their operation.

2.3.2 Graph Neural Networks

Molecular graphs serve as a valuable two-dimensional representation of chemical molecules, capturing their topological and structural characteristics, as well as atom connectivity. A molecule can be seen as a graph of $G = \{X, A, E\}$ where X represents the node matrix, indicating the atom types within the molecule. The adjacency matrix A and edge matrix E provide information about the connections between atoms. Graph Neural Networks (GNNs) have the ability to learn atom order

permutation invariant representations, encode the graph matrix representation into a latent space, and efficiently train on a graphics processing unit (GPU) and scale to large datasets. Some of these points are not unique to GNNs. However, the graph representation can naturally be expanded in applications where one would need more information than simply the identity and connectivity of atoms in a molecule.

Recently, a novel approach called HM-GNNs[71] has been developed, which integrates heterogeneous motifs into the architecture of GNNs. Motifs are small subgraphs that frequently occur in molecular structures and represent important features such as functional groups and cycles. By incorporating these motifs into the neural network architecture, HM-GNNs can capture more detailed structural features and enhance the expressiveness of the learned representations. Another approach for pretraining molecular graph representations with 3D geometry involves the use of graph convolutional neural networks (GCNNs) [72]. GCNNs are a type of neural network designed specifically for processing graph-structured data, such as molecular structures. They operate by passing messages between neighboring atoms in a molecule, using learned weights to combine information from different atoms and edges in the graph[73].

In a GNN, there are three main tasks, that can be listed as follows:

- **Graph-level task**: In a graph-level task, the aim is to predict the property of an entire graph.
- **Node-level tasks**: Focus on nodes, which include node classification, node regression, and node clustering. This task is not explored in our research.
- **Edge-level tasks**: They are edge classification and link prediction, which require the model to predict whether or not there is a connection between two atoms. If there is, then identify the property of that vertex. A notable exam-

ple of this problem is a Kaggle Data Challenge [74] with the goal to predict magnetic interactions between a pair of atoms.

Chemoinformatics is a broad field that encompasses computer science and chemistry with the goal of utilizing computer information technology to solve problems in the field of chemistry, such as chemical information retrieval and extraction, compound database searching, and molecular graph mining [75, 76, 77, 78, 79]. Regarding the problem of VUV/UV prediction, the Graph-level task is a valid option for our research since molecules can be described as graphs, and VUV spectra output can be specified as graph properties. Details on the GNNs implemented in our research are provided in Appendix B and Chapter 3.

2.3.3 Transformer Techniques

Until recent years, architectures that utilized recurrent neural networks (RNNs) and deep learning methods (e.g., long short-term memory), were widely used for tasks such as text translation to text classification. However, in 2017, the transformer architecture was developed and demonstrated to outperform these methods [2]. The rise of the transformer architecture for various natural language processing (NLP) tasks was the catalyst for more transformer techniques [80, 81].

An encoder model would typically pass the embedding inputs via the following sublayers [82, 83, 84, 85]:

1. **Positional Embeddings**: Incorporate positional information for each token embedding.
2. **Pre-layer normalization**: Normalize each input in the batch to have zero mean and unity variance.
3. **Multi-headed attention layer**: Focus the model simultaneously on multiple sets of linear projections from the data.

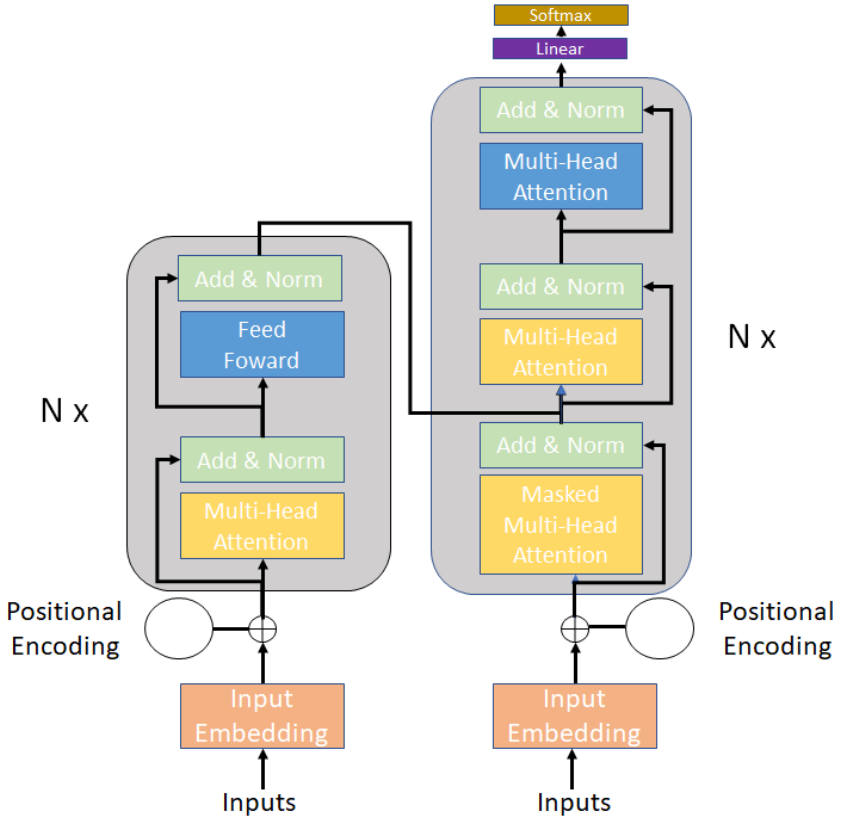

Figure 2.6. Architecture of full transformer model used in [2].

4. **Feed-forward layer**: Process each embedding input independently.

One of the most common ways to implement a self-attention layer would be to use the scaled dot-product attention, which is discussed in more detail in Appendix C. It is worth mentioning that, our VUV/UV spectral prediction problem falls in the category of supervised learning; hence, only the encoder architecture in the transformer-full model is investigated. Similar problems can be found in [86].

2.4 Visualization Toolbox by FLASK

The popularity of the FLASK toolbox has grown quickly in recent years [87, 88]. In chemistry, this toolbox is a convenient interface for predicting new molecules using

learned patterns from machine learning models trained on library data. However, the "black-box" effects of machine learning models potentially generate misleading results. The main purpose of the FLASK application is for visualization, providing

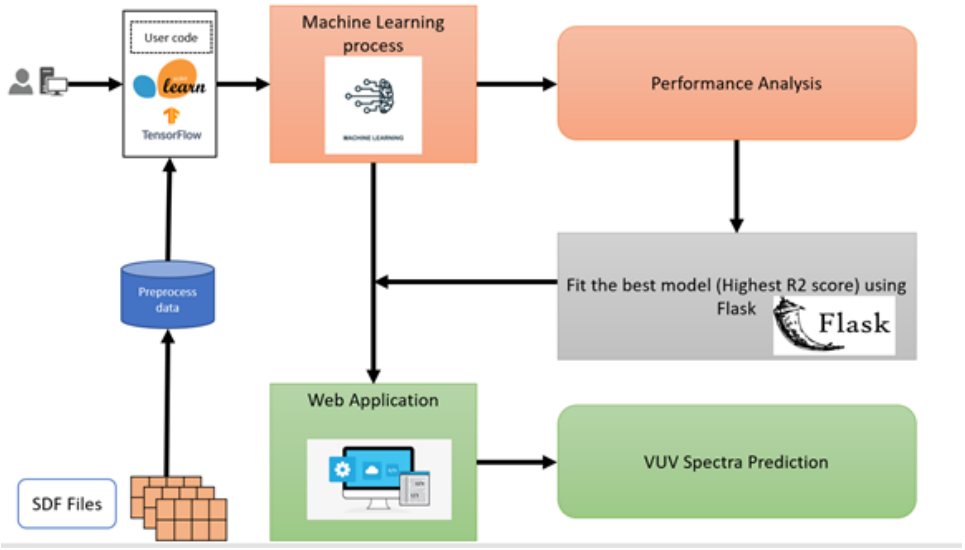

Figure 2.7. Architecture of spectral prediction toolbox developed by FLASK.

information on a new molecule that is not in the library dataset. Since the whole purpose of machine learning models is to build the relationship between molecular structures and VUV spectra, a new prediction is retrieved as the mean prediction from five models in a cross-validation procedure. After receiving the SDF file, the system checks whether the uploaded molecule is already in the database. If not, the prediction is provided. Since the pre-trained machine learning models from the training data run at the backend of the FLASK application, the users are able to get the final prediction by just submitting the SDF file of the new molecule. After receiving SDF file as input, FLASK will implement data processing, load pre-trained models, feed the input to the models, and perform prediction. It is worth noting that FLASK is a Python framework that can handle HTML templates, CSS, and

Javascript, making it convenient for machine learning developers, to deploy a user-friendly application.

CHAPTER 3

Machine Learning Molecular Feature Representation for Vacuum Ultraviolet and Ultraviolet Gas Phase Absorption Spectral Prediction

3.1 Introduction

Gas phase absorption spectroscopy in the vacuum ultraviolet and ultraviolet (VUV/UV) region of the electromagnetic spectrum probes the electronic structure of molecules in the absence of an interacting solvent. Photons in the VUV/UV promote quantized ground to excited state transitions for valence electrons of a molecule during absorption. The energies (i.e., wavelengths) and probabilities (i.e., magnitude) of molecular absorption are dictated by the chemical structure and atom connectivity of a molecule. Gas phase VUV/UV absorption spectra measured at sufficient resolution are essentially unique and diagnostic for a particular molecule. Gas phase spectra are not subject to deviations in absorption caused by the presence of a solvent, and as such, they are highly stable and reproducible from one measurement apparatus to another [1].

The relatively recent advancement of coupling gas chromatography with VUV absorption spectroscopic detection (GC-VUV) now offers routine separation of complex mixtures and speciation of separated components since most of the chemical species absorb and have unique gas phase absorption cross sections in the approximately 120 – 240 nm wavelength range monitored [1, 7]. Hence, VUV/UV has the ability to overcome some limitations of standardized techniques, such as gas chromatography-mass spectrometry (GC-MS), when compounds with isomeric, isobaric, and multiple isomers are analyzed [7, 8, 9]. The prediction of spectroscopic

absorption spectra, such as illicit drugs[89], fuels components [90, 91, 92], and synthetic products [93], have been studied in recent years. A reliable means for VUV/UV spectral prediction is valuable and could accelerate molecular design and analytical measurements since chemists can use it to characterize and identify new molecules, especially when pure standards are lacking.

Recent research demonstrates machine learning applied to different domains of data science, including pharmaceutical studies, with increasing accuracy over time. It is worth noting that, before the development of deep learning methods, there were a number of classification and regression tasks in cheminformatics that had been tackled by machine learning methods [16, 15, 18, 23, 19, 94]. In the literature, some of the predicted chemical properties have been logP [16], energetic properties [15], atomization energies [18, 19, 17], and toxicity [20, 21]. These examples all involve the prediction of a single-output value. Prediction of VUV/UV spectra, similar to mass spectra, is considered a prediction of multiple-output values [22]. Traditionally, predicting VUV/UV spectra has been a challenging task due to the complexity of spectral data and the high computational cost of underlying quantum mechanical calculations.

Molecular characterization plays a critical role in machine learning prediction of chemical properties of molecules. Machine learning algorithms rely on input features, such as molecular descriptors, to learn patterns and make predictions about new molecules. Traditional molecular characterization techniques have included the use of various techniques, such as Abraham Descriptors and Linear Solvation Energy Relationships (LSER). Efforts (often through extensive experimental measurements) are made to reduce the complexity of a chemical compound/structure to a set of values that describe the nature of physicochemical properties and interactions exhibited by the molecule [12, 13]. Since the accuracy of prediction models is highly depen-

dent on the input data of molecular descriptors and experimental values, traditional molecular screening faces difficulties when feature extraction requires laboratory experiments [14]. In addition, linear solvation energy relationships can provide some useful information on solutes having detailed descriptors, but the molecules for which these descriptors are known are limited, and the descriptors are not easy to determine for other molecules of interest. On the other hand, properties with straightforward calculations, such as pKa or $\log P$, provide limited information when the molecules are present in complex systems. Another solution for predicting measured properties from molecule structure without carrying out expensive wet-lab experiments is machine learning, which also delivers more information in diverse pools of molecules by intuitive feature engineering. Thus, there is a need to investigate molecular representations and computational techniques to map the relationship between the structural information of molecules and VUV/UV spectra. With the advent of machine learning algorithms, it is now possible to accurately predict VUV/UV spectra using computational models. Two major categories of molecular characterization are feature extraction based on Simplified Molecular-Input Line-Entry System (SMILES)[16, 15] and molecular graphs [18, 23], both of which are studied in this chapter.

Reported techniques for predicting VUV/UV spectra of molecules involved the use of quantum mechanical methods, specifically time-dependent density functional theory (TD-DFT). TD-DFT relies on several approximations, including the adiabatic approximation, the linear response approximation, and the Tamm-Dancoff approximation, which can lead to errors in the calculation of electronic properties, especially for highly excited states or strongly coupled systems [95]. In addition, TD-DFT predicts only absorption lines and does not provide the shape over a wide bandwidth. To address this, predicted absorption lines have been artificially broadened using Gaussian functions to make them appear closer in appearance to experimental spec-

tra. However, the accuracy of TD-DFT calculations can be highly sensitive to the choice of basis set and exchange-correlation functions used [96]. As TD-DFT does not fully account for electron correlation effects, it shows limitations when being applied to complex datasets where highly excited states or strongly correlated systems are presented [97]. On the other hand, the prediction of ultraviolet spectra has been restricted due to a shortage of quality data when the wavelength is below 200 nm.

Recently, Urbina et al. reported the use of machine learning models for the prediction of UV spectra [24] at a wavelength window from 220 nm – 400 nm. They implemented the well-known Extended Connectivity Fingerprint Diameter 6 (ECFP6) as the molecular representation and Long-Short Term Memory (LSTM) network as a machine learning model. In their dataset, the presence of solvents in experiments can affect and shift the spectra in their studies. The framework of Urbina et al. can serve as a spectral prediction tool for a new molecule in the wavelength range 225 – 400nm, as long as a valid SMILES is provided [98]. Our research work concentrates on the wavelength ranging from 126 nm to 240 nm, as such the prediction can be expanded to shorter wavelengths. In addition, our extended research on featurization schemes can provide a better representation to explain the relative position of a new molecule among the existing ones in the library data, and this function is not yet included in the web-based framework [98]. By extensively exploring the feature space of molecules with the aid of chemistry insights, predictions and outputs of machine learning models are more interpretable and explainable.

3.1.1 Contribution

The aim of this work is to investigate different molecular feature representations and machine learning models using a dataset of 1397 molecules with the effective wavelength 126 nm – 240 nm and a resolution of 0.15 nm. The wavelength scale

in our dataset is different from the 1 nm resolution and the wavelength range 220 nm – 400 nm studied by Urbina et al. [24]. The VUV measurements was partially carried out by the company VUV Analytics [99] and an established spectral library of VUV/UV gas phase absorption spectra, which is available commercially. In our study, all measurements were conducted in the gas phase, as opposed to the liquid phase in traditional solution-based UV/Vis measurements that involves potential interaction with the solvent. Consequently, the measured spectra in our dataset are more regular and do not include variability and potential sources of error due to solvents interacting with measuring substances. With this more reliable dataset, our goal in this chapter is to provide an investigation of machine learning based options to improve VUV spectral prediction. In the next section, we first examine options for featurizing molecular structure, starting with the overview by Elton et al. [15], and then provide background on appropriate machine learning and deep learning methods. In particular, new features are introduced that identify the aromaticity, existence of atoms in the halogen group, and bond properties of molecules to better represent molecular structures that impact VUV spectra.

3.2 Molecular representations and machine learning methods

There are two main components to building predictive models. First, the explanatory factors that may potentially yield predictive ability must be identified, and second, the predictive model is constructed. In machine learning, the factors are called features, and the prediction is generated by the machine learning model. For characterizing features, we discuss how information from a molecular structure data file (.sdf) is translated into feature vectors, representative images, and molecular graphs by manipulating SMILES and their relative positions in 3D coordinates. For machine learning models, we discuss both standard machine learning models that

take the features as a input and deep learning models whose algorithms internally conduct featurization.

3.2.1 Molecule represented as a feature vector

The feature representations in our study explicitly avoid descriptors that require physical measurements. Rather, our focus in on features based on domain knowledge and computational methods. We start the discussion with features in the problem domain of energetic materials [15] and then introduce new features specifically intended to improve VUV spectral prediction.

3.2.1.1 Existing featurization scheme

There are three categories of features discussed by Elton et al. [15]: fingerprints, custom descriptor set, and counts over chemical bonds.

- **Fingerprint**: Fingerprinting algorithms represent molecular structure as a set of numerical values or binary bits. These numerical or binary values, known as molecular descriptors, encode the chemical and physical properties of the molecule, such as size, shape, polarity, and electronic structure [100]. In Elton et al. [15], different schemes of fingerprints, such as Atom-Pair [101], Topological Torsion [102, 103], Morgan's fingerprint [104], and E-state [105] were examined. Molecular representation in Urbina et al. [24] also falls in this category since they implemented the Extended Connectivity Fingerprint Diameter 6 (ECFP6), which is similar to Morgan's fingerprint. In Elton et al., it was also demonstrated that E-state has the best performance among the pool of fingerprints. Hence E-state fingerprint truncated to 32 atom types is the only fingerprint fragment studied in our feature set.

- **Sum over bonds (SOB)**: Based on the intuition that almost all of the latent heat energy is stored in chemical bonds, the bond counts feature vector was introduced [105]. The bond counts vectors find all bond types that occur in the overall set of molecules considered and then count how many of each bond are present in each molecule. Each entry in this vector has a magnitude that corresponds to the number of one specific bond type that exists in the molecule ensuring a unique representation. In the end, all the bond features are concatenated to form the final vector.
- **Custom Descriptor (CDS)**: A set of custom descriptors [15] was chosen based on physical intuition and computational efficiency, mapping a molecule to a scalar value and ignoring the descriptors that require physics computation and measurement. It is worth noting that the functional groups in Elton et al. [15] contain a lot of nitrogen.

The reaction of oxygen with the fuel atoms carbon and hydrogen is represented in a descriptor called oxygen balances (**OB**) [15], and this feature type is also included in our study.

3.2.1.2 New features as complementary information

Our proposed new features are generated in similarly to the count used by the **custom descriptor** and **sum over bonds** techniques. Our features are based on the theory that the aromaticity of a molecule contributes a significant amount of information to the VUV spectral prediction problem [106]. Aromatic compounds often exhibit strong UV absorption due to their conjugated pi-electron system. The pi electrons are distributed evenly over a cycle, making it less reactive than other types of compounds. This stability affects the chemical reactivity and overall properties of the molecule [107]. Our new features also account for the presence of olefin atoms and

conjugated non-aromatic double bonds, since the existence of the double bond also leads to a planar, sp2 hybridized geometry around the carbon atoms, which influences the molecule's reactivity and stability [108]. In addition, contiguous rotatable bond groups, also known as rotatable bonds or torsional angles, refer to a set of two or more adjacent single bonds that can rotate relative to each other. The presence of these rotatable bond groups can significantly influence the molecular properties of a compound, such as its conformation, stability, and reactivity [109]. Finally, halogen group atoms are highly electronegative, meaning they have a strong ability to attract electrons toward themselves. The presence of a halogen atom in a molecule can increase the overall electronegativity of the molecule [110]. consequently, our new "ABOCH" features are proposed as follows:

1. Number of aromatic (A) and olefin (O) atoms.
2. Number of conjugated double bonds, aromatic bonds, and contiguous rotatable bond groups (B).
3. Number of saturated cycles and aromatic cycles (C).
4. Number of each type of different halogen atom (H)

In order to count aromatic and olefin atoms, our method is based on data manipulation using RDKit [111], an open-source software toolkit for cheminformatics. RDKit transforms molecules into hierarchical structures, then accesses all atoms and bonds in each molecule to extract their properties. The count of aromatic cycles and benzene cycles is implemented in a similar fashion since we can access all the cycles within molecules and extract the information. Regarding the count of contiguous rotatable bonds, our proposed method finds all groups of contiguous rotatable bonds and sorts them by decreasing size. Since different molecules can have different contiguous rotatable bond groups, the final feature is the count of all possible contiguous rotat-

able bond groups. Regarding the count of halogen atoms, we propose three separate features corresponding to the number of fluorine, chlorine, and bromine atoms.

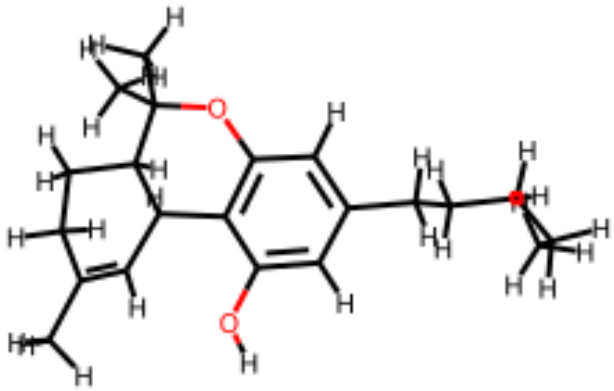

Figure 3.1. Example molecule delta9-trans-tetrahydrocannabinol.

To demonstrate the counts for our ABOCH features, we use the delta9-trans-tetrahydrocannabinol molecule as an example. This molecule contains one cycle that may be mistaken as an aromatic cycle. Our featurization identifies three saturated cycles and one aromatic cycle. There are six aromatic atoms and two olefin atoms. However, there are no conjugated double bonds. Additionally, the molecule has six aromatic bonds. Figure 3.1 shows the structure of the molecule. Notably, there are no halogen atoms, resulting in zero values for the features indicating the number of chlorine, bromine, and fluorine atoms.

3.2.1.3 Machine learning models

There are multiple readily available machine learning models that can map the relationship between X: input feature vectors to Y: VUV spectra output [38, 39, 40, 41, 42, 43]. Three machine learning algorithms are considered for this molecular

representation. From the Pytorch package [112], we implement a Multi-Layer Perceptron Neural Network (MLP), with two hidden layers of 256 and 128 nodes. From the scikit-learn framework [113], we employ the multiple-output Random Forest Regressor (RF) and the multiple-output Gradient Boosted Tree Regressor (GBT). PyTorch and scikit-learn are widely used open-source machine learning frameworks that provide users with the necessary tools and libraries for constructing machine learning models. These frameworks offer a range of functionalities, including data preprocessing, model construction, training, and evaluation, facilitating the development and implementation of machine learning algorithms to diverse application domains.

3.2.2 Deep learning approaches

In this section, three deep-learning approaches are discussed for VUV spectral prediction. Unlike traditional machine learning methods discussed earlier, deep learning embeds featurization within a complex algorithm that, in theory, can automatically learn nonlinear relationships between raw input data and the target variable, automatically identifying transformations and combinations of the inputs that improve prediction. Because the deep learning process conducts a comprehensive search to uncover relationships, domain experts are not needed to define input features, such as those features described in Section 3.2.1. This is particularly useful when dealing with high-dimensional data, such as images or text, or molecules as graph-based structures, where it may be difficult to manually design relevant features. However, deep learning feature extraction is computationally expensive and may require large amounts of training data to achieve good performance. In addition, it may be difficult to interpret the resulting models due to the lack of transparency, which can be particularly problematic in cheminformatics. Unlike traditional machine learning algorithms, where features are explicitly defined and interpretable, deep learning

models automatically learn hierarchical representations through multiple layers in the model. Furthermore, deep learning models often consist of millions of parameters, making it difficult to pinpoint the exact contribution of each model component (or "neuron") to the final prediction. This lack of interpretability raises concerns in cheminformatics, where understanding the underlying molecular features is crucial for decision-making, understanding structure-activity relationships, or identifying important substructures. In our VUV/UV spectral prediction problem, deep learning feature extraction involves using a deep neural network to learn relevant features directly from the raw input data. The neural network consists of multiple layers of interconnected nodes, each of which performs a nonlinear transformation of the input data. Below, three specific deep learning models are described, each of which has been adapted to represent aspects of molecular structure.

3.2.2.1 Molecules as multi-dimensional images and convolutional neural networks

In recent years, the use of deep learning for image recognition has been increasing significantly. In the chemistry domain, these advances have led to the development of Chemception, which is trained to predict chemical properties by encoding molecules as multi-dimensional images. For this encoding, first, SMILES strings of molecules are translated to their respective 2D coordinates and then encoded into multi-channel images, with each layer containing molecular information [20, 21]. Each layer is then used to encode different information from the molecule. In our implementation, layer zero is encoded with the bond order, and the next two layers are encoded with the atomic number and hybridization (not Gastieger charge as in [21]). It should be emphasized that augmented images of molecules help the deep learning models generalize chemical structures better. Augmentation techniques involve applying transformations, such as rotation, scaling, or flipping, to existing images to create new

variations of the same image [114]. By introducing these variations into the training data, machine learning models are forced to learn features that are invariant to such transformations. This can help to reduce overfitting and improve the generalization performance of the resulting prediction model. After molecules are transformed into multi-dimensional images, DenseNet [115], a deep learning structure that consists of multiple convolutional operators and pooling layers, is implemented to recognize the molecular patterns and predict VUV spectra.

3.2.2.2 Graph representation and graph neural network

A molecular graph is a two-dimensional representation of a chemical molecule and accounts for its topo-structural features and atom connectivity. A molecule can be seen as a graph of $G = \{X, A, E\}$ where $X = \{x_i \in \mathbb{R}^F \mid i = 1, 2, ..., N\}$, and N is the maximum number of atoms for one molecule in the dataset. The binary adjacency matrix A has the dimension $N \times N$, in which 1 denotes there is a connection between two atoms, and 0 denotes when there is no connection. It is worth noting that the dimension of the node matrix is $N \times F$, where F is the number of node features. For cheminformatics studies, the node feature is typically the atom type, in which the atom can be represented by the corresponding number in the periodic table. The edge feature matrix has the dimension $N \times N \times S$, where S denotes the number of edge features. Similarly to the adjacency matrix, edge features provide the characteristic of a connection between two atoms in the molecule: single bond, double bond, or triple bond. The graph neural network framework in this research uses the *Spektral* toolbox [18], consisting of message passing and graph pooling functions. Message passing is similar to the role of the convolutional operator in the convolutional neural network where X, the node representation of one molecule,

is updated iteratively. Details of graph neural network operations can be found in different studies citeGrattarola2021[116, 117, 118].

3.2.2.3 Molecular graph representation and transformer

In the last few years, the transformer framework, with its unique mechanism of "self-attention," has been successfully used in the field of natural language processing as an automated translator between languages [2] and in the domain of computer vision [86]. A similarity between natural language processing and molecular structures is that they can both be represented as sequences of data. In natural language processing, the text is often represented as a sequence of words, where the order of the words matters and affects the meaning of the text. Similarly, in molecular structures, the sequence of atoms and bonds within a molecule determines its properties and behavior. Recently, Wu et al. [23] proposed the molformer framework, with a distinguished self-attention mechanism that can estimate interactions between multi-level nodes. In their framework, the multi-scale mechanism to capture local patterns with increasing contextual scales provides a perspective into how distances can influence the score between center atoms and the rest in one molecule, similar to the use of deep learning to approximate quantum chemical simulations [19]. The molformer framework takes atomic numbers and their 3D corresponding positions as the input for the training process to capture how different chemical compositions, structures, and conformations affect the behavior of VUV/UV spectra.

3.3 Comparisons of machine learning approaches for VUV spectral prediction

Since VUV spectra reflect the intensity of absorption of VUV light at different wavelengths, the output can be seen as a vector with a length that depends on the resolution of the absorption measurement. Experimental spectra from the VUV

spectrometer can be from the range of 123 nm – 450 nm; however, the wavelength window selected is 126 nm – 240 nm since relatively few entries in the spectral library contain data in the range from 240 – 450 nm. It is worth noting that the resolution of wavelength measurement is 0.15 nm which results in each molecule having 746 VUV spectra outputs in the selected wavelength window. Our dataset comprises 1397 distinct molecules. To compare different machine learning algorithms, the dataset of 1397 molecules was divided into five groups, and 5-fold cross-validation with a fixed randomization seed was employed to calculate prediction accuracy. Cross-validation is a well-known approach for addressing possible biased selections of training and testing sets [119]. After training, out-of-sample molecules, which are not included in the training process, are used to investigate the goodness of the proposed method.

There are multiple methods to evaluate the efficiency of machine learning models such as mean absolute error (MAE), root mean square error, and coefficient of determination. In this research, we employ the coefficient of determination, as defined in the following equation, where $0 \leq R^2 \leq 1$ and a higher R^2 indicates better prediction accuracy:

$$R^2 = 1 - \frac{\sum_i \left(y_i{}^{true} - y_i{}^{predict}\right)^2}{\sum_i \left(y_i{}^{true} - \bar{y}_i^{true}\right)^2} \tag{3.1}$$

As mentioned in the previous section, this problem is a multiple-output regression problem; hence, each sample/instance in the cross-validation operation has a corresponding R^2 score. In order to compare the different methods, the mean of R^2 score was obtained through 5-fold cross-validation.

Comparison of molecular feature representations and traditional machine learning methods

Our new set of ABOCH features consists of ten integer values: number of cycles, number of aromatic cycles, number of aromatic atoms, number of olefin atoms,

number of contiguous rotatable bond groups, number of conjugated double bonds, number of aromatic bonds and separate features accounting for the counts of different halogen atoms. In Table 4.1, performance for various combinations of the feature sets from Section 3.2.1 are presented. The following abbreviations are used to represent the feature sets:

- Estate = E-state fingerprint (31 features).
- CDS = Custom Descriptor (19 features).
- SOB = Sum over bonds (45 features).
- OB = oxygen balances (1 feature).

Combining all the features together yields the best performance across all the machine learning methods. Among the existing feature sets from Section 3.2.1, Estate+CDS+SOB performs best, notably better than Estate+CDS+SOB+OB. This appears to be a non-intuitive result since both sets contain Estate+CDS+SOB. However, in general, adding unimportant features hampers predictive modeling, so this result indicates that the OB feature set is not beneficial for VUV spectral prediction. Regarding the machine learning models, Random Forest Regressor provided the highest R^2 values. For the subsequent comparisons in this section, we employ the overall best model using the combination of all features with a Random Forest Regressor.

The Random Forest Regressor model additionally provides an importance ranking of the input features, which is given in Figure 3.2. Features from our ABOCH feature set are marked by asterisks ($*$). Seven of the ten ABOCH features appear in the figure, with the number of aromatic atoms emerging as the most important. The other ABOCH features listed include the number of conjugated double bonds, the number of olefins, the number of aromatic cycles and bonds, and the number of chlorine atoms. Notably, two features derived from the E-state fingerprint, namely *aCa* and *aaCa*, indicate the presence of a carbon atom with two aromatic bonds and

Table 3.1. Averages of 5-fold cross-validated R^2 scores for various combinations of molecular feature sets across three machine learning methods.

	Multi-layer NN	RF	GBT
SOB + OB	0.485	0.635	0.621
CDS + SOB	0.543	0.637	0.614
Estate + SOB	0.607	0.663	0.643
Estate + CDS + SOB	0.617	0.665	0.650
Estate + CDS + SOB + OB	0.52	0.663	0.654
ABOCH features alone	0.524	0.575	0.556
ABOCH combined features	**0.630**	**0.691**	**0.665**

three aromatic bonds in its vicinity [105]. This observation further strengthens our hypothesis that molecular aromaticity plays a prominent role in the absorption of VUV/UV spectra within the wavelength window ranging from 126 nm to 240 nm.

To illustrate the effectiveness of learned patterns by machine learning algorithms, four predictions of four "new" molecules namely hexahydrothymol, 4-methylethcathinone, 2,4-dimethylphenol, and 2,2,3,3,5,6,6-heptachlorobiphenyl are shown in Figure 4.5. We consider them "new" molecules because they are not in the training database that was used to generate our machine learning models. They have measured spectra in the library, and they were chosen to represent a reasonably diverse group of compounds across the molecules used to formulate the training set. Hexahydrothymol is a terpenoid compound lacking aromaticity, 4-methyl methcathinone is an illicit stimulant drug designed to mimic amphetamine, 2,4-dimethylphenol is a simple substituted aromatic, and 2,2,3,3,5,6,6-heptachlorobiphenyl represents a polychlorinated biphenyl (PCB) environmental contaminant. Figure 4.5 shows that high quality VUV spectral predictions were achieved for all four of these "new" molecules.

To illustrate the performance of machine learning vs. TD-DFT, VUV spectral predictions for three molecules are presented in Figure 3.4. The three molecues 1,2-

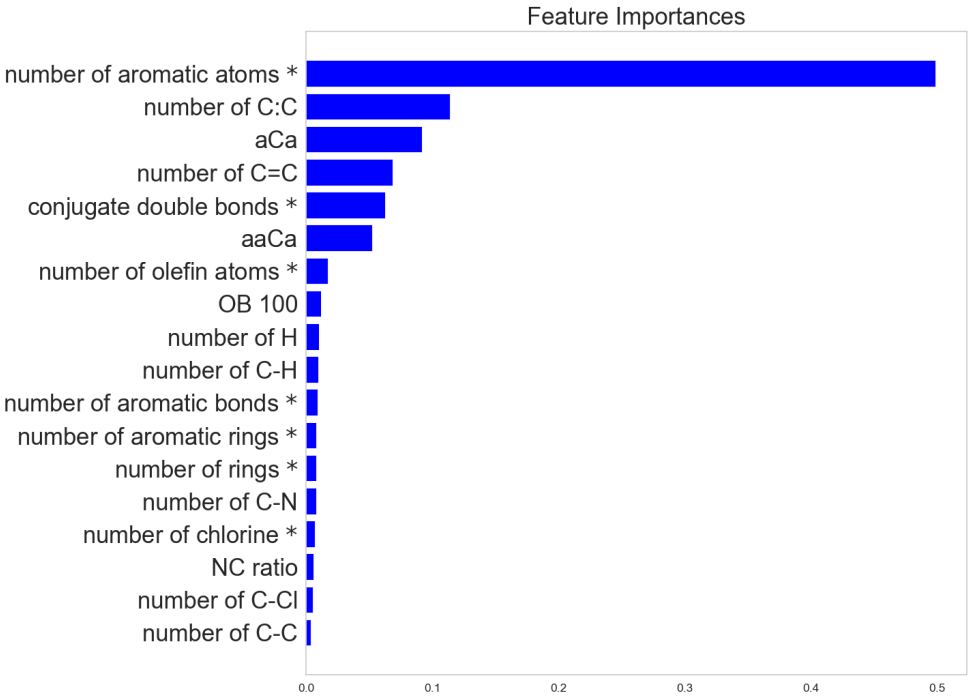

Figure 3.2. Feature importance from Random Forest Regressor model.

dimethylnaphthalene, a-PVP, and naphthalene were chosen because their spectral predictions based on TD-DFT were previously presented in the literature [91]. However, it should be noted that these molecules are members of the training database, as opposed to being "new" molecules (referring to previous paragraph). Consequently, it is not surprising to see strong performance by the machine learning approach. Regardless, TD-DFT fails to capture some critical peaks and valleys in the VUV pattern, with some severely overestimated or underestimated. This is in part due to the TD-DFT process that artificially broadens calculations over the wavelength range by using overlapping Gaussian functions.

Finally, to provide perspective on poor performance by the machine learning approach, Figure 3.5 provides the VUV spectral predictions with the lowest R^2 score. These molecules were contained in the training database, but their molecular structure

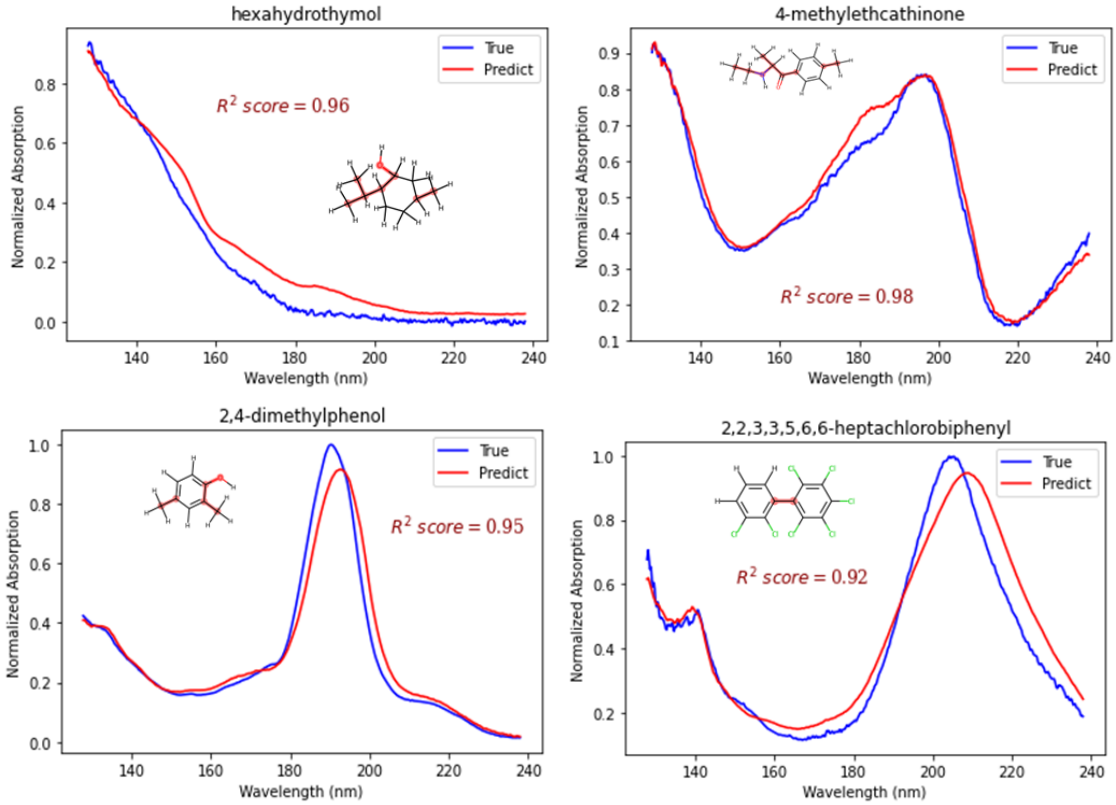

Figure 3.3. Prediction of new molecules not in the training database using the combination of all features with the Random Forest Regressor: (a) hexahydrothymol (b) 4-methylethcathinone HCl (c) 2,4-dimethylphenol (d) 2,2,3,3,5,6,6-heptachlorobiphenyl..

includes characteristics that were not well represented in the database, namely, more than three stacked aromatic cycles (chrysene, benzo[b]chrysene, 3-nitrophenanthrene) or the inclusion of sulfur within the cycle (dibenzothiphene sulfone). With the knowledge gained from our study, future research can introduce additional features that better characterize the member and bond properties within aromatic cycles that are currently absent from our studied feature sets.

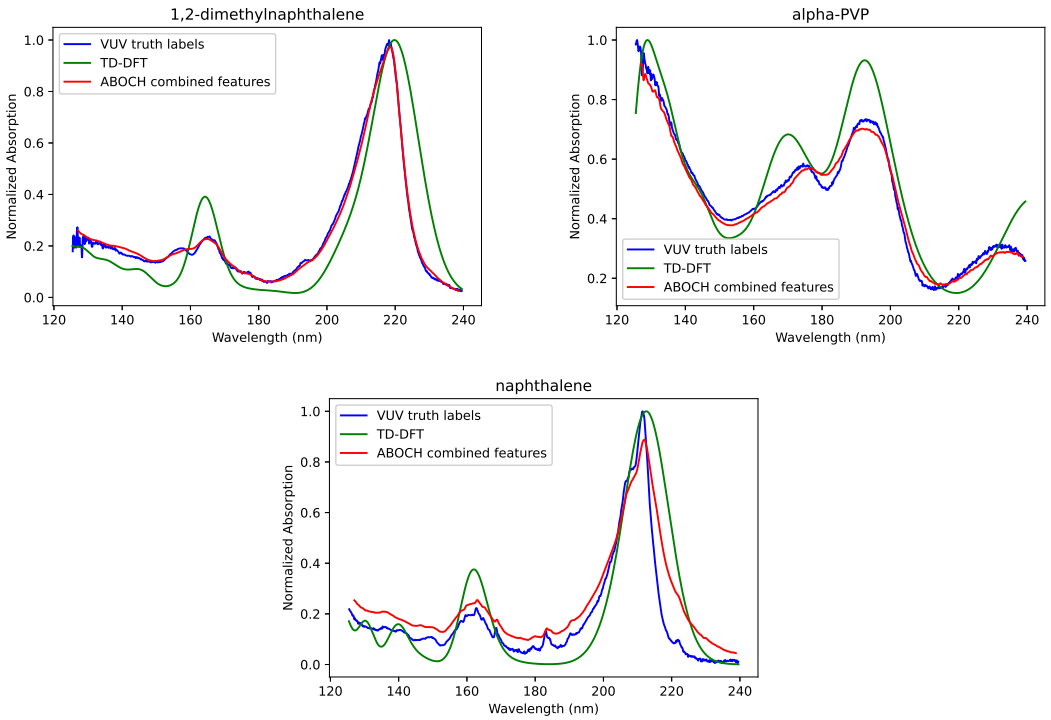

Figure 3.4. Prediction using TD-DFT vs. the combination of all features with the Random Forest Regressor method: (a) 1,2-dimethylnaphthalene (b) a-PVP (c) naphthalene..

Comparison of deep learning methods

Two variants of the graph neural network (GNN) were implemented, namely the Graph Attention (GAT) and edge-conditioned convolutional (ECC) neural networks. Both GAT and ECC use the features of the nodes at the endpoints of an edge, but ECC additional uses edge attributes as inputs to its convolutional operation. In our implementation, we employ the configuration of GAT and ECC as in the *spektral* toolbox [18] with two layers of *GATConv* and *ECCConv* with 64 output channels for each layer, with a batch size of 32. Because GAT does not use properties of edges as inputs, bond information is absent from its modeling. As can be seen from the feature importance of the Random Forest model in Figure 3.2, bond properties, including aromatic and double bonds, are important for VUV spectral prediction.

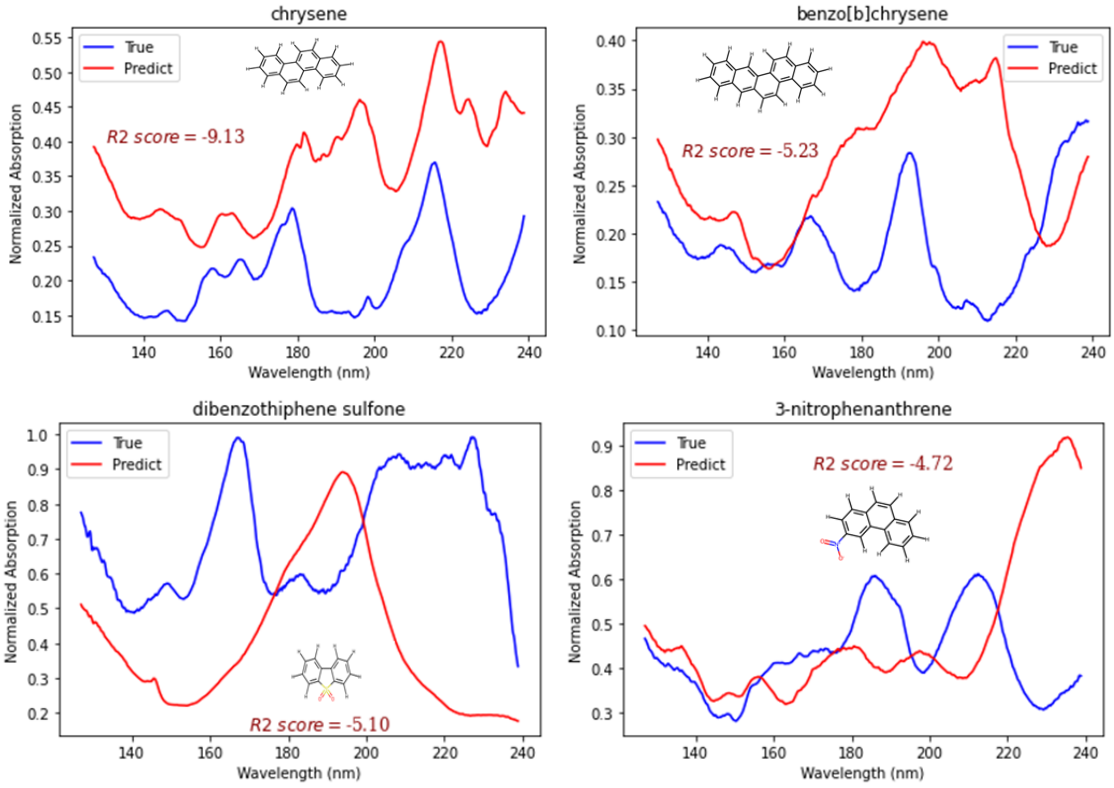

Figure 3.5. Lowest R^2 score predictions using the combination of all features with the Random Forest Regressor method for molecules in the training database: (a) chrysene (b) benzo[b]chrysene (c) dibenzothiphene sulfone (d) 3-nitrophenanthrene..

Table 3.2 compares the performance of GAT, ECC, and the Chemception model with DenseNet [115]. It is seen that Chemception performs best, followed by ECC. The absence of bond information in GAT is likely the primary reason for its extremely poor performance.

Table 3.2. Averages of 5-fold cross-validated R^2 scores for GNN methods and Chemception.

	Graph attention GNN	Edge Convolutional GNN	Chemception
Average R^2 score	-0.510	0.418	0.519

In the Molformer framework proposed by Wu et al. [23], a set of five variants with distinct modeling concepts is presented. The initial variant, referred to as a 3D-Transformer with Sinusoidal Position Encoding (SPE), leverages sinusoidal encoding to effectively utilize the 3D molecular geometry for molecular representation learning. The second approach, a 3D-Transformer with Convolutional Position Encoding (CPE), employs convolutional operations on the pairwise distance matrix of molecules. The third variant, termed 3D-Transformer with Multi-scale Self-attention (MSA), incorporates a distance-based constraint within the self-attention mechanism to extract multi-scaled patterns from the global 3D coordinates of atoms in a given molecule. The fourth variant, 3D-Transformer with Attentive Farthest Point Sampling (AFPS), employs an algorithm to group the most significant atoms around a designated starting atom within a molecule, forming the final representation of the molecular graph. Finally, the complete model encompasses all the features and characteristics derived from CPE, MSA, and AFPS.

A comparison of the five molformer variants is given in Table 3.3. It can be observed that the SPE model exhibits the lowest R^2 score. This result indicates that utilizing sinusoidal positional encoding to embed the 3D relative positions of atoms into their representation is not effective for VUV spectral prediction. While the self-attention mechanism employed in the SPE model demonstrates the ability to capture global data patterns, it falls short in capturing information regarding local context, specifically the interactions between atoms. Similar performance is observed among the CPE, MSA, and AFPS models, highlighting the effectiveness of applying convolutional operations to analyze the pairwise distance matrix of the molecules. Notably, the full model, encompassing all the characteristics of CPE, MSA, and AFPS, outperforms other variants, achieving an average R^2 score of 0.603. As can be seen from Table 4.1, the performance of the full model is comparable to

that of conventional features when applied to the traditional MLP neural network model.

Table 3.3. Averages of 5-fold cross-validated R^2 scores for five variants of Molfomer methods.

	SPE	CPE	MSA	AFPS	Full
Average R^2 score	0.370	0.481	0.478	0.494	0.603

As a final look at the results, Table 3.4 compares the average performance and computational effort of the deep learning methods against the best performing traditional machine learning method from Table 4.1, namely the combination of all features from Section 3.2.1 with the Random Forest Regressor method. The traditional method is seem to be superior to deep learning, from both the prediction performance and computational effort perspectives. All the simulations were performed on the server consisting of the following components: Intel(R) Core(TM) i7-2960X CPU, 65536MB RAM. Ultimately, it does not appear worthwhile to implement a deep learning method that requires many hours of training, is not interpretable, and does not achieve improved prediction performance.

Table 3.4. Averages of 5-fold cross-validated R^2 scores comparing deep learning methods against the combination of all features with the Random Forest Regressor method. Computational times for model training are also shown.

Methods	Average R^2 score	Training time
ABOCH combined features + RF	**0.691**	**25 min**
Molformer Full model	0.603	32 hours
Chemception + DenseNet	0.519	50 hours
Edge Conditional Convolutional Graph	0.418	22 hours

Figure 3.6 provides the complete distribution of R^2 scores for the entire training database in order to illustrate the variability in performance for the methods compared in Table 3.4. In order to separate the improvement in performance achieved with the addition of our new ABOCH feature set, Figure 3.6 includes the performance of the Random Forest Regressor using only the existing conventional feature sets (Estate, CDS, SOB, OB). All distributions are left-skewed with the majority of molecules yielding R^2 scores towards the maximum value of 1.0, and a long lower whisker with outlier cases of extremely poor (negative R^2) performance. The worst outlier cases are seen for ECC, followed by Chemception, while the outlier cases are similar for the Full Molformer and the two traditional machine learning methods. In the zoomed view of the boxes, including our ABOCH feature set with Random Forest Regressor yields the smallest box and the shortest whiskers, which corresponds to the smallest variability in performance.

3.4 Concluding remarks and future research

In this paper, we have presented and compared molecular feature representations and machine learning methods for the application of VUV spectral prediction. Our results demonstrate the benefit of our proposed ABOCH feature set and identified the combination of all features with the Random Forest Regressor method to be the best performing approach, from the perspectives of prediction accuracy, interpretability, and computational effort. Overall, for VUV spectral prediction, it is recommended to utilize feature sets that are based on molecular structure and chemical intuition. Hence, future research should conduct further examination of interpretable features, possibly conducting feature selection to eliminate redundant descriptors. Further, given the benefit of these feature sets, developing a complemen-

tary deep learning structure to incorporate interpretable features could improve the prediction performance of deep learning approaches.

Given the discussion for Figure 3.5 and through the examination of outliers in Figure 3.6, it becomes apparent that there is room for expanding our ABOCH feature set to enhance the representation of molecules. Specifically, this can be achieved by investigating additional characteristics, such as the number of members present in saturated cycles. Furthermore, it is beneficial to separately account for N- and O-containing heterocycles with 5 and 6-membered cycles, as well as incorporating separate features to capture the count of heterocyclic aromatic atoms and carbon aromatic atoms. Moreover, it is worth exploring the properties of aromatic cycles in more depth, including features that account for the number of non-aromatic bonds between aromatic atoms. In addition, an additional set of features was considered, specifically the count of halogen atoms connected directly to aromatic carbons and the count of halogen atoms connected to non-aromatic carbons. The inclusion of these features is justified by the significant influence of halogens' electronegativity on the pi-electron cloud of aromatic cycles, as they tend to attract electron density towards themselves. These expansions to our feature set will contribute to a more comprehensive and accurate representation of molecular structures.

On the other hand, in the ECC implementation, edge features are retrieved from the SDF file, with a double bond and aromatic bond considered the same while in chemistry, these two bonds clearly have different functionality. This may lead to the poor performance of ECC following the implementation of *spektral* toolbox [18]. A research avenue and possible improvement could be reprocessing the molecular structures and distinguishing the aromatic bonds from regular double bonds, which exist outside of aromatic cycles.

3.5 FLASK for visualization with chemical insights

The main purpose of the FLASK application is for visualization, providing information on new molecules that are not in the library dataset. Since the purpose of machine learning models is to build the relationship between molecular structures and VUV spectra, a new prediction is retrieved as the mean prediction from five models from the cross-validation procedure. If the uploaded molecule is not in the database, the prediction is provided. For instance, the new molecule **hexahydrothymol** is not in the database but has measured data in the library. Figure 3.7 shows the prediction versus the truth labels of the spectra. In normal practice, for an arbitrary molecule, only the VUV prediction is provided. The application also has the ability to provide visualization when users want an image of prediction in a specific wavelength range. An example shown in Figure 3.8.

To demonstrate the example, Figure 3.9 shows the prediction of VUV spectra for a new molecule hexahydrothymol, the R^2 score for this validation molecule is quite high at 0.99, showing the ability to capture VUV pattern by machine learning models, in this case, is the best model we have discussed in the previous section: Random Forest Regressor with ABOCH combined features scheme.

In order to understand the position of the new molecule in the chemical domain together with library data, 2D Principal Component Analysis is proposed as in Figure 3.9. By featurizing all data, which includes both the existing library and the new molecule, we can identify the two nearest neighbors by sorting Euclidean distances between feature vectors of molecules. The two most similar to hexahydrothymol in the database are tetrahydrocannabivarin and (+)-3-methoxymorphinan HCl, with their corresponding VUV spectra shown in the Flask app.

The last function of the FLASK application is to show the information by processing the SDF file, as in Figure 3.10. The left-side of the figure is the 3D layout

of atoms in the new molecule following the mechanism in xyz2graph [120] and the right-side of the figure is a 2D layout. In addition, the most important and reasonable features of the new molecule are also given in the web application.

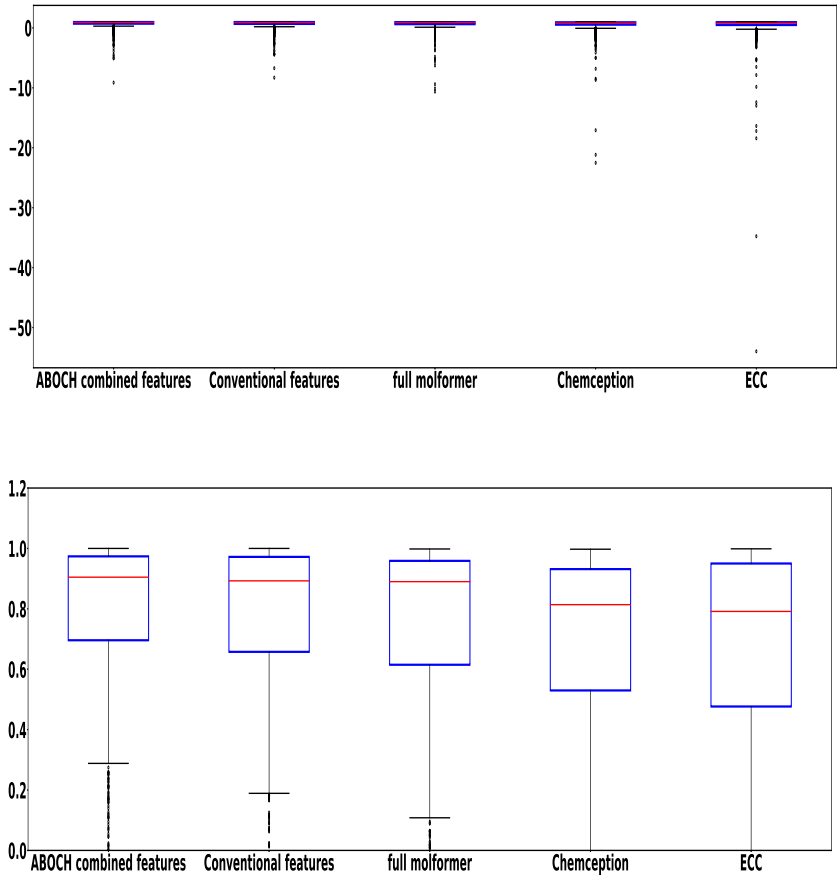

Figure 3.6. Box-and-whisker plots illustrating the distribution of 5-fold cross-validated R^2 scores across the 1397 molecules in the training database, for comparing deep learning methods against the combination of all features with the Random Forest Regressor method. Top figure shows the full box-and-whisker plots, and bottom figure excludes lower outliers and extended whiskers to allow a better view of the box representing the interquartile range (middle 50%) of the distribution..

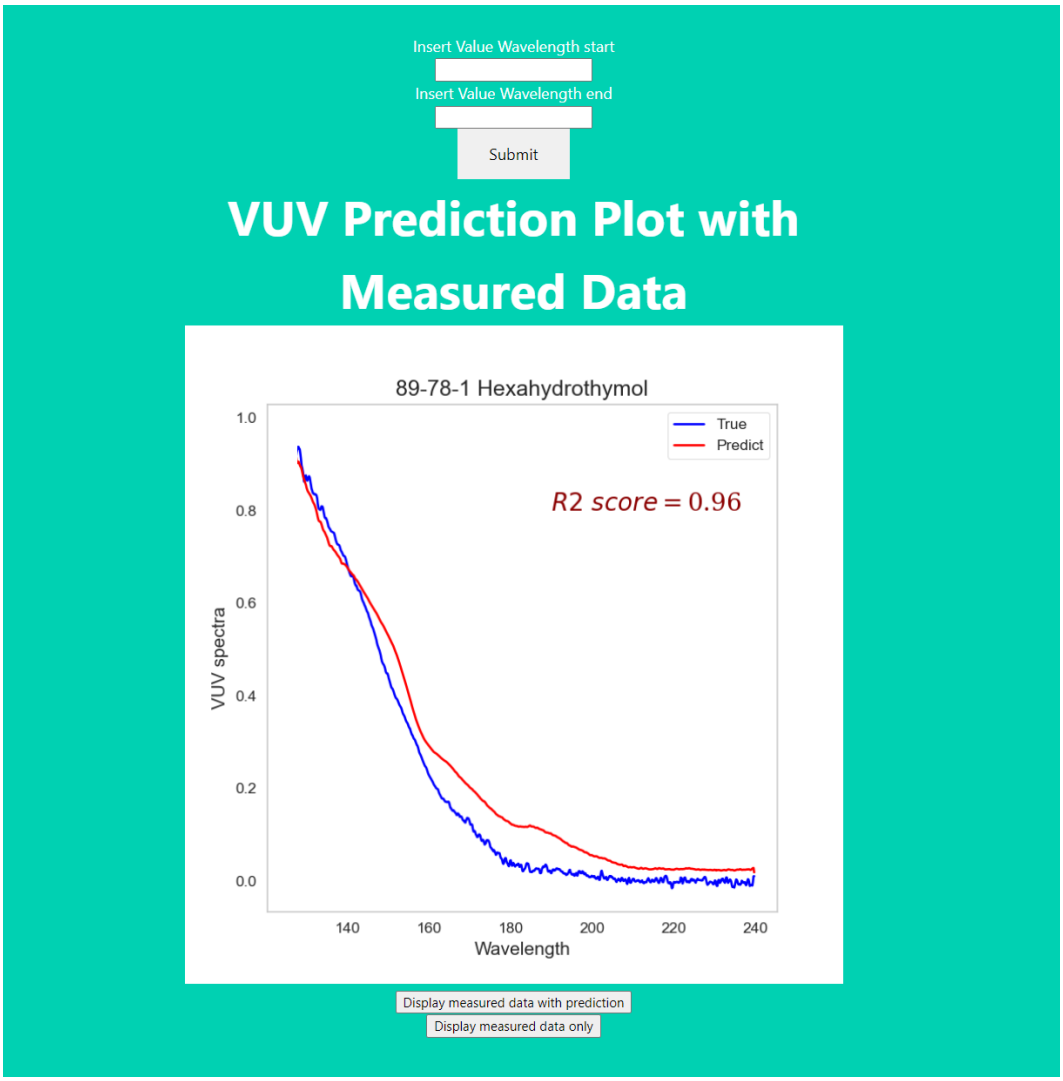

Figure 3.7. The plot of the VUV prediction using neural network.

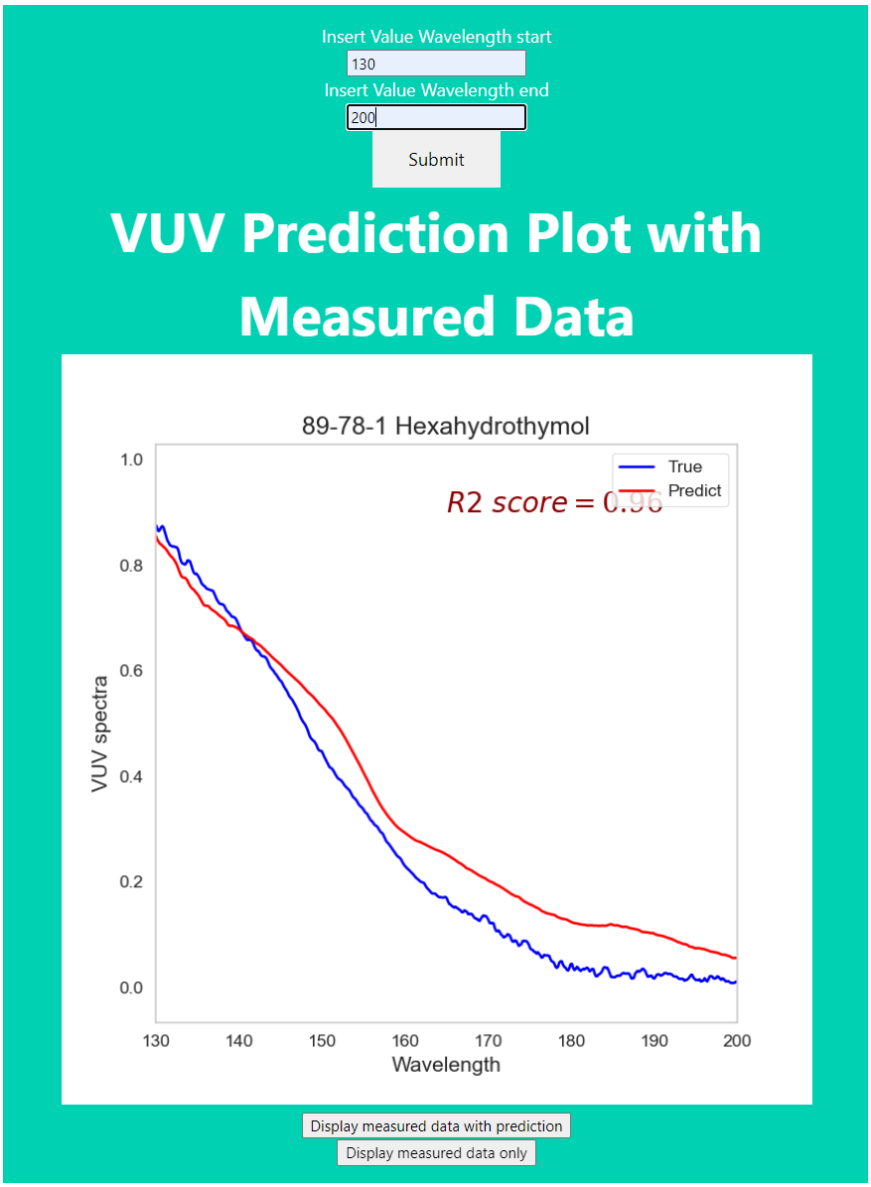

Figure 3.8. Zoom-in mode for specific wavelength range by VUV prediction toolbox by FLASK.

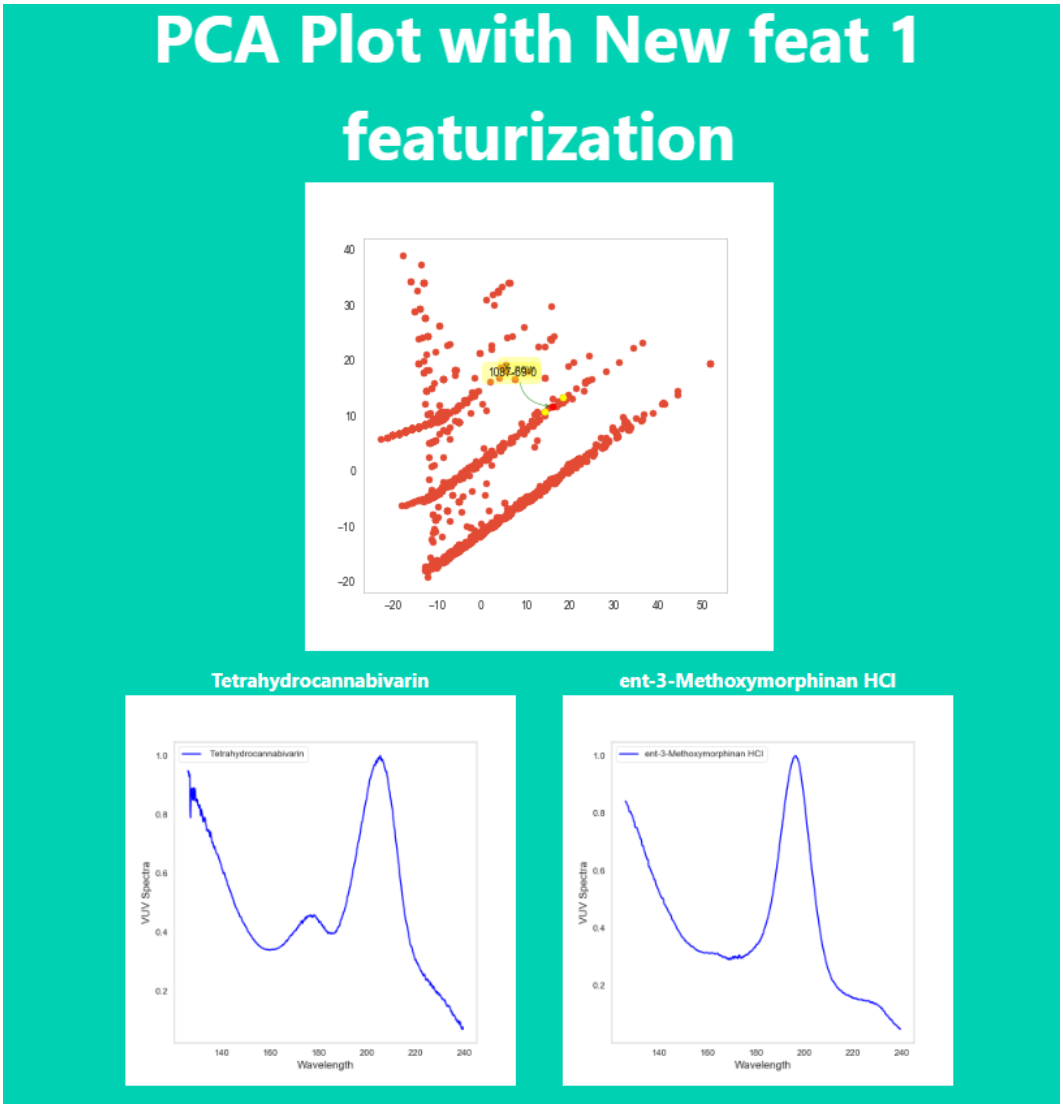

Figure 3.9. PCA plot of new sample: hexahydrothymol with library chemical domain.

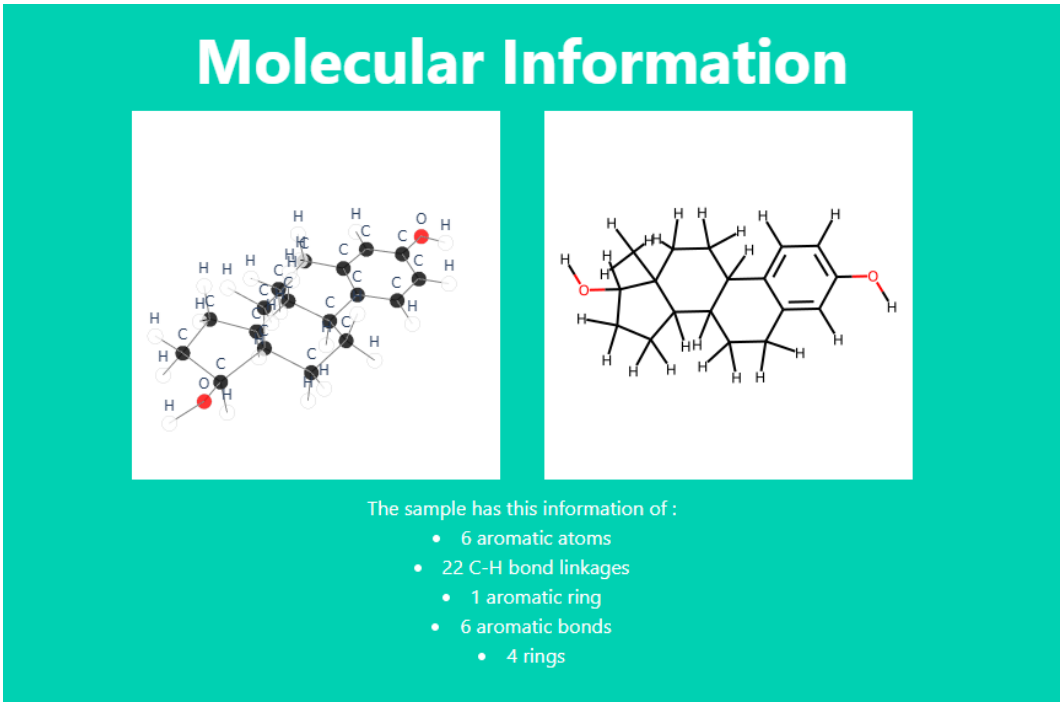

Figure 3.10. Molecular information of new sample: hexahydrothymol.

CHAPTER 4

Complementary Deep Learning Architecture for Vacuum Ultraviolet Spectral Prediction

4.1 Abstract

Vacuum ultraviolet spectroscopy (VUV) plays a critical role in elucidating the electronic structure and optical properties of molecules, thereby finding widespread applications in materials science, chemistry, and physics. Having a reliable and trustworthy system in place to forecast VUV/UV spectra holds significant value, given its potential to accelerate molecular design and analytical measurement processes. It would allow chemists to identify distinctive characteristics of new molecules even prior to their synthesis. Given the difficulties in creating new compounds and obtaining VUV spectra measurements, chemistry researchers would greatly benefit from a highly precise spectral prediction model. In recent years, the application of graph neural networks and transformer-based techniques in Cheminformatics has garnered significant attention as a means to address this challenge. This research presents a novel approach that modifies the graph representation of molecules and employs enhanced deep-learning structures, integrating feature engineering for improved molecular analysis. Our proposed complementary deep learning architectures outperform existing deep learning approaches by a minimum of 25% in terms of average R^2 score.

4.2 Introduction

Gas phase absorption spectroscopy in the vacuum ultraviolet and ultraviolet (VUV/UV) region of the electromagnetic spectrum is a powerful tool for studying

the electronic structure of molecules without the influence of a solvent. In VUV/UV absorption, photons induce discrete transitions from the ground state to the excited states of the valence electrons within a molecule. The absorption energies (i.e., wavelengths) and probabilities are determined by the molecule's chemical structure and atom connectivity. A notable advantage of VUV absorption spectroscopy lies in its ability to directly observe molecular electronic transitions, which exhibit sharp and well-resolved features. This characteristic facilitates the identification of specific electronic states and their corresponding energies. Moreover, VUV absorption spectroscopy demonstrates high sensitivity, enabling the detection of even minute quantities of molecules in the gas phase. This sensitivity arises from the unique gas phase absorption cross-sections exhibited by various chemical species within the monitored wavelength range of approximately 120 to 240 nm [1, 3, 4, 5, 6].

Accurate prediction of molecular properties from molecular structures plays a crucial role in diverse fields like fuel research, forensics, and pharmaceutical drug discovery, where efficient identification of promising drug candidates can significantly reduce time and costs. As illustrated in Chapter 3, machine learning models offer a promising approach for achieving accurate predictions of new molecules by deploying machine learning models with learned patterns from library data. Hence, the use of machine learning can overcome limitations associated with traditional techniques, such as time-dependent density functional theory [95] and quantitative structure-activity relationship [31, 32]. Graph-based neural networks (GNNs), which represent molecules as graphs, have gained considerable attention and popularity due to their ability to handle complex molecular structures without the need for extensive feature engineering. GNNs have demonstrated remarkable success in various domains, including natural language processing [121, 122], pattern recognition [123, 124], and

these methods are especially promising in Cheminformatics and material science, as in the prediction of physicochemical properties [71, 125, 18].

The idea behind the molecular graph representation lies in mapping the atoms and bonds that make up a molecule into sets of nodes and edges. Intuitively, one could imagine treating the atoms in a molecule as nodes and the bonds as edges [126] [127]. In Chapter 3, it was noted that features representing bond information for molecules are important. In particular, the GNN variant that does not incorporate bond information performed poorly, while the edge-conditioned convolutional (ECC) neural network [18] performed reasonably because it encodes bond information as edge features. Grattarola et al. [18] presented the implementation of ECC in both classification and regression tasks for QM9 public data. ECC is a type of neural network that is designed to work with graph-structured data, where each data point (or node) is connected to other nodes via edges. Its purpose is to aggregate information from neighboring nodes in the graph, allowing them to learn meaningful representations of the data. To improve upon the ECC structure, one contribution in this chapter modifies the edge feature representation to distinguish aromatic cycles.

More generally, the results in Chapter 3 identified the benefit of the feature sets from Elton et al. [15] with our new ABOCH feature set towards improving the prediction of VUV spectra. Because deep learning embeds feature engineering within their algorithms, they are not designed to take advantage of externally defined features. In order to enable the integration of feature engineering from domain expert knowledge, we introduce a complementary deep learning architecture. To study this and the modified ECC edge features, we employ the same VUV spectra data set as Chapter 3, with 1397 molecules in the effective wavelength range of 126 nm – 240 nm and a resolution of 0.15 nm. The library of reference absorption spectra was provided by VUV Analytics company [99], which is available commercially. As in Chapter 3,

it should be noted that all measurements of the molecules were conducted in the gas phase, excluding the presence of a liquid medium and any interacting medium.

4.2.1 Contribution

Our main contribution introduces a complementary deep learning architecture to leverage the benefits of external features as part of the deep learning algorithm. This approach enables the prediction to take advantage of both the feature engineering conducted by deep learning algorithms and features derived from domain knowledge. Our new architecture embeds the external features via the feature encoder from a Multilayer Perceptron (MLP) neural network. This approach is studied for ECC and molformer [23] deep learning methods, but is generalizable to any deep learning approach.

A secondary contribution of this research modifies the edge features of the graph representation for molecules, specifically incorporating a distinct edge feature to represent aromatic bonds. Traditionally, chemistry literature classifies bonds based on the number of electron pairs shared between atoms, distinguishing between single, double, and triple bonds. In this study, our proposal entails the recognition of aromatic bonds as a novel type of bond, distinct from both single bonds and double bonds found outside of aromatic cycles.

4.3 Deep learning toolboxes

Two deep learning methods are studied in this paper, GNNs, and molformer. An overview of these methods can be found in Chapters 2 and 3. Here we describe the toolboxes employed for implementing deep learning.

Spektral Toolbox

For implementing GNNs, we employed the Spektral toolbox based on Keras. This toolbox is becoming increasingly popular thanks to its accessibility and intuition, even to non-technical audiences. Specifically, the Spektral toolbox is designed to provide tools for graph representation learning, which is the process of learning low-dimensional embeddings of nodes in a graph that capture important structural information about the graph. It provides a wide range of functionalities for working with graph data, including tools for building GNNs, preprocessing graph data, and visualizing graphs [18]. One of the most important options that this toolbox offers is a wide range of graph convolutional layers that can be used to build GNNs. These layers can be used to learn low-dimensional embeddings of nodes in a graph, which can then be used for downstream tasks such as node classification, link prediction, and graph classification. It is worth mentioning that our problem of prediction of VUV/UV spectra falls into the category of graph prediction when molecules are represented as graphs and VUV/UV spectra are considered as graph properties.

Molformer Toolbox

In recent years, the Transformer model has been successfully applied in the applications in the realm of natural language processing [2], and computer vision [84]. In the domain of Cheminformatics, molformer proves as a feasible framework to capture long-range interactions and dependencies within molecules. Leveraging self-attention mechanisms, molformer framework possesses the ability to attend to various segments of the molecule while incorporating information from the entire sequence [23]. This attention mechanism facilitates the identification of pertinent molecular features and the capture of subtle correlations that contribute to specific properties, such as VUV spectra in our study. Within the Molformer toolbox, multiple

techniques, including Sinusoidal Position Encoding, Convolutional Position Encoding, and Attentive Farthest Point Sampling, are implemented to extract information from two inputs: atom types and 3D coordinates, enabling the learning of relevant features for molecular representation.

4.4 Complementary Deep Learning Architecture

In terms of feature engineering, our proposed scheme incorporates additional features beyond the extensively studied ones described in Elton's work [15]. These include custom descriptor sets, sum over bond counts, E-state fingerprints, and oxygen balance. In total, our scheme introduces ten new integer-valued features, denoted as ABOCH. These encompass the number of cycles, number of aromatic cycles, number of aromatic atoms, number of olefin atoms, number of contiguous rotatable bond groups, number of conjugated double bonds, number of aromatic bonds, as well as individual features representing the counts of different halogen atoms. Details of how these features contribute to the characterization of molecules, and how the retrieve them were discussed in detail in Chapter 3.

The goal of complementary architecture is to incorporate external features within a deep learning algorithm. In Chapter 3, molecular feature representations derived from expert knowledge were found to be important for improving VUV spectral prediction. These features included custom descriptor sets, sum over bond counts, E-state fingerprints, and oxygen balances described by Elton et al. [15] and our new ABOCH features that encompass the number of cycles, number of aromatic cycles, number of aromatic atoms, number of olefin atoms, number of contiguous rotatable bond groups, number of conjugated double bonds, number of aromatic bonds, the counts of different halogen atoms. To incorporate these external features, we integrated the deep learning structure with an MLP structure as the external feature

encoder. Our complementary ECC architecture is depicted in Figure 4.1. The MLP layer enhances the model's capacity by introducing additional trainable parameters and enables the combination of learned patterns from different representations. Similar to the ECC architecture, we combine the features extracted from the deep learning structure of Molformer with our set of measurable features in the final layer, which leads to the generation of the predicted values for VUV spectra, as illustrated in Figure 4.2.

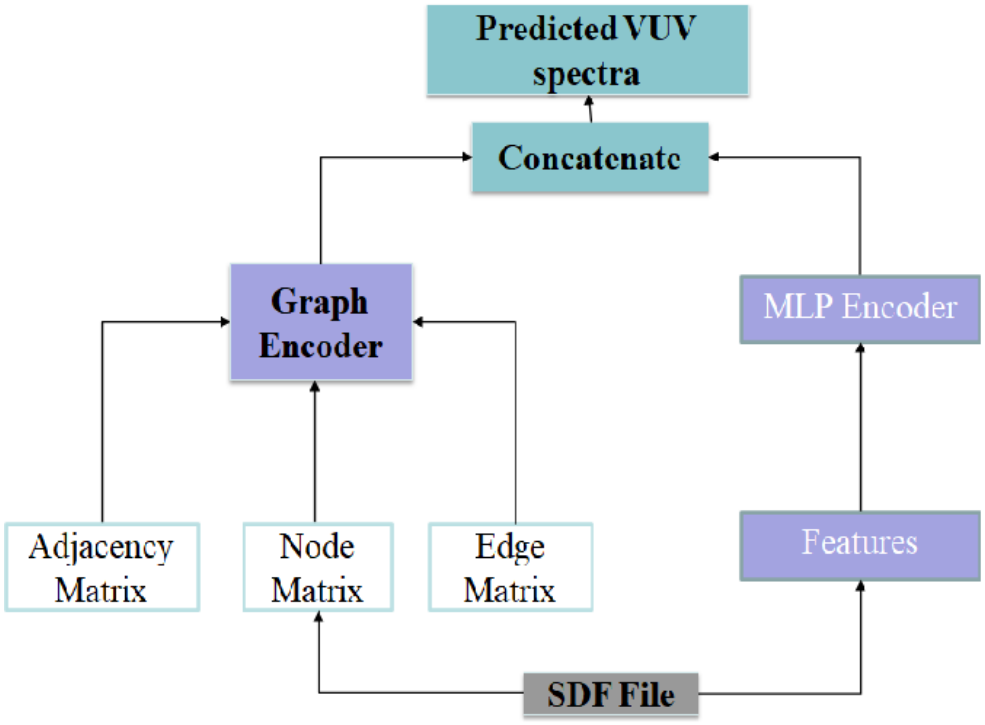

Figure 4.1. Architecture of modified ECC with the participation of molecular feature engineering.

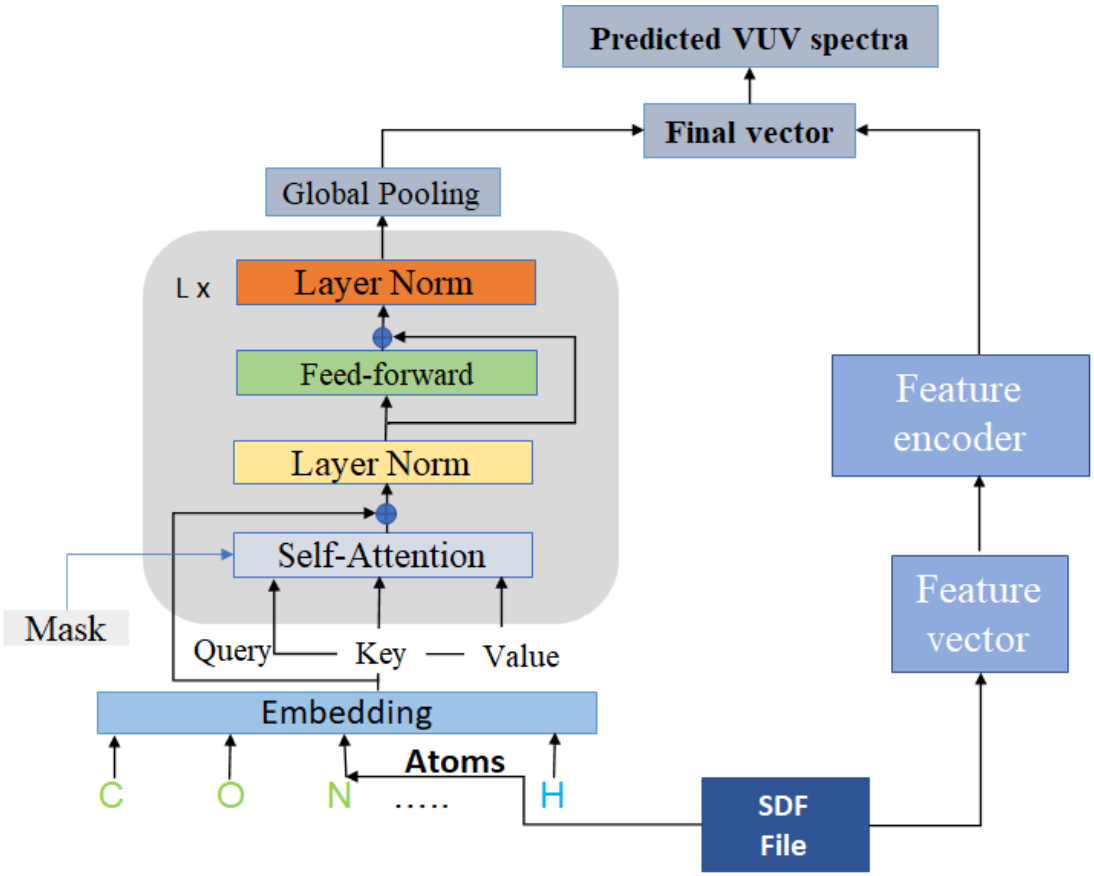

Figure 4.2. Architecture of modified molformer framework with the participation of molecular feature engineering.

4.5 Modified edge features for graph-based deep learning

ECC takes the node feature matrix, the adjacency matrix, and the edge feature matrix as inputs, and generates a new node feature matrix based on the ECC operation. The EdgeConditionedConv layer uses a neural network to generate the filter weights based on the edge and node features. In the conventional Spectral toolbox, edge features are extracted from original sdf files, wherein aromatic bonds within a cycle are depicted as three double bonds with single bonds adjacent to each other. However, it is known that the pi electrons are evenly delocalized over the cycle,

leading to its reduced reactivity compared to other compound types. This inherent stability has a significant impact on the chemical reactivity and overall properties of the molecule [107]. To address this, we introduce a unique edge feature to represent aromatic bonds. In this study, we propose the inclusion of aromatic bonds as a new bond type within a molecule, denoted as artificial type 4, as illustrated in Figure 4.3

Figure 4.3. Illustration of new edge features taking cypermethrin (CAS number: 52315-07-8) as an example (a) Original edge features as in spektral (b) Proposed edge features.

4.6 Computational results

The VUV dataset is the same as Chapter 3, with 1397 molecules in the training database and VUV spectral measurements over the effective wavelength range of 126 nm – 240 nm. Also as in Chapter 3, coefficient of determination (R^2 score) derived from 5-fold cross validation is employed.

Figure 4.4 illustrates the implementation of the ECC framework with complementary deep learning architecture represented by the MLP feature encoder on the right, and the modified edge features at the upper right (where the "4" indicates the aromatic cycle information for edges). Table 4.1 presents the comparison among the variants of ECC, including the original ECC framework, ECC with modified edge features, and the complementary ECC architecture with externally defined features

from Chapter 3 for both the original ECC and modified edge features. As a benchmark, the last row of the table presents the prediction performance using only the externally defined features from Chapter3 with a traditional MLP model.

ECC Original, representing the basic form of ECC without any modifications or additional features, exhibited a moderate R^2 score of 0.418, suggesting its partial ability to predict VUV spectra. However, introducing the aromatic cycle information in ECC Modified resulted in a higher R^2 score of 0.548, indicating the significant benefit of identifying aromatic cycles. The best performance is achieved using our complementary ECC architecture. While ECC Modified with the complementary structure shows a slightly higher R^2 score of 0.657 vs. 0.647, this difference is not considered significant. This result is logical since the ABOCH features from Chapter 3 also incorporate aromatic cycle information, so the modified ECC does not provide additional information. Finally, the complementary ECC deep learning structure does provide a small benefit above the traditional MLP structure (0.657 vs. 0.630), indicating that the feature engineering within the ECC algorithm is contributing to the prediction of VUV spectra.

Table 4.1. Averages of 5-fold cross-validated R^2 scores for ECC variants

Methods	R^2 Score
ECC Original	0.418
ECC Modified	0.548
ECC original + ABOCH combined features	0.647
ECC modified + ABOCH combined features	0.657
ABOCH combined features alone	0.630

To illustrate the effectiveness of learned patterns by machine learning algorithms, Figure 4.5 presents VUV spectral predictions for four "new" molecules that

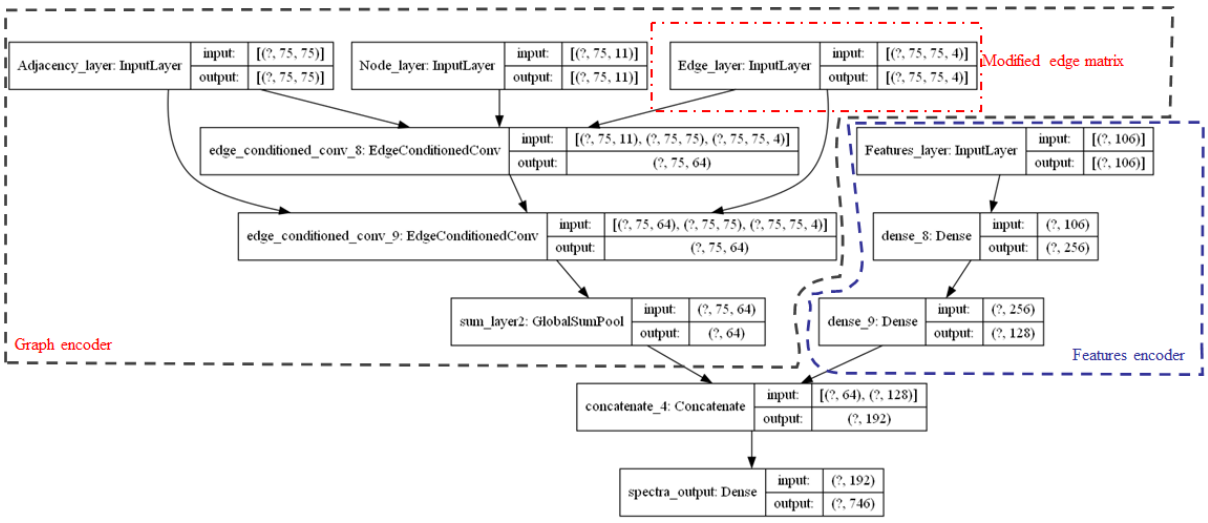

Figure 4.4. The modified ECC framework in TensorFlow module.

were not in the training database, namely hexahydrothymol, 4-methylethcathinone, 2,4-dimethylphenol, and 2,2,3,3,5,6,6-heptachlorobiphenyl. It is seen that the original ECC model successfully captures the underlying pattern and yielded comparable results to the modified edge features version (Aromatic ECC) and complementary ECC architecture (Aromatic ECC + ABOCH combined features) version of graph neural networks for hexahydrothymol and heptachlorobiphenyl. However, it is evident that the improved graph versions outperformed the original ECC model in terms of peak positions and magnitudes for 4-methylethcathinone and 2,4-dimethylphenol. Overall, our modifications yielded greater accuracy in predicting the precise locations of peaks and accurately estimating their intensities.

For the Molformer-based deep learning framework, Chapter 3 studied five variants proposed originally by Wu et al. [23]: Sinusoidal Position Encoding (SPE), Convolutional Position Encoding (CPE), Multi-scale Self-attention (MSA), Attentive Farthest Point Sampling (AFPS), and the Full model with all of the above. Table 4.2 presents the comparison of these five Molformer variants without and with our comple-

Table 4.2. Averages of 5-fold cross-validated R^2 scores for Molformer variants

	R^2 Score	Training time
SPE model	0.370	18.50 hours
SPE + ABOCH combined features	0.513	32.50 hours
CPE model	0.481	21.20 hours
CPE + ABOCH combined features	0.545	35.36 hours
MSA model	0.478	25.50 hours
MSA + ABOCH combined features	0.595	49.68 hours
AFPS model	0.494	30.50 hours
AFPS + ABOCH combined features	0.574	52.25 hours
Full model	0.603	45.50 hours
Full+ ABOCH combined features	0.608	72 hours
ABOCH combined features alone	0.630	2.50 hours

mentary deep learning architecture. The bottom row is the same benchmark as Table 4.1. Computational times are also provided for reference. While it can be seen that our complementary deep learning architecture yields a clear benefit, the molformer structure does not demonstrate improvement over the traditional MLP benchmark. This indicates that the Molformer deep learning structure is over-complicating the modeling task and critical information for predicting VUV structure is not effectively discovered by the algorithm.

This disparity in performance can be attributed to the challenges associated with monitoring a large-scale neural network model like Molformer, which poses inherent difficulties in the context of machine learning. The conventional perception of machine learning models primarily emphasizes measurable properties derived from feature engineering. The computational times also indicate the complexity of the Molformer (up to 72 hours) vs. the traditional MLP (2.5 hours).

4.6.1 Concluding remarks

In this paper, we implemented a modification to the graph-based representation of molecules by differentiating double bonds in aromatic cycles from regular double bonds. Our approach results in an average R^2 score improvement of 25%. Furthermore, by incorporating external features related to molecule properties using our complementary deep learning architecture, we achieved a further enhancement in the average R^2 score by 20%. A similar concept was also applied to the Molformer architecture. While the inclusion of external features improved the performance of the original models, it did not surpass the performance of using only the molecular features from Chapter 3 with a traditional MLP neural network model. It is worth noting that the Molformer architecture is relatively new and has the potential for unexplored modifications and variations. The training times of the models vary, depending on factors such as the framework's size and complexity. We speculate that the selection of hyperparameters and optimization algorithms plays a crucial role in model performance. Additionally, training deep learning models like Molformer can be computationally expensive, posing challenges in conducting a thorough hyperparameter optimization. Consequently, our work primarily focuses on conceptual advancements, leaving room for further research on hyperparameter exploration.

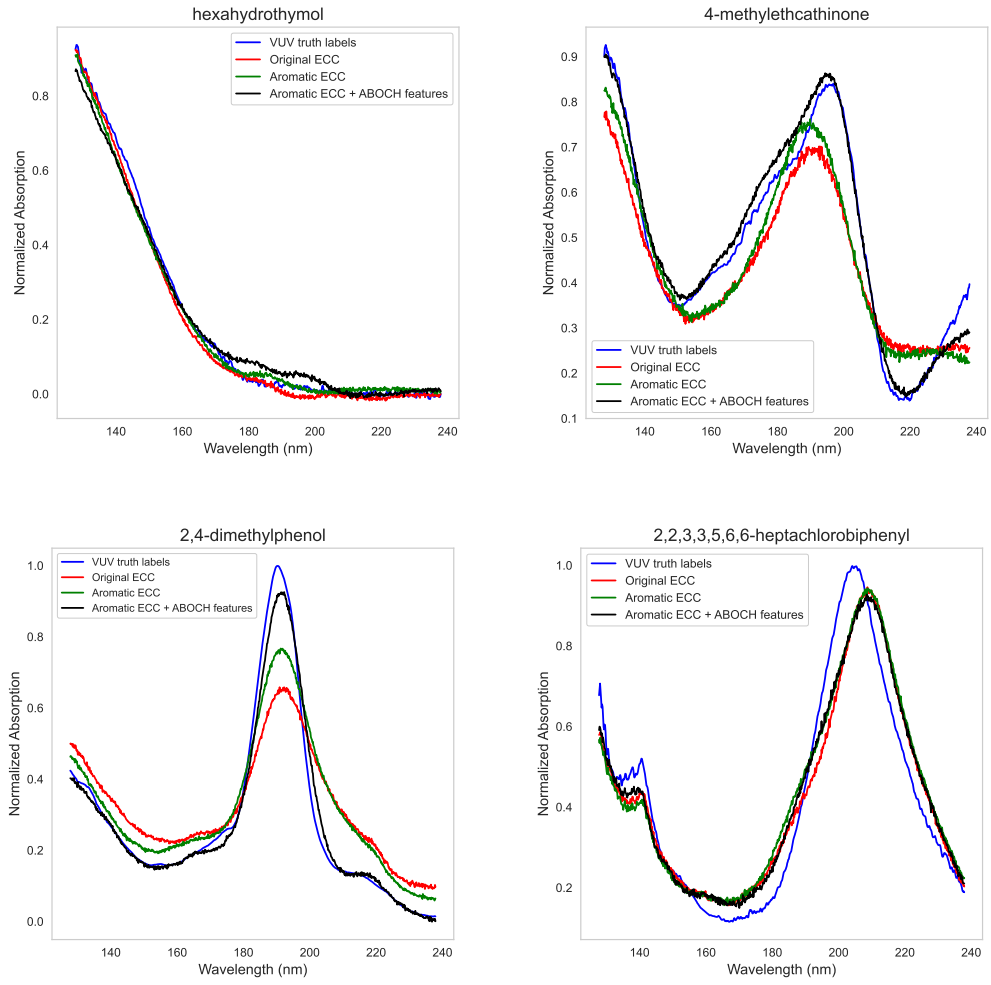

Figure 4.5. Prediction of new, novel molecules not in the database by three versions of graph neural networks: original ECC, aromatic ECC, and aromatic ECC with ABOCH combined features (a) hexahydrothymol (b) 4-methylethcathinone (c) 2,4-dimethylphenol (d) 2,2,3,3,5,6,6-heptachlorobiphenyl.

CHAPTER 5

Final Discussion and Future Work

The research presented in this dissertation addresses the development of machine-learning models for predicting VUV spectra. The findings relevant to advancement of machine learning and feature engineering for applications in chemistry, biotechnology, pharmaceutical studies, energy research, and fuel characterization. In this study, the VUV spectral measurements were obtained in the gas phase, so as to avoid errors due to interactions with a solvent. The specific wavelength range was from 126 nm to 240 nm at a resolution of 0.15 nm, yielding 746 VUV spectra outputs per molecule for this multiple-output prediction problem. Chapter 3 considered both traditional machine learning and deep learning and introduced a new "ABOCH" set of molecular feature representations, explicitly for improving VUV spectral prediction. Across all the comparisons, the recommended approach was using traditional machine learning with molecular feature representations derived from chemistry domain knowledge. This approach was not only superior for prediction but was also significantly faster computationally and provided interpretable models. Of the traditional machine learning methods implemented, the Random Forest Regressor had the highest R^2 performance metric, but the Gradient Boosted Tree Regressor performed similarly. In addition, it is revealing to note that seven of the ten ABOCH features were identified as important for VUV spectral prediction.

In Chapter 4, an examination of the deep learning architecture was conducted to explore potential improvement in VUV spectral prediction. Two modifications were studied. The main contribution was a complementary deep learning architecture that

enabled the inclusion of the molecular feature sets from Chapter 3 within deep learning algorithms. The secondary contribution was the edge encoding of double bonds in aromatic cycles in graph-based ECC models. Both modifications demonstrated superior predictive performance over the original deep learning structures. However, these improvements did not outperform the traditional and interpretable machine learning approaches from Chapter 3.

Given the success of featurization derived from chemical intuition in this dissertation, future research should continue this direction for feature engineering, including addressing directions identified in Chapter 3 following the examination of outlier cases and molecules with poor VUV spectral predictions. In future studies, the training database can be expanded to include more large molecules, with more aromatic cycles and cyclic groups, and corresponding features to describe these large molecules can be studied. In addition, given the already high number of molecular features and the fact that too many features can degrade the performance of machine learning, it would be useful to examine relationships between the features and identify redundant descriptors. An initial look at this is given in the Appendix with a correlation analysis of the features from Chapter 3. For improving the machine learning algorithms, hyperparameter optimization could be conducted to tune the modeling parameters. However, we do expect significant improvement from hyperparameter optimization since our findings demonstrate that the critical directions for improving prediction are molecular feature representations.

Finally, an interesting application of machine learning in this domain is computer-aided molecular design with the goal of generating new molecules with desired properties. This problem can be viewed as the reverse of the problem studied in this dissertation. Specifically, if we are given a desired VUV spectral pattern, how we can identify a set of molecules that match. A common strategy from a practical point

of view is to select the top molecules out of millions of cases in the public dataset and verify them experimentally. In theory, this challenge can be overcome by the use of a variational autoencoder [128, 129, 130, 131, 132], where the target properties can be incorporated with latent space in the encoding process and manipulated in the decoding process to retrieve the final set of molecules matching the target VUV spectra.

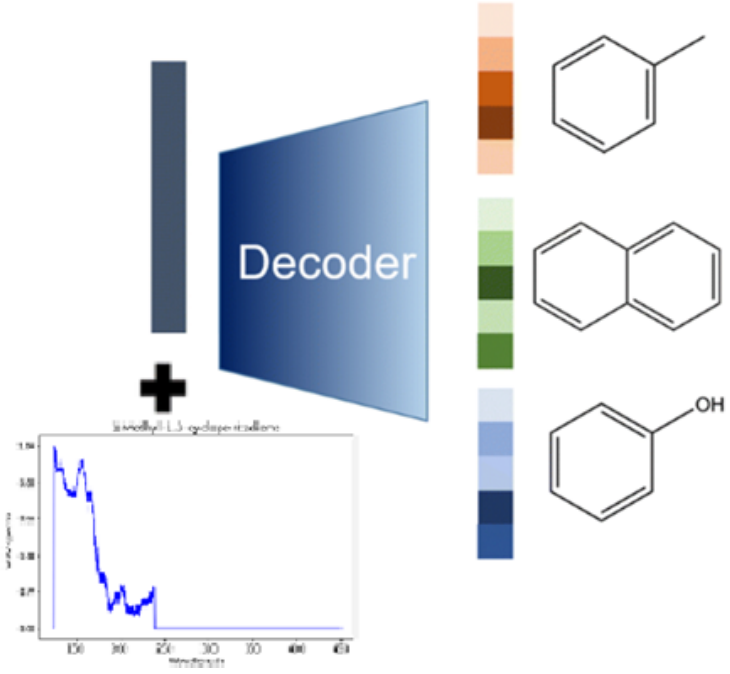

Figure 5.1. Generation of new molecules by the decoder with the measured VUV spectra as a constraint.

Figure 5.1 illustrates the conceptual framework for generating novel molecules using a trained variation autoencoder (VAE). Following the training process, the encoder maps the molecular structures onto a multi-dimensional latent space characterized by a predefined range. To generate a diverse set of solutions with desired target properties, namely a measured spectrum, the decoder employs two inputs. The first

input is the "artificial" latent space, which is predefined before the training process, to represent the target molecules (unknown), while the second input corresponds to the measured vacuum ultraviolet (VUV) spectra, acting as a conditioning factor.

Additional information related to this dissertation is provided in the Appendix. Additional information on molecular fingerprints and correlation analysis for ABOCH features are given in Appendix A. Appendix B provides background on graph-based deep learning, and Appendix C provides background on the Molformer approach.

APPENDIX A

Molecular Feature Representations

A.1 Molecular fingerprints

Molecular fingerprints were originally created to solve the problem of identifying isomers and later found to be useful for rapid substructure searching and the calculation of molecules in large molecular databases. In the past two decades, fingerprints have been used as an alternative to descriptors for QSPR studies. Fingerprinting algorithms transform the molecular graph into a vector populated with bits or integers. The RDKit graph fingerprints are a set of circular and path-based molecular fingerprints implemented in the RDKit cheminformatics toolkit, The RDKit graph fingerprints include several types of fingerprints as they can be generated using a single function call in the RDKit Python API. One advantage of the RDKit graph fingerprints is their computational efficiency and ease of use. They have been widely used in various cheminformatics applications, such as compound similarity search, and QSAR modeling, among others.

In this work, a brief description of several fingerprints found in RDKit, a popular cheminformatics package– Atom-Pair, Topological Torsion, Extended, E-state fingerprints, Avalon fingerprints, ErG fingerprints, and physiochemical property fingerprints.

A.1.1 Atom Pair fingerprints

Atom Pair fingerprint is a commonly used method in cheminformatics for generating a molecular descriptor that encodes the presence or absence of pairs of atoms within a molecule. This method involves enumerating all pairs of atoms within a molecule and hashing the resulting pairs into a fixed-length bit vector. This type of fingerprint has been shown to be effective in predicting molecular properties such as solubility, lipophilicity, and biological activity [133]. It is widely used in drug discov-

ery and other areas of cheminformatics for pattern searching, clustering, and machine learning.

The Atom Pair fingerprint algorithm starts by defining a "radius" around each atom in the molecule [101]. The radius is the number of bonds that need to be traversed from the atom to reach the neighboring atom. The algorithm then enumerates all pairs of atoms within the molecule that are separated by a distance less than or equal to a predefined maximum radius. For each pair of atoms, a hash function is applied to generate a unique hash code. The hash code is then used to set a bit in the fingerprint vector [134].

The resulting Atom Pair fingerprint is a binary vector, which corresponds to a particular pair of atoms within the molecule in each bits. A value of 1 indicates the presence of the pair of atoms, while a value of 0 indicates their absence [135].

A.1.2 Topological Torsion fingerprints

Topological Torsion (TT) fingerprint is another widely used molecular descriptor, which encodes the topological features of a molecule related to its three-dimensional shape. The descriptor is based on the concept of molecular torsion, which is defined as the angle between two sets of three consecutive bonds within a molecule. Compared to other fingerprinting methods, TT fingerprint is particularly effective in capturing the 3D structural features of a molecule, making it a powerful tool for the virtual screening of compounds for drug discovery [136].

The Topological Torsion fingerprint algorithm starts by identifying all possible sets of four contiguous atoms within a molecule that form a planar quadrilateral. For each set of four atoms, the torsion angle is calculated where each bit corresponds to a particular torsion angle[137].

A.1.3 E-state fingerprints

E-state fingerprint is a molecular descriptor that encodes the electronic properties of a molecule. The descriptor is based on the concept of atom types, where each atom in a molecule is assigned a unique electronic state or E-state [138]. The E-state of an atom is determined by its local chemical environment, including the types and identities of its neighboring atoms and the bond orders between them.

The E-state fingerprint algorithm involves identifying all the atoms in a molecule and assigning them unique E-states. The E-states are then hashed into a fixed-length bit vector using a hash function, resulting in the E-state fingerprint. The resulting E-state fingerprint is a binary vector of fixed length, where each bit corresponds to a particular E-state [139].

A.1.4 Avalon fingerprints

Avalon fingerprint is designed by the circular substructures or "patterns" in a molecule, which is defined as sets of atoms and bonds that form a closed loop. The size of the circular substructures can be varied, allowing for different levels of detail and complexity to be captured in the fingerprint [140]. The most common sizes used in practice are 2, 3, 4, and 6, corresponding to patterns with 3, 4, 5, and 7 atoms respectively. The Avalon fingerprint is robust to changes in atom order, stereochemistry, and tautomerism, making it a reliable descriptor for diverse sets of molecules. It is also highly scalable, allowing for the efficient processing of large databases of compounds. This makes it a popular choice for drug discovery.

A.1.5 Extended Reduced Graph (ErG) fingerprints

Extended Reduced Graph is fingerprint is an extension of the Reduced Graph (RG) fingerprint, which represents a molecule as a graph by removing its atomic

details and focusing on its structural connectivity. In ERG fingerprint, each atom and bond in the graph is labeled based on its properties, such as its hybridization state, formal charge, and atom type. These labels are used to capture the environment of each atom and bond in the molecule. The labeled graph is then hashed into a fixed-length binary vector using a hash function, resulting in the ERG fingerprint [141].

Compared to other fingerprinting methods, ERG fingerprint has several advantages. It is robust to different tautomeric forms and can handle a wide range of molecular sizes. It also has a high information content, allowing for the capture of a large number of substructures and their combinations.

A.2 Multicollinearity in proposed ABOCH feature set

Figure A.1 illustrates a strong correlation between the number of aromatic bonds, the number of aromatic atoms, and the number of aromatic rings. This correlation arises from the resonance-based bonding system formed by the delocalized π-electrons in aromatic compounds. Typically, a normal aromatic ring contains one aromatic ring, six aromatic atoms, and six aromatic bonds.

However, certain fused ring systems, such as naphthalene-1,8-dione, exhibit deviations from the typical aromatic bonding pattern. In this molecule, which comprises two benzene rings fused together, carbonyl groups (C=O) are attached to each ring. As a result, two aromatic atoms in close proximity do not form an aromatic bond.

Despite this deviation, the three distinct features associated with aromaticity remain valid in characterizing molecules. Conversely, other features in the dataset demonstrate minimal correlation, thus highlighting their independent and valid contributions to the feature set.

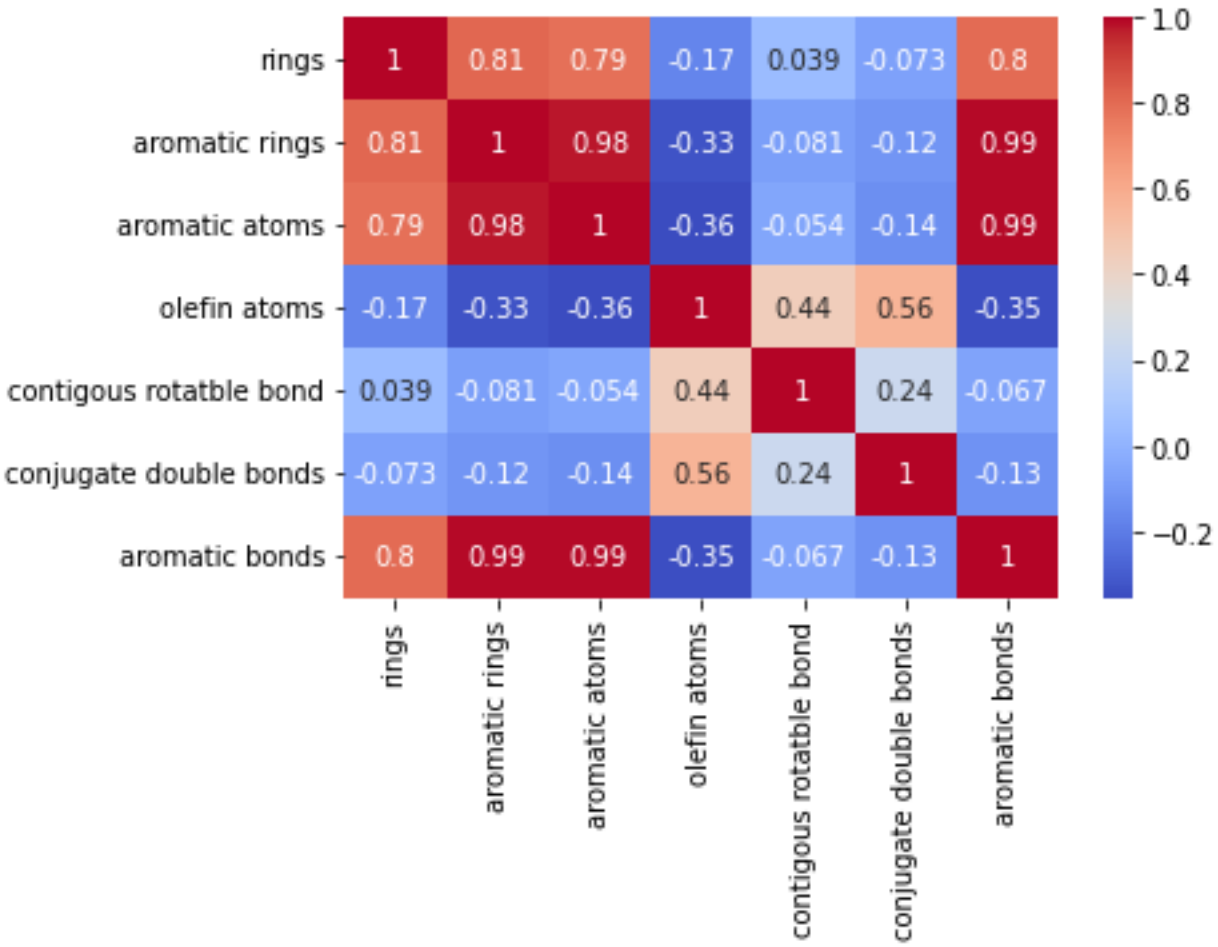

Figure A.1. **Correlation analysis to test multicollinearity in proposed feature set.**

APPENDIX B

Graph Neural Networks

B.1 Data processing for graph neural network models

Atoms in one molecule can be represented by characters such as **C**: Carbon, **H**: Hydrogen, **O**: Oxygen,... and then they are encoded to their corresponding number in the periodic table, forming input node matrix for graph neural network. Features in a node matrix of a corresponding molecule is its atom type As depicted in Figure

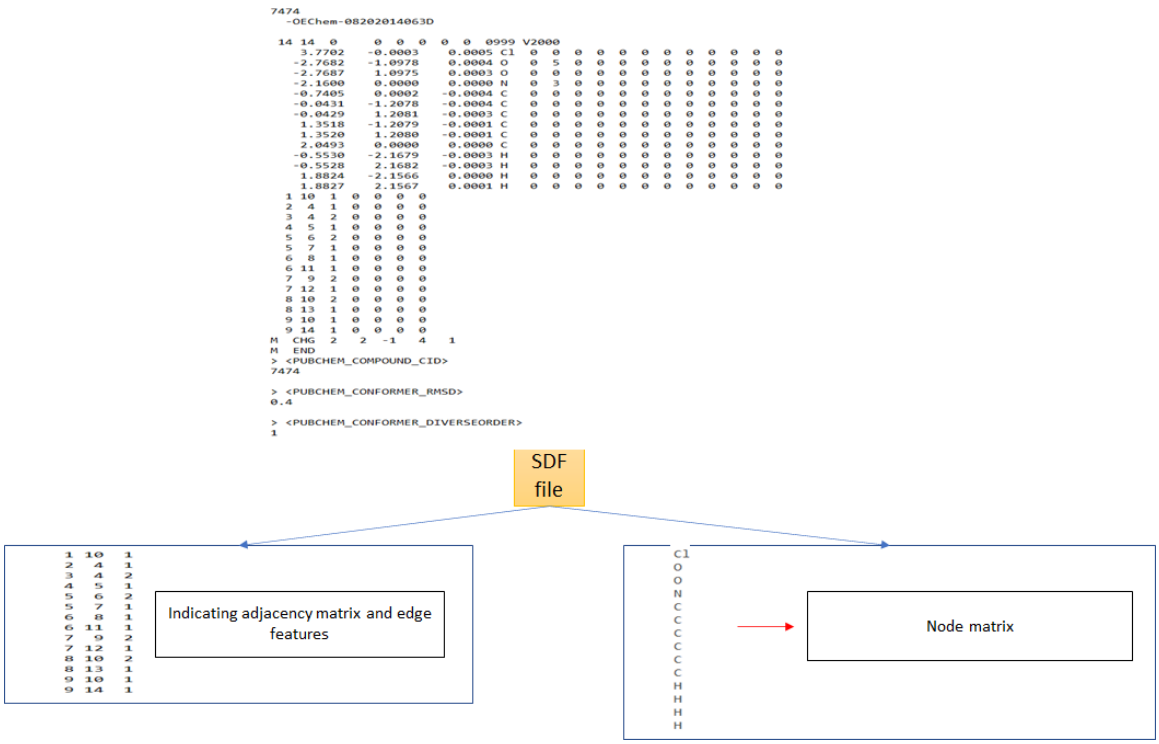

Figure B.1. **Processing graph representative inputs from a raw SDF file**.

B.1, the adjacency matrix gives information on the connection between atoms in a molecule and edge features are the properties of bond types, which are extracted from .sdf file.

B.2 Graph Attention Network

Graph Attention Network is an attention-based architecture to perform node classification of graph-structured data.

The idea is to compute the hidden representations of each node in the graph, by computing the scores attaching the node to its neighbors, following an attention strategy. The attention architecture has several interesting properties: (1) the operation is efficient since it is parallelizable across node neighbor pairs; (2) it can be applied to graph nodes having different degrees by specifying arbitrary weights to the neighbors; and (3) the model is directly applicable to inductive learning problems, including tasks where the model has to generalize to completely unseen graphs.

- $Z = \alpha XW + b$

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(a^{\top}[(XW)_i \,\|\, (XW)_j]\right)\right)}{\sum_{k \in \mathcal{N}(i)\cup\{i\}} \exp\left(\text{LeakyReLU}\left(a^{\top}[(XW)_i \,\|\, (XW)_k]\right)\right)}$$

Notation:

1. X: node attribute matrix
2. W: trainable weights matrices

B.3 Edge-Conditioned Convolution

A mathematically sound definition of the convolution operator makes use of the spectral analysis theory, where it corresponds to the multiplication of the signal on vertices transformed into the spectral domain by graph Fourier transform.

- This layer computes for each node i:

$$x'_i = \sum_{j \in \mathcal{N}(i)} \text{MLP}\big(x_i \| x_j - x_i\big)$$

where MLP is a multi-layer perceptron.

- Deep Networks with Edge-Conditioned Convolution (ECC): The information from the local neighborhoods gets combined over successive layers to gain context [18]. The information from the edges in the case of molecular graph representation is extracted from the original sdf where edge (vertex) properties are bond types: single, double or triple and encoded by corresponding real numbers of 1, 2 and 3.

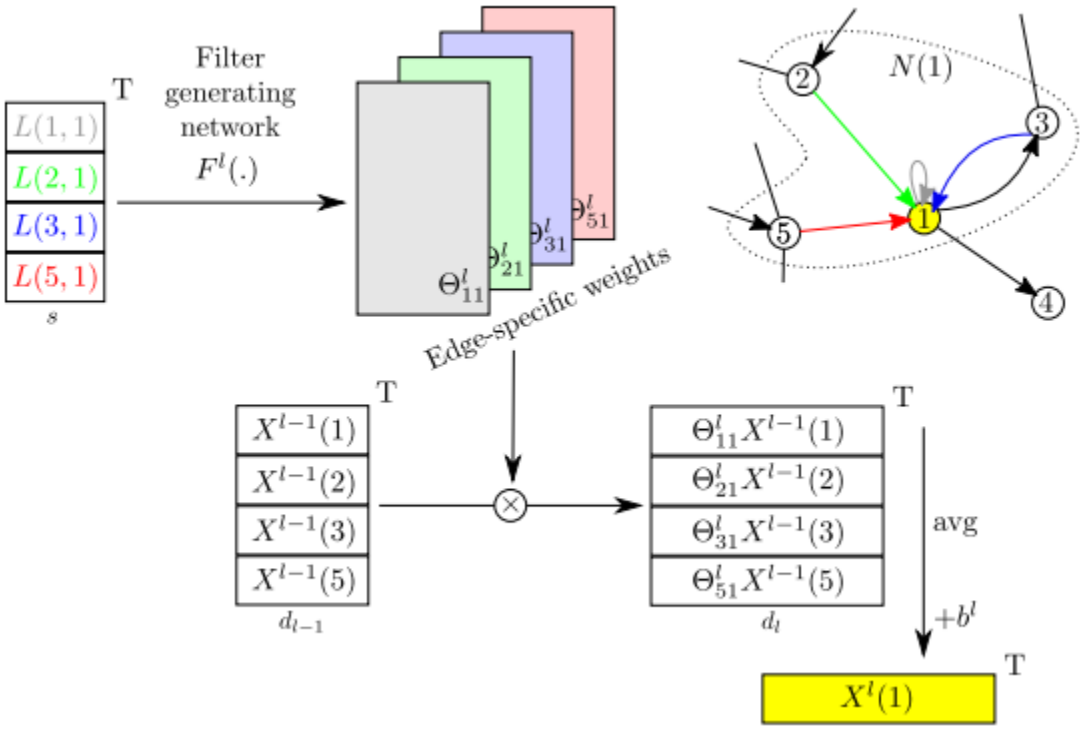

Figure B.2. Illustration of edge-conditioned convolution as presented in [18].

APPENDIX C

Molformer techniques

C.1 Introduction to molformer framework

The framework in Molformer exploits 3D molecular geometry, as depicted in Figure C.1, finding 3D translations and rotations is an underlying principle for molecular representation learning. The idea of the Transformer encoder and AlphaFold2 is to apply a convolutional operation to the pairwise distance matrix with the kernel size of (1,1). Consequently, the attention score is computed to control the impact of interatomic distance over the attention score. This molecular representation learning has been deployed to a number of drug discovery and material design problems. A similar framework can be adopted to predict VUV spectra.

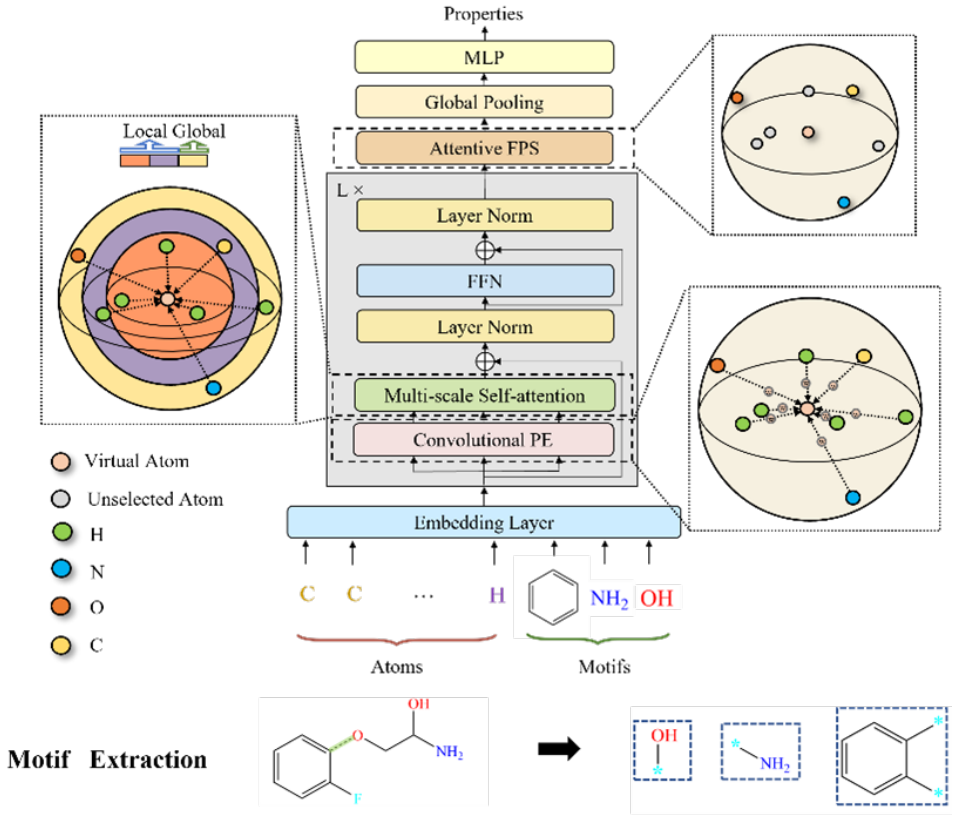

Figure C.1. **An overview of molformer framework as presented in Wu's work [23].**

C.2 Data processing and positional embedding

A molecule has n atoms and c atom classes, which contain the one-hot atom representations of the 3D coordinates of each atom. Atoms in one molecule can be represented by characters such as **C**: Carbon, **H**: Hydrogen, **O**: Oxygen,... and then they are encoded to their corresponding number in the periodic table, forming two inputs **atomic charges** and **positions** respectively. **N** is so-called the maximum number of atoms for one molecule in the dataset, if one molecule has a smaller number of atoms than the maximum number, 0 are imputed to standardize the data, making each molecule has an atomic charge with the size of $(N, 1)$ and positions of $(N, 3)$ The embedding layer, a well-known module in language-processing machine learning

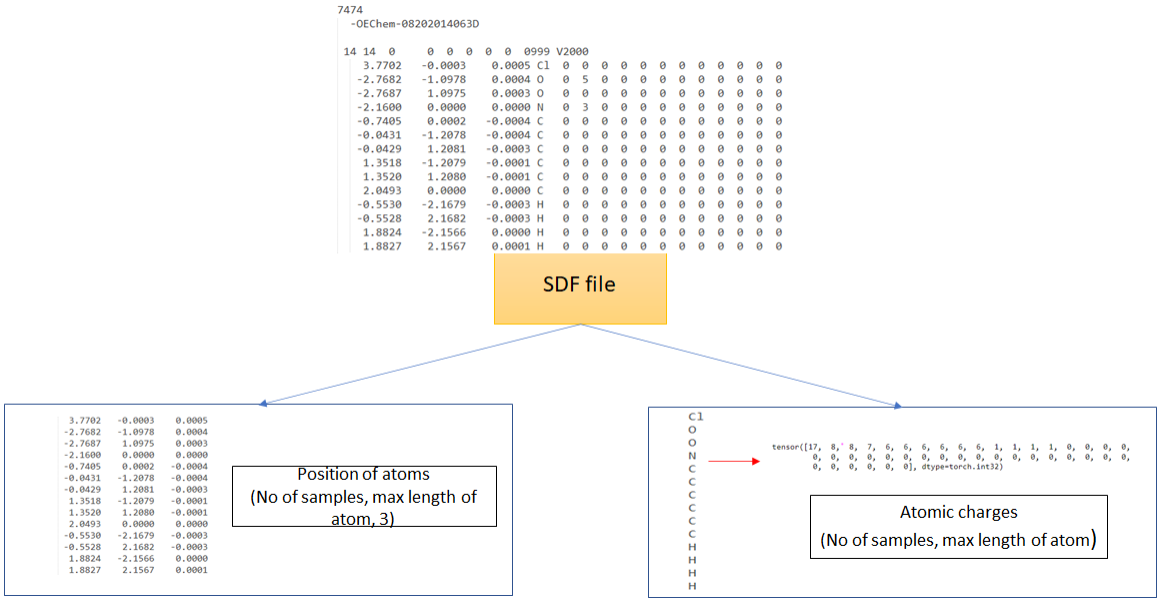

Figure C.2. **Processing inputs for molformer models from original sdf**.

models, transform the dimension of atomic charges from 1 to $embed_dim$, which is set to be 512 by default. It is worth noting that the embedding dimension should be

divisible by the number of heads in the multi-head mechanism is presented in the following section.

C.3 Details of molformer

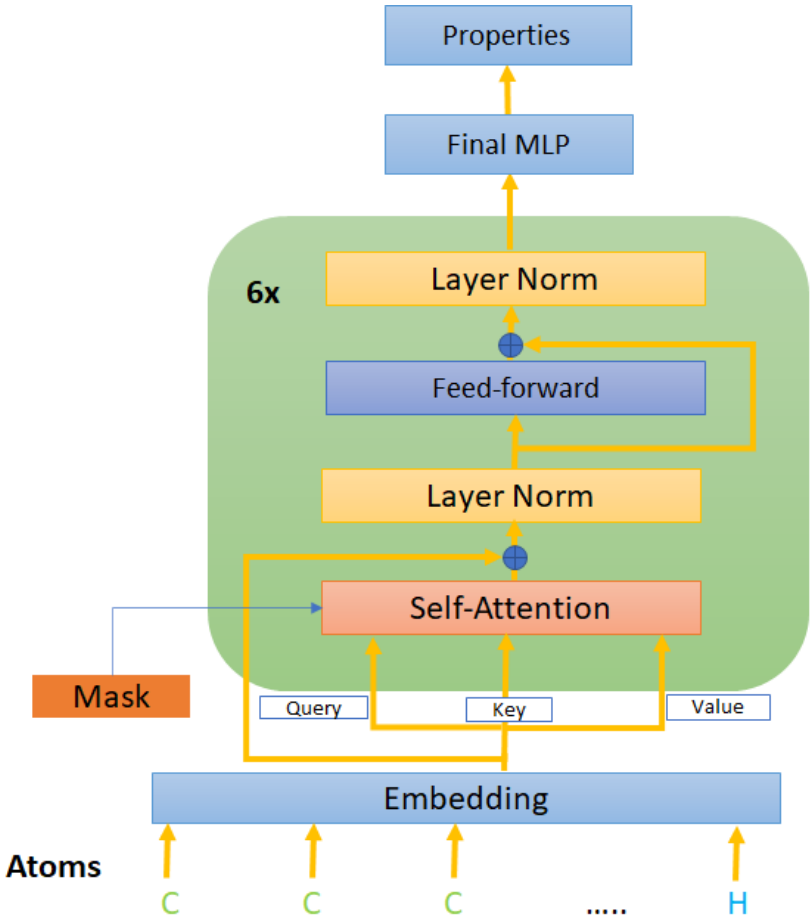

Figure C.3. **Diagram of one variant of molformer model with Sinudoial Position Encoding (SPE)**.

Details of positional encoding for SPE model can be found in Figure C.4 and details of multi-head mechanism are illustrated in Figure C.5

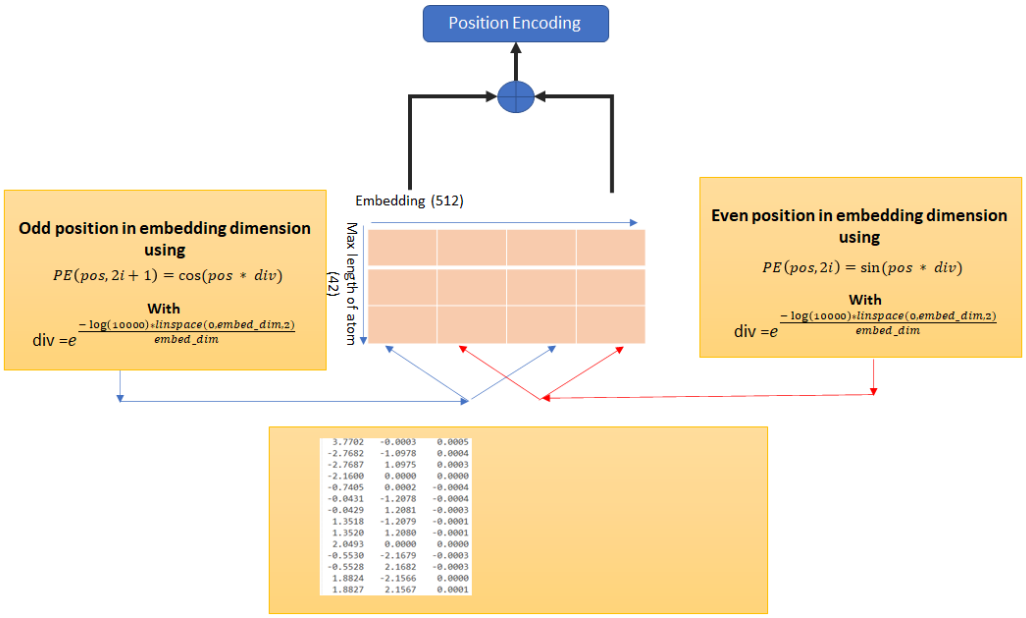

Figure C.4. **Illustration of positional encoding in SPE model.**

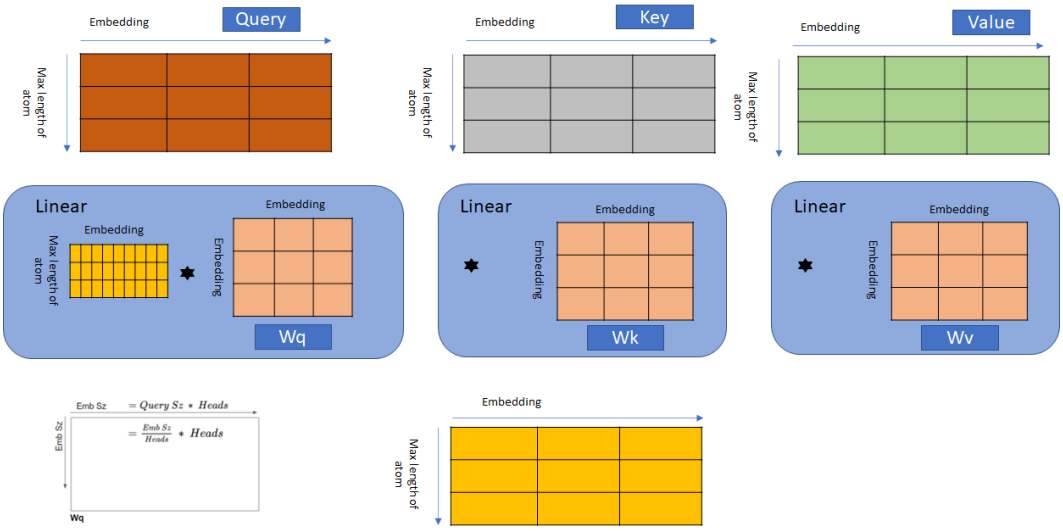

Figure C.5. **Illustration of multi-head mechanism.**

After combining the output of the embedding layer and positional encoding layer, an attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The

output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. The input consists of queries and keys of dimension d_k. This is fed to all three parameters, Query, Key, and Value in the Self-Attention which then also produces an encoded representation for each atom in the molecule sequence. The Self-Attention module also adds its own attention scores to each word's representation. We compute the dot products of the query with all keys, divide by $\sqrt{d_k}$, and apply a softmax function to obtain the weights on the values, as depicted in Figure C.6

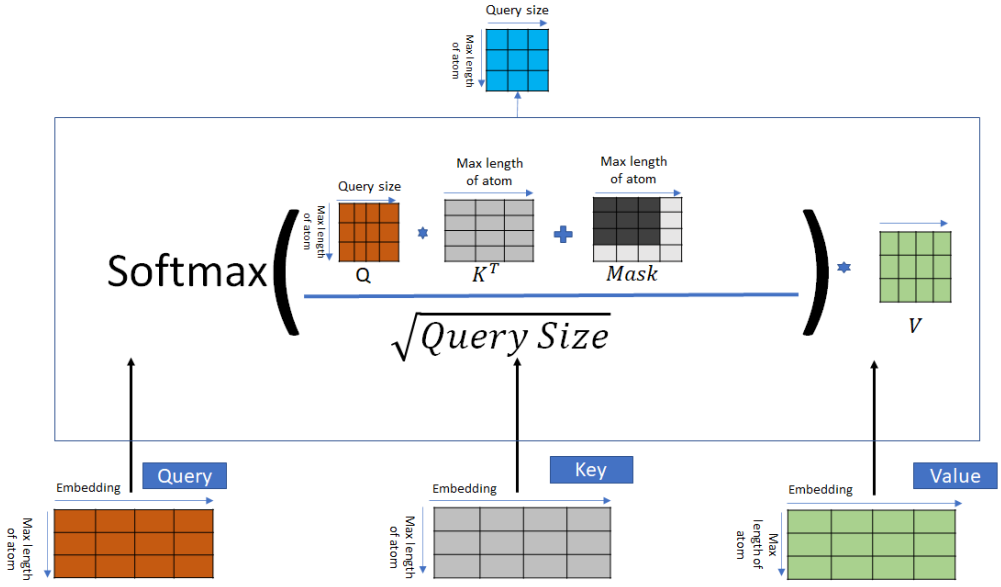

Figure C.6. **Illustration of scale-dot product with Q, K, V as inputs**.

Instead of performing a single attention function with model-dimensional keys, values, and queries, It is beneficial to linearly project the queries, keys, and values h times with different, learned linear projections. On each of these projected versions of queries, keys, and values we then perform the attention function in parallel, yielding

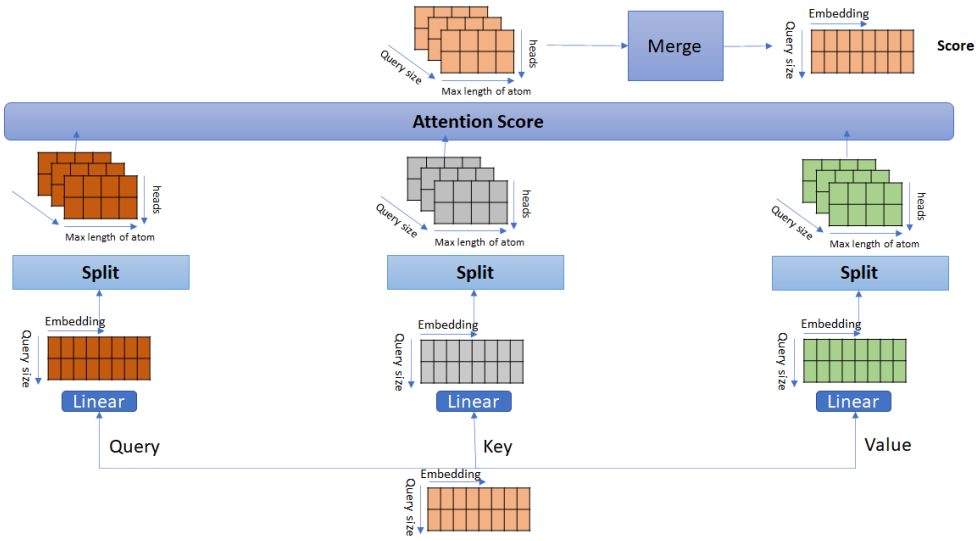

Figure C.7. **Illustration of score update after multi-head separation**.

number of heads dimensional output values. These are concatenated and once again projected, resulting in the final values, as depicted in Figure C.7.

Multi-scaled attention (MSA) model

Transformers are powerful sequence models capable of passing information from any position to any other position. However, they are not trivially applied to a set of aligned sequences. The main contribution MSA model is to extend the transformer to operate better while respecting its structure matrix. Guo et al. (2020b) propose to use integer-based distance to limit attention to local word neighbors, which cannot be used in molecules. This is because different types of molecules have different densities and molecules of the same type have different spatial regularity, which results in the nonuniformity of interatomic distances. To address that, a new multi-scale methodology is designed to robustly capture details. An illustration of the MSA model can be found in Figure C.8.

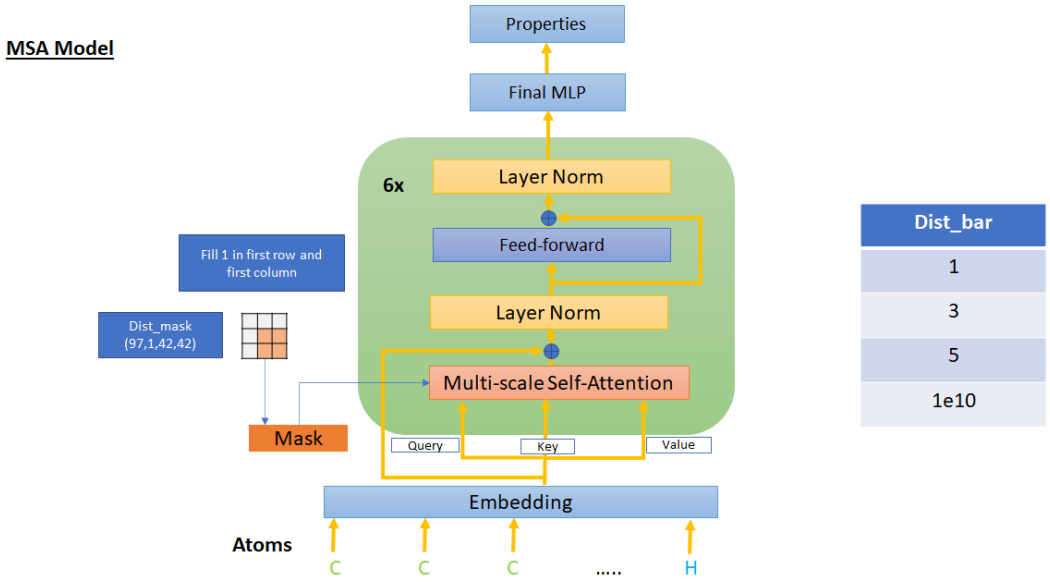

Figure C.8. **Illustration of MSA model as one variant of molformer-based techniques**.

Attentive furthest point sampling (AFPS) model

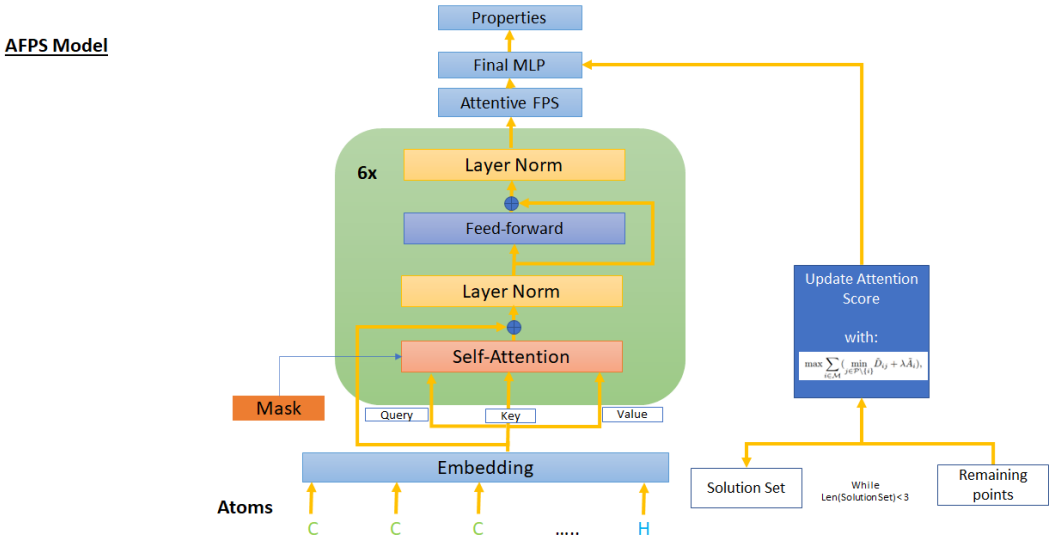

Figure C.9. **Illustration of AFPS model as one variant of molformer-based techniques**.

The model not only characterizes the atomic local environment by propagating node information from nearby nodes to more distant ones but also allows for non-local effects at the intramolecular level. For input, we assume that a molecule, its positions, and its atomic charge features are extracted like other models. Because the model is atom-centric, each atom has its own neighbor features that concatenate both neighboring atoms and the connecting bond features. An illustration of the AFPS model can be found in Figure C.9.

REFERENCES

[1] K. A. Schug, I. Sawicki, D. D. Carlton, H. Fan, H. M. McNair, J. P. Nimmo, P. Kroll, J. Smuts, P. Walsh, and D. Harrison, "Vacuum ultraviolet detector for gas chromatography," *Analytical Chemistry*, vol. 86, pp. 8329–8335, 8 2014.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 6 2017.

[3] C. Maggi, J. Elder, W. Fundamenski, R. Giannella, L. Horton, K. Lawson, A. Loarte, A. Maas, R. Reichle, M. Stamp, P. Stangeby, and H. Summers, "Measurement and analysis of radiated power components in the jet mki divertor using vuv spectroscopy," *Journal of Nuclear Materials*, vol. 241-243, pp. 414–419, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022311597800748

[4] N. J. Mason, A. Dawes, P. D. Holtom, R. J. Mukerji, M. P. Davis, B. Sivaraman, R. I. Kaiser, S. V. Hoffmann, and D. A. Shaw, "Vuv spectroscopy and photo-processing of astrochemical ices: an experimental study," *Faraday Discuss.*, vol. 133, pp. 311–329, 2006. [Online]. Available: http://dx.doi.org/10.1039/B518088K

[5] F. Wien, A. J. Miles, J. G. Lees, S. Vrønning Hoffmann, and B. A. Wallace, "VUV irradiation effects on proteins in high-flux synchrotron radiation circular dichroism spectroscopy," *Journal of Synchrotron Radiation*, vol. 12, no. 4, pp. 517–523, Jul 2005. [Online]. Available: https://doi.org/10.1107/S0909049505006953

[6] J. Schenk, D. D. Carlton, J. Smuts, J. Cochran, L. Shear, T. Hanna, D. Durham, C. Cooper, and K. A. Schug, "Lab-simulated downhole leaching of formaldehyde from proppants by high performance liquid chromatography (hplc), headspace gas chromatography-vacuum ultraviolet (hs-gc-vuv) spectroscopy, and headspace gas chromatography-mass spectrometry (hs-gc-ms)," *Environ. Sci.: Processes Impacts*, vol. 21, pp. 214–223, 2019. [Online]. Available: http://dx.doi.org/10.1039/C8EM00342D

[7] I. C. Santos and K. A. Schug, "Recent advances and applications of gas chromatography vacuum ultraviolet spectroscopy," *Journal of Separation Science*, vol. 40, pp. 138–151, 1 2017.

[8] H. Fan, J. Smuts, P. Walsh, D. Harrison, and K. A. Schug, "Gas chromatography–vacuum ultraviolet spectroscopy for multiclass pesticide identification," *Journal of Chromatography A*, vol. 1389, pp. 120–127, 4 2015.

[9] C. Qiu, J. Smuts, and K. A. Schug, "Analysis of terpenes and turpentines using gas chromatography with vacuum ultraviolet detection," *Journal of Separation Science*, vol. 40, pp. 869–877, 2 2017.

[10] R. Zimmermann, C. Lermer, K. Schramm, A. Kettrup, and U. Boesl, "Three-dimensional trace analysis: Combination of gas chromatography, supersonic beam uv spectroscopy and time-of-flight mass spectrometry," *European Mass Spectrometry*, vol. 1, no. 4, pp. 341–351, 1995. [Online]. Available: https://doi.org/10.1255/ejms.118

[11] "Uv-vis spectroscopy: Principle, strengths and limitations and applications." [Online]. Available: https://www.technologynetworks.com/analysis/articles/uv-vis-spectroscopy-principle-strengths-and-limitations-and-applications-349865

[12] P. Gedeck, B. Rohde, and C. Bartels, "Qsar how good is it in practice? comparison of descriptor sets on an unbiased cross section of corporate data sets," *Journal of Chemical Information and Modeling*, vol. 46, pp. 1924–1936, 9 2006.

[13] G. E. Dahl, N. Jaitly, and R. Salakhutdinov, "Multi-task neural networks for qsar predictions," 6 2014.

[14] L. K. Tsou, S.-H. Yeh, S.-H. Ueng, C.-P. Chang, J.-S. Song, M.-H. Wu, H.-F. Chang, S.-R. Chen, C. Shih, C.-T. Chen, and Y.-Y. Ke, "Comparative study between deep learning and qsar classifications for tnbc inhibitors and novel gpcr agonist discovery," *Scientific Reports*, vol. 10, p. 16771, 10 2020.

[15] D. C. Elton, Z. Boukouvalas, M. S. Butrico, M. D. Fuge, and P. W. Chung, "Applying machine learning techniques to predict the properties of energetic materials," *Scientific Reports*, vol. 8, p. 9059, 6 2018.

[16] M. Korshunova, B. Ginsburg, A. Tropsha, and O. Isayev, "Openchem: A deep learning toolkit for computational chemistry and drug design," *Journal of Chemical Information and Modeling*, vol. 61, pp. 7–13, 1 2021.

[17] G. A. Pinheiro, J. Mucelini, M. D. Soares, R. C. Prati, J. L. F. D. Silva, and M. G. Quiles, "Machine learning prediction of nine molecular properties based on the smiles representation of the qm9 quantum-chemistry dataset," *The Journal of Physical Chemistry A*, vol. 124, pp. 9854–9866, 11 2020.

[18] D. Grattarola and C. Alippi, "Graph neural networks in tensorflow and keras with spektral [application notes]," *IEEE Computational Intelligence Magazine*, vol. 16, pp. 99–106, 2 2021.

[19] M. Tsubaki and T. Mizoguchi, "Quantum deep field: Data-driven wave function, electron density generation, and atomization energy prediction and extrapolation with machine learning," *Physical Review Letters*, vol. 125, p. 206401, 11 2020.

[20] G. B. Goh, C. Siegel, A. Vishnu, N. O. Hodas, and N. Baker, "How much chemistry does a deep neural network need to know to make accurate predictions?" 10 2017.

[21] G. B. G. et all, "Chemception: A deep neural network with minimal chemistry knowledge matches the performance of expert-developed qsar/qspr models," 6 2017.

[22] S. Wang, T. Kind, D. J. Tantillo, and O. Fiehn, "Predicting in silico electron ionization mass spectra using quantum chemistry," *Journal of Cheminformatics*, vol. 12, p. 63, 12 2020.

[23] F. Wu, D. Radev, and S. Z. Li, "Molformer: Motif-based transformer on 3d heterogeneous molecular graphs," 10 2021.

[24] F. Urbina, K. Batra, K. J. Luebke, J. D. White, D. Matsiev, L. L. Olson, J. P. Malerich, M. A. Z. Hupcey, P. B. Madrid, and S. Ekins, "Uv-advisor: Attention-based recurrent neural networks to predict uv–vis spectra," *Analytical Chemistry*, vol. 93, pp. 16 076–16 085, 12 2021.

[25] E.-R. Kenawy, A. Ghazy, A. Al-Hossainy, H. Rizk, and S. Shendy, "Synthesis, characterization, td-dft method, and optical properties of novel nanofiber conjugated polymer," *Synthetic Metals*, vol. 291, p. 117206, 12 2022.

[26] M. Lebel, T. Very, E. Gloaguen, B. Tardivel, M. Mons, and V. Brenner, "Excited states computation of models of phenylalanine protein chains: Td-dft and composite cc2/td-dft protocols," *International Journal of Molecular Sciences*, vol. 23, no. 2, 2022. [Online]. Available: https://www.mdpi.com/1422-0067/23/2/621

[27] N. Almutlaq, M. M. Elshanawany, M. Sayed, O. Younis, M. Ahmed, J. Wachtveitl, M. Braun, M. S. Tolba, A. F. Al-Hossainy, and A. A. Abozeed, "Synthesis, structural, td-dft, and optical characteristics of indole derivatives,"

Current Applied Physics, vol. 45, pp. 86–98, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S156717392200270X

[28] E.-R. kenawy, A. Ghazy, A. F. Al-Hossainy, H. F. Rizk, and S. Shendy, "Synthesis, characterization, td-dft method, and optical properties of novel nanofiber conjugated polymer," *Synthetic Metals*, vol. 291, p. 117206, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0379677922002004

[29] Y. Wu, X. Su, C. Xie, R. Hu, X. Li, Q. Zhao, G. Zheng, and J. Yan, "First cycloruthenation of 2-alkenylpyridines: synthesis, characterization and properties," *RSC Advances*, vol. 11, pp. 4138–4146, 2021.

[30] M. S. Zoromba, F. Alharbi, A. F. Al-Hossainy, and M. H. Abdel-Aziz, "Preparation of hybrid conducting polymers blend nanocomposite for energy conversion using experimental data and td-dft/dmol3 computations," *Journal of Materials Research and Technology*, vol. 23, pp. 2852–2867, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2238785423002090

[31] V. Kumar, S. Kumar, and P. Rani, "Pharmacophore modeling and 3d-qsar studies on flavonoids as - glucosidase inhibitors," *Der Pharma Chem*, vol. 2, 01 2010.

[32] F. Sadeghi, A. Afkhami, T. Madrakian, and R. Ghavami, "Qsar analysis on a large and diverse set of potent phosphoinositide 3-kinase gamma (pi3k) inhibitors using mlr and ann methods," *Scientific Reports*, vol. 12, p. 6090, 04 2022.

[33] V. Bastikar, A. Bastikar, P. Alpana, P. Khadke, and D. Alessandro, "Qsar of topoisomerase i inhibitors using cluster based descriptor procedure for camptothecin analogs," vol. 2, pp. 28–48, 01 2013.

[34] W. Tong, H. Hong, Q. Xie, L. Shi, H. Fang, and R. Perkins, "Assessing qsar limitations - a regulatory perspective," *Current Computer - Aided Drug Design*, vol. 1, pp. 195–205, 04 2005.

[35] Y. Pan, J. Jiang, R. Wang, H. Cao, and J. Zhao, "Quantitative structure–property relationship studies for predicting flash points of organic compounds using support vector machines," *QSAR Combinatorial Science*, vol. 27, pp. 1013 – 1019, 08 2008.

[36] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, and A. Tropsha, "Qsar modeling: Where have you been? where are you going to?" *Journal of Medicinal Chemistry*, vol. 57, pp. 4977–5010, 6 2014.

[37] A. Tropsha, "Best practices for qsar model development, validation, and exploitation," *Molecular Informatics*, vol. 29, pp. 476–488, 7 2010.

[38] X. Li and Y. Sui, "Multiple regression and k-nearest-neighbor based algorithm for estimating missing values within sensor," in *2021 7th Annual International Conference on Network and Information Systems for Computers (ICNISC)*, 2021, pp. 613–618.

[39] K. Pliakos, P. Geurts, and C. Vens, "Global multi-output decision trees for interaction prediction," *Machine Learning*, vol. 107, pp. 1257–1281, 9 2018.

[40] H. Gensheng and L. Dong, "Multi-output support vector machine regression and its online learning," in *2008 International Conference on Computer Science and Software Engineering*, vol. 4, 2008, pp. 878–881.

[41] P. Hajek and R. Henriques, "Correction: Modelling innovation performance of european regions using multi-output neural networks," *PLOS ONE*, vol. 12, p. e0189746, 12 2017.

[42] M. Breskvar, D. Kocev, and S. Džeroski, "Ensembles for multi-target regression with random output selections," *Machine Learning*, vol. 107, pp. 1673–1709, 11 2018.

[43] Z. Zhang and C. Jung, "Gbdt-mo: Gradient-boosted decision trees for multiple outputs," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 7, pp. 3156–3167, 2021.

[44] "Bagging- 25 questions to test your skills on random forest algorithm." [Online]. Available: https://www.analyticsvidhya.com/blog/2021/05/bagging-25-questions-to-test-your-skills-on-random-forest-algorithm/

[45] T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer, "Robust and accurate shape model fitting using random forest regression voting," in *Computer Vision – ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 278–291.

[46] A. P. Lind and P. C. Anderson, "Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties," *PLOS ONE*, vol. 14, p. e0219774, 7 2019.

[47] D. S. Palmer, N. M. O'Boyle, R. C. Glen, and J. B. O. Mitchell, "Random forest models to predict aqueous solubility," *Journal of Chemical Information and Modeling*, vol. 47, no. 1, pp. 150–158, 2007, pMID: 17238260. [Online]. Available: https://doi.org/10.1021/ci060164k

[48] J. Chen, K. de Hoogh, J. Gulliver, B. Hoffmann, O. Hertel, M. Ketzel, G. Weinmayr, M. Bauwelinck, A. van Donkelaar, U. A. Hvidtfeldt, R. Atkinson, N. A. H. Janssen, R. V. Martin, E. Samoli, Z. J. Andersen,

B. M. Oftedal, M. Stafoggia, T. Bellander, M. Strak, K. Wolf, D. Vienneau, B. Brunekreef, and G. Hoek, "Development of europe-wide models for particle elemental composition using supervised linear regression and random forest," *Environmental Science & Technology*, vol. 54, no. 24, pp. 15 698–15 709, 2020, pMID: 33237771. [Online]. Available: https://doi.org/10.1021/acs.est.0c06595

[49] B. Singh, P. Sihag, and K. Singh, "Modelling of impact of water quality on infiltration rate of soil by random forest regression," *Modeling Earth Systems and Environment*, vol. 3, pp. 999–1004, 9 2017.

[50] L. Schmid, A. Gerharz, A. Groll, and M. Pauly, "Tree-based ensembles for multi-output regression: Comparing multivariate approaches with separate univariate ones," *Computational Statistics Data Analysis*, vol. 179, p. 107628, 3 2023.

[51] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[52] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. [Online]. Available: http://www.jstor.org/stable/2699986

[53] J. Cai, K. Xu, Y. Zhu, F. Hu, and L. Li, "Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest," *Applied Energy*, vol. 262, p. 114566, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261920300787

[54] L. Wei, Z. Yuan, Y. Zhong, L. Yang, X. Hu, and Y. Zhang, "An improved gradient boosting regression tree estimation model for soil heavy metal (arsenic) pollution monitoring using hyperspectral remote sensing," *Applied Sciences*, vol. 9, no. 9, 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/9/1943

[55] C. Persson, P. Bacher, T. Shiga, and H. Madsen, "Multi-site solar power forecasting using gradient boosted regression trees," *Solar Energy*, vol. 150, pp. 423–436, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038092X17303717

[56] "Gradient boost for regression explained." [Online]. Available: https://www.numpyninja.com/post/gradient-boost-for-regression-explained

[57] M.-X. Wang, D. Huang, G. Wang, and D.-Q. Li, "Ss-xgboost: A machine learning framework for predicting newmark sliding displacements of slopes," *Journal of Geotechnical and Geoenvironmental Engineering*, vol. 146, 9 2020.

[58] I. Yilmaz and O. Kaynar, "Multiple regression, ann (rbf, mlp) and anfis models for prediction of swell potential of clayey soils," *Expert systems with applications*, vol. 38, no. 5, pp. 5958–5966, 2011.

[59] N. K. Manaswi and N. K. Manaswi, "Regression to mlp in keras," *Deep Learning with Applications Using Python: Chatbots and Face, Object, and Speech Recognition With TensorFlow and Keras*, pp. 69–89, 2018.

[60] "Multilayer perceptron explained with a real-life example and python code: Sentiment analysis." [Online]. Available: https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analysis-cb408ee93141

[61] I. Colombet, A. Ruelland, G. Chatellier, F. Gueyffier, P. Degoulet, and M.-C. Jaulent, "Models to predict cardiovascular risk: comparison of cart, multilayer perceptron and logistic regression." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2000, p. 156.

[62] M. Taki, A. Rohani, F. Soheili-Fard, and A. Abdeshahi, "Assessment of energy consumption and modeling of output energy for wheat production by neural

network (mlp and rbf) and gaussian process regression (gpr) models," *Journal of Cleaner Production*, vol. 172, pp. 3028–3041, 2018.

[63] F. Murtagh, "Multilayer perceptrons for classification and regression," *Neurocomputing*, vol. 2, no. 5-6, pp. 183–197, 1991.

[64] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[65] T. Guo, J. Dong, H. Li, and Y. Gao, "Simple convolutional neural network on image classification," in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*. IEEE, 2017, pp. 721–724.

[66] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation*, vol. 29, no. 9, pp. 2352–2449, 2017.

[67] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical image classification with convolutional neural network," in *2014 13th international conference on control automation robotics & vision (ICARCV)*. IEEE, 2014, pp. 844–848.

[68] F. Sultana, A. Sufian, and P. Dutta, "Advancements in image classification using convolutional neural network," in *2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*. IEEE, 2018, pp. 122–129.

[69] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 558–567.

[70] "Introduction to densenets (dense cnn)." [Online]. Available: https://www.analyticsvidhya.com/blog/2022/03/introduction-to-densenets-dense-cnn/

[71] Z. Yu and H. Gao, "Molecular representation learning via heterogeneous motif graph neural networks," 2 2022.

[72] O. Wieder, S. Kohlbacher, M. Kuenemann, A. Garon, P. Ducrot, T. Seidel, and T. Langer, "A compact review of molecular property prediction with graph neural networks," *Drug Discovery Today: Technologies*, vol. 37, pp. 1–12, 12 2020.

[73] J. Y. Choi, P. Zhang, K. Mehta, A. Blanchard, and M. L. Pasini, "Scalable training of graph convolutional neural networks for fast and accurate predictions of homo-lumo gap in molecules," *Journal of Cheminformatics*, vol. 14, p. 70, 10 2022.

[74] "Predicting molecular properties." [Online]. Available: https://www.kaggle.com/c/champs-scalar-coupling

[75] V. Fung, J. Zhang, E. Juarez, and B. G. Sumpter, "Benchmarking graph neural networks for materials chemistry," *npj Computational Materials*, vol. 7, no. 1, p. 84, 2021.

[76] C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, and K. F. Jensen, "A graph-convolutional neural network model for the prediction of chemical reactivity," *Chemical science*, vol. 10, no. 2, pp. 370–377, 2019.

[77] Z. Yang, M. Chakraborty, and A. D. White, "Predicting chemical shifts with graph neural networks," *Chemical science*, vol. 12, no. 32, pp. 10 802–10 809, 2021.

[78] J. Xiong, Z. Xiong, K. Chen, H. Jiang, and M. Zheng, "Graph neural networks for automated de novo drug design," *Drug Discovery Today*, vol. 26, no. 6, pp. 1382–1393, 2021.

[79] M. Sakai, K. Nagayasu, N. Shibui, C. Andoh, K. Takayama, H. Shirakawa, and S. Kaneko, "Prediction of pharmacological activities from chemical structures with graph convolutional neural networks," *Scientific reports*, vol. 11, no. 1, p. 525, 2021.

[80] "An overview of encoder transformers — part 1." [Online]. Available: https://medium.com/geekculture/an-overview-of-encoder-transformers-part-1-78524ca5a784

[81] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu, "On layer normalization in the transformer architecture," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 524–10 533.

[82] C. Zhang, Y. He, B. Du, L. Yuan, B. Li, and S. Jiang, "Transformer fault diagnosis method using iot based monitoring system and ensemble machine learning," *Future generation computer systems*, vol. 108, pp. 533–545, 2020.

[83] D. Saravanan, A. Hasan, A. Singh, H. Mansoor, and R. N. Shaw, "Fault prediction of transformer using machine learning and dga," in *2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON)*. IEEE, 2020, pp. 1–5.

[84] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *International conference on machine learning*. PMLR, 2018, pp. 4055–4064.

[85] P. Karpov, G. Godin, and I. V. Tetko, "A transformer model for retrosynthesis," in *Artificial Neural Networks and Machine Learning–ICANN 2019: Work-*

shop and Special Sessions: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings. Springer, 2019, pp. 817–830.

[86] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 10 2020.

[87] N. Ahmed, R. Ahammed, M. M. Islam, M. A. Uddin, A. Akhter, M. A.-A. Talukder, and B. K. Paul, "Machine learning based diabetes prediction and development of smart web application," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 229–241, 2021.

[88] O. Taylor, P. Ezekiel, and F. Deedam-Okuchaba, "A model to detect heart disease using machine learning algorithm," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 11, pp. 1–5, 2019.

[89] L. Skultety, P. Frycak, C. Qiu, J. Smuts, L. Shear-Laude, K. Lemr, J. X. Mao, P. Kroll, K. A. Schug, A. Szewczak *et al.*, "Resolution of isomeric new designer stimulants using gas chromatography–vacuum ultraviolet spectroscopy and theoretical computations," *Analytica Chimica Acta*, vol. 971, pp. 55–67, 2017.

[90] J. X. Mao, P. Kroll, and K. A. Schug, "Vacuum ultraviolet absorbance of alkanes: an experimental and theoretical investigation," *Structural Chemistry*, vol. 30, pp. 2217–2224, 2019.

[91] J. X. Mao, P. Walsh, P. Kroll, and K. A. Schug, "Simulation of vacuum ultraviolet absorption spectra: paraffin, isoparaffin, olefin, naphthene, and aromatic hydrocarbon class compounds," *Applied Spectroscopy*, vol. 74, no. 1, pp. 72–80, 2020.

[92] J. Schenk, J. X. Mao, J. Smuts, P. Walsh, P. Kroll, and K. A. Schug, "Analysis and deconvolution of dimethylnaphthalene isomers using gas chromatography vacuum ultraviolet spectroscopy and theoretical computations," *Analytica chimica acta*, vol. 945, pp. 1–8, 2016.

[93] T. T. Ponduru, C. Qiu, J. X. Mao, A. Leghissa, J. Smuts, K. A. Schug, and H. R. Dias, "Copper (i)-based oxidation of polycyclic aromatic hydrocarbons and product elucidation using vacuum ultraviolet spectroscopy and theoretical spectral calculations," *New Journal of Chemistry*, vol. 42, no. 24, pp. 19 442–19 449, 2018.

[94] Y.-C. Lo, S. E. Rensi, W. Torng, and R. B. Altman, "Machine learning in chemoinformatics and drug discovery," *Drug Discovery Today*, vol. 23, pp. 1538–1546, 8 2018.

[95] M. Thiele, E. Gross, and S. Kümmel, "Adiabatic approximation in nonperturbative time-dependent density-functional theory," *Physical review letters*, vol. 100, no. 15, p. 153004, 2008.

[96] M. E. Casida, "Time-dependent density-functional theory for molecules and molecular solids," *Journal of Molecular Structure: THEOCHEM*, vol. 914, no. 1-3, pp. 3–18, 2009.

[97] Z.-L. Cai, K. Sendt, and J. R. Reimers, "Failure of density-functional theory and time-dependent density-functional theory for large extended π systems," *The Journal of chemical physics*, vol. 117, no. 12, pp. 5543–5549, 2002.

[98] "Spectra prediction, uv advisor." [Online]. Available: https://spectra.collaborationspharma.com/

[99] "Vuv analytics company." [Online]. Available: https://vuvanalytics.com/

[100] N. M. O'Boyle and R. A. Sayle, "Comparing structural fingerprints using a literature-based similarity benchmark," *Journal of Cheminformatics*, vol. 8, p. 36, 12 2016.

[101] R. E. Carhart, D. H. Smith, and R. Venkataraghavan, "Atom pairs as molecular features in structure-activity studies: definition and applications," *Journal of Chemical Information and Computer Sciences*, vol. 25, pp. 64–73, 5 1985.

[102] P. Skoda and D. Hoksza, "Exploration of topological torsion fingerprints." IEEE, 11 2015, pp. 822–828.

[103] R. Nilakantan, N. Bauman, J. S. Dixon, and R. Venkataraghavan, "Topological torsion: a new molecular descriptor for sar applications. comparison with other descriptors," *Journal of Chemical Information and Computer Sciences*, vol. 27, pp. 82–85, 5 1987.

[104] H. L. Morgan, "The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service." *Journal of chemical documentation*, vol. 5, no. 2, pp. 107–113, 1965.

[105] L. H. Hall and L. B. Kier, "Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information," *Journal of Chemical Information and Computer Sciences*, vol. 35, pp. 1039–1045, 11 1995.

[106] Z. Benkhedda, P. Landais, J. Dereppe, J. Kister, and M. Monthioux, "Spectroscopic characterization of aromatic fractions from maturation series of coal," in *1991 International Conference on Coal Science Proceedings*, International Energy Agency Coal Research Ltd, Ed. Butterworth-Heinemann, 1991, pp. 56–59. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780750603874500176

[107] A. T. Balaban, D. C. Oniciu, and A. R. Katritzky, "Aromaticity as a cornerstone of heterocyclic chemistry," *Chemical reviews*, vol. 104, no. 5, pp. 2777–2812, 2004.

[108] H. Mu, L. Pan, D. Song, and Y. Li, "Neutral nickel catalysts for olefin homo- and copolymerization: relationships between catalyst structures and catalytic properties," *Chemical Reviews*, vol. 115, no. 22, pp. 12 091–12 137, 2015.

[109] T. Kinnibrugh, S. Bhattacharjee, P. Sullivan, C. Isborn, B. Robinson, and B. Eichinger, "Influence of isomerization on nonlinear optical properties of molecules," *The Journal of Physical Chemistry B*, vol. 110, no. 27, pp. 13 512–13 522, 2006.

[110] Z. Xu, Z. Yang, Y. Liu, Y. Lu, K. Chen, and W. Zhu, "Halogen bond: its role beyond drug–target binding affinity for drug discovery and development," *Journal of chemical information and modeling*, vol. 54, no. 1, pp. 69–78, 2014.

[111] RDKit, "Open-source cheminformatics." [Online]. Available: https://www.rdkit.org/

[112] Pytorch, "Open-source code for pytorch." [Online]. Available: https://github.com/pytorch/pytorch

[113] , "scikit-learn machine learning in python." [Online]. Available: https://scikit-learn.org/stable/

[114] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[115] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[116] R. Kojima, S. Ishida, M. Ohta, H. Iwata, T. Honma, and Y. Okuno, "kgcn: a graph-based deep learning framework for chemical structures," *Journal of Cheminformatics*, vol. 12, p. 32, 12 2020.

[117] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," 10 2017.

[118] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," 12 2018.

[119] R. D. King, O. I. Orhobor, and C. C. Taylor, "Cross-validation is safe to use," *Nature Machine Intelligence*, vol. 3, no. 4, pp. 276–276, 2021.

[120] xyz2graph, "Open-source to plot xyz file." [Online]. Available: https://github.com/zotko/xyz2graph

[121] L. Wu, Y. Chen, K. Shen, X. Guo, H. Gao, S. Li, J. Pei, B. Long *et al.*, "Graph neural networks for natural language processing: A survey," *Foundations and Trends® in Machine Learning*, vol. 16, no. 2, pp. 119–328, 2023.

[122] T. Gui, Y. Zou, Q. Zhang, M. Peng, J. Fu, Z. Wei, and X.-J. Huang, "A lexicon-based graph neural network for chinese ner," in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 2019, pp. 1040–1050.

[123] D. Yao, Z. Zhi-li, Z. Xiao-feng, C. Wei, H. Fang, C. Yao-ming, and W.-W. Cai, "Deep hybrid: multi-graph neural network collaboration for hyperspectral image classification," *Defence Technology*, vol. 23, pp. 164–176, 2023.

[124] M. Adnan, S. Kalra, and H. R. Tizhoosh, "Representation learning of histopathology images using graph neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 988–989.

[125] S. Ishida, T. Miyazaki, Y. Sugaya, and S. Omachi, "Graph neural networks with multiple feature extraction paths for chemical property estimation," *Molecules*, vol. 26, no. 11, p. 3125, 2021.

[126] Y. Hou, S. Wang, B. Bai, H. C. S. Chan, and S. Yuan, "Accurate physical property predictions via deep learning," *Molecules*, vol. 27, p. 1668, 3 2022.

[127] T. Kasanishi, X. Wang, and T. Yamasaki, "Edge-level explanations for graph neural networks by extending explainability methods for convolutional neural networks," in *2021 IEEE International Symposium on Multimedia (ISM)*. Los Alamitos, CA, USA: IEEE Computer Society, dec 2021, pp. 249–252. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ISM52913.2021.00049

[128] D. P. Kingma, M. Welling *et al.*, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[129] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[130] H. Shao, S. Yao, D. Sun, A. Zhang, S. Liu, D. Liu, J. Wang, and T. Abdelzaher, "Controlvae: Controllable variational autoencoder," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8655–8664.

[131] M. J. Kusner, B. Paige, and J. M. Hernández-Lobato, "Grammar variational autoencoder," in *International conference on machine learning*. PMLR, 2017, pp. 1945–1954.

[132] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin, "Variational autoencoder for deep learning of images, labels and captions," *Advances in neural information processing systems*, vol. 29, 2016.

[133] R. P. Sheridan and B. P. Feuston, "Comparison of topological, shape, and docking methods in virtual screening," *Journal of chemical information and computer sciences*, vol. 36, no. 4, pp. 682–694, 1996.

[134] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. S. Pande, "Applications and advancements of machine learning in structure-based drug discovery: methods and examples," *Journal of chemical information and modeling*, vol. 58, no. 3, pp. 498–511, 2018.

[135] A. Heifetz, J. Barker, and P. Banker, "Generating qsar models using atom pairs: application to aqueous solubility prediction," *Journal of chemical information and computer sciences*, vol. 35, no. 4, pp. 717–723, 1995.

[136] R. S. Bohacek and C. McMartin, "Torsion angle dependence of molecular shape descriptors and their application in qsar studies," *Journal of chemical information and computer sciences*, vol. 36, no. 4, pp. 870–878, 1996.

[137] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Supervised learning in quantitative structure-activity relationship studies," *Nature biotechnology*, vol. 27, no. 10, pp. 951–956, 2008.

[138] C. Hansch, A. Leo, S. Unger, K. Kim, D. Nikaitani, and E. Lien, "Structure-activity relationships in substituted benzenes," *Journal of medicinal chemistry*, vol. 7, no. 1, pp. 71–79, 1964.

[139] R. Montanari, E. Gancia, L. Pignatti, and R. Silvestri, "Prediction of the anti-hiv activity of a large series of 2, 3-diaryl-1, 3-thiazolidin-4-ones: a comparison between e-state and hansch approaches," *Journal of medicinal chemistry*, vol. 42, no. 3, pp. 386–391, 1999.

[140] A. Bender and R. C. Glen, "Bayesian de novo design: retrosynthetic analysis using predicted product distributions," *Journal of chemical information and modeling*, vol. 45, no. 3, pp. 700–709, 2005.

[141] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of chemical information and modeling*, vol. 50, no. 5, pp. 742–754, 2010.