Chapter – 1

INTRODUCTION

1. INTRODUCTION

Drug discovery involves the identification and creation of novel medications to address and prevent diseases. Drug discovery playing a crucial role in impacting both human health and society. In the Drug discovery mainly considerable properties are solubility, metabolism and toxicity. In that Solubility is a main factor influencing drug related researches. Solubility holds significant importance in drug discovery across various aspects. It plays a crucial role in influencing the bioavailability, synthesis, and manufacturing processes of drugs, impacting different stages of drug design. Chemists aim to enhance the solubility of molecules by optimizing their molecular structures during the drug design phase. Once a drug-like compound exhibits satisfactory properties , it becomes a candidate for further development into a new medication. The solubility of a drug significantly affects its absorption into the body, making it a key factor in this aspect of drug[1]. So that in Drug discovery , Solubility plays a vital role. We need to find the solubility of the each molecule or chemical compound . But in traditional way , time - consuming and expensive. Traditional analytical methods are insufficient for handling extensive datasets, therefore it is necessary to processing and converting such data into valuable knowledge[2]. We can achieve this by using Machine Learning techniques.

A machine learning (ML) algorithm capable of precisely characterizing the compositions of behavioral components can meet this requirement. By employing ML techniques, it becomes possible to assess a considerable number of materials without the need for physical samples and to efficiently ascertain their physical properties, like solubility. Machine Learning Techniques such as Random forest , Multi linear regression and some other regression model were used previously. But the main obstacle is the final output RMSE (root mean square error) is 0.7 to 1.0. By using ML approaches the error is more . At this difficulties we can use Graph Neural Network(GNN) which is mainly for the Graph Learning. A Graph Neural Network (GNN) is a neural network specifically crafted for processing graph-structured data, where graphs are composed of nodes representing entities and edges signifying relationships between these entities. In the realm of GNNs, the network is trained to execute tasks on such graph-structured data by incorporating information from adjacent nodes during the learning process. When we are dealing with Graphs is Drug related researches then definetly we come across a word called SMILES(Simplified Molecular

Input Line Entry System). SMILES serves as a prevalent notation to depict the structure of chemical molecules through ASCII strings. Within a SMILES string, individual characters symbolize particular atoms or bonds, enabling a textual portrayal of molecular structures.Here SMILES are not directly the inputs to the Graph Neural network ,It contains some approaches to convert the SMILES into the Molecular Graphs or Molecular Descriptors these are the input to the Graph neural network. At this context we can use RDKit and Featurization. RDKit allows the representation of chemical molecules in different formats. It can read and write files in various chemical structure formats, providing the exchange of molecular data. When it comes to Featurization , In the field of cheminformatics and machine learning, featurization refers to the extraction of meaningful features or descriptors from chemical data, often represented as molecular structures. This process involves capturing relevant information to enable effective analysis and modelling.Before the Graph neural network there is Quantitative structure property relationships (QSPR) and Quantitative structure activity relationships(QSAR) were presented for finding the activity and property of the SMILES.Here we were used Quantitative structure property relationships (QSPR) which is same as QSAR but this approach mainly focusing on the property of the chemical compound ,our aim is to predicting the Solubility. Therefore we were used Quantitative structure property relationships (QSPR).

We are using Graph Convolution Neural network for the solubility prediction(ESOL).In GCNN it takes the input from after the featurization of SMILES. Our input is molecular graph. Graph Convolutional Neural Networks (GCNs) work by adapting traditional convolutional neural network principles to operate on graph-structured data. Here we also used the Grid Search Hyperparameter tuning, it is a method for tuning hyperparameters in deep learning, where a systematic exploration is conducted across a predetermined grid of hyperparameter values to identify the most effective combination for a model. It help us to optimize the Root mean square error (RMSE).

1.1 Quantitative Structure-Property Relationship (QSPR):

Quantitative Structure-Property Relationship (QSPR) in the context of water solubility refers to the development of predictive models that quantitatively relate the molecular structure of chemical compounds to their solubility in water. Water solubility is a critical physicochemical property that influences various aspects of chemistry, pharmacology, and environmental science. QSPR models for water solubility aim to establish a

mathematical relationship between molecular descriptors (quantitative representations of molecular structure) and the experimentally determined solubility values of compounds in water. Quantitative Structure-Property Relationship (QSPR) models enhanced by these

Graph Neural Networks (GNNs) represent a cutting-edge approach to understanding and predicting molecular properties. Traditional QSPR models rely on molecular descriptors to encode structural information, but GNNs take advantage of the inherent graph-like nature of molecular structures. In this paradigm, each atom becomes a node, and bonds between atoms become edges, forming a molecular graph. GNNs operate by iteratively aggregating information from neighbouring atoms, allowing them to capture complex spatial relationships within the molecular structure. This enables the model to learn intricate patterns and dependencies that may be challenging for traditional descriptors to represent explicitly. The integration of GNNs into QSPR leverages the strengths of deep learning in handling graph-structured data, providing a more flexible and data-driven approach to understanding the complex relationships between molecular structures and properties. The GNN-enhanced QSPR models demonstrate remarkable efficacy in predicting properties such as solubility, toxicity, or bioactivity, showcasing their potential to advance the field of computational chemistry and cheminformatics.

1.2 QSAR - Quantitative Structure-Activity Relationship:-

Quantitative Structure-Activity Relationship (QSAR) is a computational modeling technique used in cheminformatics and drug discovery to predict the biological activity or other properties of chemical compounds based on their molecular structure.The main aim of this QSAR is to establish a mathematical relationship between the structural properties(Molecular weight,size,shape , hydrophilicity , electronegativity) and their observed biological activities (Absorptioon , distribution , metabolism).

QSAR analysis begins with converting SMILES representations of chemical compounds into molecular structures. SMILES (Simplified Molecular Input Line Entry System) can serve as a crucial component representing the chemical structures of compounds. SMILES is a text-based notation for describing the structure of chemical molecules using ASCII characters. This conversion can be done using software libraries or tools (RDkit) capable of interpreting SMILES strings and generating 2D or 3D

representations of the molecules.In QSAR modeling, various molecular descriptors are calculated from the chemical structure of compounds using tools like RDkit .These molecular descriptors are numerical or categorical representations of chemical compounds that capture various physicochemical properties , structural features and molecular characteristics .These descriptors are derived from the chemical structure of the molecule and serve as input

features for quantitative structure-activity relationship(QSAR) modeling. Once the molecular descriptors are calculated, statistical or machine learning models are trained using a dataset of compounds with known biological activity or property values. The model learns the relationship between the descriptors and the target property, allowing it to make predictions for new compounds with unknown activity.QSAR modeling is essential for drug discovery, but it has many constraints.Ensemble-based machine learning approaches have been used to overcome constraints and obtain reliable predictions.Ensemble-based machine learning is a powerful technique that involves combining multiple individual models to produce a stronger predictive model. The fundamental idea behind ensemble methods is to leverage the diversity among the base models to improve overall predictive performance, robustness, and generalization ability. The Delaney dataset, also known as the ESOL (Estimated Solubility) dataset, is a widely used benchmark dataset in QSAR modeling. It contains chemical structures of organic molecules along with their experimentally measured water solubility values. QSAR models are built using this dataset by first calculating molecular descriptors (features) from the chemical structures, such as topological, geometric, constitutional, or physicochemical properties. These descriptors serve as input features for the model.

1.3 Graph Neural Network:

Graph Neural Networks (GNNs) have emerged as a powerful tool for analysing structured data, especially in domains where relationships between entities are key. In recent years, the application of GNNs has extended to various fields such as social network analysis, recommendation systems, drug discovery, and material science. One particularly promising application of GNNs lies in predicting molecular properties, including solubility. The solubility of a compound is a crucial factor in drug development, material design, and various industrial processes. Traditional methods for predicting solubility often rely on handcrafted features and lack the ability to capture

complex molecular interactions effectively. Graph Neural Networks offer a data-driven approach to learn representations directly from molecular graphs. By treating atoms as nodes and chemical bonds as edges in a graph representation, GNNs can effectively capture the structural information of molecules and their interatomic relationships. This allows for the development of more accurate and versatile models for predicting solubility. Traditional methods for predicting solubility often rely on handcrafted features and lack the ability to

capture complex molecular interactions effectively. Graph Neural Networks offer a data-driven approach to learn representations directly from molecular graphs

1.3.1 Key Components of GNNs:

- **Node Embeddings:** GNNs start by representing each node in the graph with a feature vector, often called a node embedding. These embeddings encode information about the node and its neighborhood.
- **Message Passing:** GNNs operate through a process known as message passing. In each layer, a node aggregates information from its neighbors, updating its own embedding. This is typically done through a weighted combination of neighbor embeddings.
- **Aggregation Functions:** Various aggregation functions can be used for message passing, such as mean pooling, sum aggregation, or more sophisticated techniques like graph convolution. The choice of aggregation function influences how information is propagated through the graph.
- **Layer Stacking:** GNNs typically consist of multiple layers, and information is passed through each layer. This allows the model to capture increasingly complex relationships as it goes deeper into the network.
- **Output Generation**: The final layer of a GNN generates the output. Depending on the task, this could be classification, regression, or any other relevant prediction.

1.1.2. Applications of GNNs GNNs have found applications in various fields:

- **Social Networks:** They can be used for link prediction, community detection, and recommendation systems.

- **Biology:** GNNs are vital for predicting protein-protein interactions, drug discovery, and metabolic pathway analysis.
- **Recommendation Systems:** GNNs can provide more personalized and accurate recommendations in e-commerce and content recommendation platforms.
- **Natural Language Processing:** GNNs can be employed to analyze semantic relationships in text data, such as co-authorship networks in academic literature. GNNs typically consist of multiple layers, and information is passed through each layer. This allows the model to capture increasingly complex relationships as it goes deeper into the network.

1.4 Dataset Overview:

The Delaney dataset, a cornerstone in cheminformatics, comprises a diverse collection of organic compounds paired with their experimentally measured aqueous solubility values. Widely used for developing and evaluating Quantitative Structure-Property Relationship (QSPR) models, this dataset offers insights into the relationship between molecular structure and solubility—a critical parameter influencing drug absorption, distribution, metabolism, and excretion. With its varied chemical structures and standardized solubility measurements, the Delaney dataset serves as a benchmark for assessing the accuracy and generalization ability of predictive models in drug discovery, environmental chemistry, and beyond, enabling advancements in computational chemistry and rational drug design. The attributes of the dataset follows:

1.4.1 Attributes of Dataset:

- **Compound ID:** This attribute represents the identifier or name assigned to each chemical compound in the dataset.
- **ESOL Predicted Log Solubility in mols per litre:** This attribute represents the estimated logarithm of the octanol-water partition coefficient (log P) for each chemical compound. Log P is a measure of a compound's hydrophobicity or lipophilicity and can be related to the compound's solubility in water. A positive log P value suggests that the compound is more soluble in lipids (hydrophobic). A negative log P value suggests that the compound is more soluble in water (hydrophilic).
- **Minimum Degree:** This attribute represents the minimum degree of connectivity for each chemical compound. It is a numerical variable that

describes the minimum number of bonds a single atom in the compound forms with other atoms. The degree of a vertex in a molecular graph corresponds to the number of bonds it forms.

- **Molecular Weight:** This attribute represents the molecular weight of each chemical compound. It is a numerical variable that quantities the mass of a molecule. The molecular weight is the sum of the atomic weights of all atoms in a molecule. It is often measured in atomic mass units (amu) or g/mol.
- **Number of H-Bond Donors:** This attribute represents the count of hydrogen bond donors in each chemical compound. It is a numerical variable indicating how many hydrogen atoms in a compound can act as hydrogen bond donors.

  Hydrogen bond donors are atoms with hydrogen atoms directly bonded to more electronegative atoms (e.g., nitrogen or oxygen) capable of forming hydrogen bonds.
- **Number of Rings:** This attribute represents the count of rings in the molecular structure of each chemical compound. It is a numerical variable indicating how many ring structures are present in a compound.
- **Number of Rotatable Bonds**: This attribute represents the count of rotatable bonds in each chemical compound. It is a numerical variable indicating the number of bonds in a compound that allow for free rotation around them. Rotatable bonds contribute to the edibility and conformational freedom of a molecule.
- **Polar Surface Area**: This attribute represents the polar surface area of each chemical compound. It is a numerical variable indicating the surface area of a compound that is polar in nature. Polar surface area is often associated with the ability of a molecule to form hydrogen bonds and interact with polar solvents.
- **Measured Log Solubility in mols per litre**: This attribute represents the experimentally measured logarithm of the solubility of each chemical compound in mols per litre. It is a numerical variable that provides actual measured data on the solubility of the compounds. The values are determined through laboratory experiments.
- **SMILES (Simplified Molecular Input Line Entry System):** This attribute represents the chemical structure of each compound using the SMILES

notation. SMILES is a text-based notation that allows a user to represent a chemical structure in a way that can be used by a computer. It is a string or sequence of characters that encodes the connectivity of atoms in a molecule.

- **Target Attribute:**

  ESOL Predicted Log Solubility in mols per litre:

  This attribute represents the estimated logarithm of the octanol-water partition coefficient (log P) for each chemical compound. Log P is a measure of a compound's hydrophobicity or lipophilicity and can be related to the compound's solubility in water. A positive log P value suggests that the compound is more soluble in lipids (hydrophobic).

- **Input Attribute:**

  SMILES (Simplified Molecular Input Line Entry System):

  This attribute represents the chemical structure of each compound using the SMILES notation. This attribute represents the chemical structure of each compound using the SMILES notation. SMILES is a text-based notation that allows a user to represent a chemical structure in a way that can be used by a computer. It is a string or sequence of characters that encodes the connectivity of atoms in a molecule.

Chapter – 2

PROFILE OF THE COMPANY

2.PROFILE OF THE COMPANY

Andhra University was constituted in the year 1926 by the Madras Act of 1926. Andhra University is approved by UGC. It was ranked 71 in India overall by the National Institutional Ranking Framework in 2022 and 36th among universities. Since its establishment in 1926, Andhra University follows its noble vision and mission as inscribed in the logo, “Thejasvina Vadhitamastu ", which means," May the Divine Light Illuminate Our Studies”. Ever since its inception in 1926 Andhra University has had an impeccable record of catering to educational needs and addressing the sociological problems of the region. The University is relentless in its efforts in maintaining standards in teaching and research, ensuring proper character building and development among the students, encouraging community developmental programs, nurturing leadership and patriotism in young men and women and imbibing a sense of responsibility to become good citizens of the country. Vision of the Company “Create New Frontiers of Knowledge in Quest for Development of the Humane and Just Society”. The vision of the university is inextricably linked to its teaching, learning, research, consultancy, processes, industrial and societal interactions and community outreach activities. Mission of the Company “To undertake quality related research studies, consultancy and training programs”. The mission of Andhra University is to leverage global knowledge networks to help India and International Community in developing human resources capable of leading creative developments by upholding intellectual traditions and human values.

2.1 Criteria wise Summary

- Curricular Aspects
- Teaching-learning and Evaluation
- Research, Innovations and Extension
- Infrastructure and Learning Resources
- Student Support and Progression
- Governance, Leadership and Management
- Institutional Values and Best Practices

Curricular Aspects

Andhra University offers 178 (UG, Integrated, PG and Ph.D.) programs through its 58

departments apart from School of Distance Education programs. Outcome based education (OBE) is at the heart of all programs in the University. Research methodology and techno-entrepreneurship are incorporated in the curriculum to imbibe innovation and entrepreneurship culture among students. The University has taken up the measures to include field projects/research projects/internships during their study as part of the curriculum. All the programs have project work/internship/field project in the curriculum. By adopting NEP – 2020, promoting skill-based courses, outcomebased education and analysis of feedback, the curriculum of AU has set a benchmark for many higher education offering institutes.

Teaching-learning and Evaluation

The admission policy is strictly according to the Constitutional provision for reservations for all the programs and admissions are done through a centralized online counselling process based on the common entrance examination conducted at State Level by regulatory body. Most of the PG and Ph.D. programs have project work/internship/practical training/Field/Society benefitted projects with industry and R&D Labs exposure. AU has a good student- teacher ratio of 18:1 which ensures an effective teaching-learning process. The mentor-mentee system that has been in practice in the institution since 2015 has benefited the students enormously. IT Integrated automatic examination management system is in place and the publication of end semester results are done on an average of 21 days. To maintain transparency in the evaluation process, re-evaluation is made available to the students, and they are also encouraged to get Digi Locker facility.

Research, Innovations and Extension

AU has a well-defined policy for the promotion of research, innovation and entrepreneurship ecosystem. Most of the departments are recognized by funding agencies such as UGC-SAP, DST-FIST, DST- SEED, DBT. The Institution provides seed money to the faculty to strengthen basic research. A well-defined research promotion policy is effectively implemented. Around 16 (sixteen) Industry Endowed professor chairs including 2 international sponsored chairs have been created during the assessment period. The campus has sixteen research centers such as AERC, Center for studies in Bay of Bengal, PRC. Workshops/seminars are organized on research methodology, intellectual copyrights, entrepreneurships, and skill development. Around 89 faculty members have received awards for their contributions towards research during the assessment period. The University has implemented plagiarism software

check and developed code of ethics for research which is monitored by research advisory and other

institutional ethics committees. 66 faculty members have received national and international awards, and the university has in turn felicitated them with incentives. University Published 81 patents out of which 14 are awarded, 3 are Copyrights, 1 Technology transfer and 67 patents are in publication stage.

Infrastructure and Learning Resources

Andhra University made substantial investment in creating infrastructure and Learning Resources to promote academic excellence and research. The university is covered with flora and fauna, over 425 acres of land located in the heart of Visakhapatnam city. The university is ecofriendly and has a great ambience having state of art infrastructure for teaching, learning, cultural activities, sports, yoga and other facilities. The University Library possesses a rich collection of e-books, e-journals, and physical editions of books and journals. Efficient transport system, safe drinking water RO plants, sports facilities, strong energy and water management systems, differently abled and genderequity ecosystem, startup and incubation centers, adequate hostel facilities for students including international students, multi-cuisine canteens, healthcare and hospital facilities, vigilant security systems make the campus versatile.

Student Support and Progression

Andhra University has the unique distinction of creating an optimum student-friendly ambience on campus: scholarships and free ships to majority of PG students, fellowships to PhD/PDFs, higher rate of employment, transparent and efficient grievance redressal mechanism, due student's representation in academic bodies, vibrant alumni associations and wide avenues for co-curricular activities. The student's grievance redressal cell, Internal Complaints Cell, SC/ST Complaint Cell and AntiRagging Committee/Squad, function efficiently on the University. A participative democratic academic ambience is ensured through students' representation in many statutory/non-statutory bodies including BOS, Grievance Committee, Hostel Monitoring Committee, Hostels' Committee and University has three commendable student unions namely united Student club, GITA, KALAW, provide platforms for academic/artistic/sports endeavors, career guidance, Canteen Advisory Committee.

Governance, Leadership and Management

The University management believes in decentralization and participative management

and leadership. It takes policy decisions through its statutory bodies viz., Academic Council for academic matters, Finance Committee for fund management, and Building

Committee for infrastructure development. The Executive Council is the apex body of the

University to consider and approve the decisions taken by other statutory bodies and act on policy matters. The leadership team consists of the Vice-Chancellor and Rector, and the Registrar. All academic matters, such as conduct of entrance examinations, semester exams, and results declaration are handled by the Office of the Controller of Examinations in close cooperation with all the departments and colleges.

Institutional Values and Best Practices

The AU campus has a conducive environment for gender equity which is reflected in the composition of students and staff. There is no gender discrimination and equal opportunities are given to men and women. The University has also taken steps to ensure the welfare of a transgender student who took admission recently. The University has established the Durgabai Deshmukh center for women's studies and they regularly conduct gender sensitization programs to promote cooperation between male and female students. Yet another best practice in Andhra University is further strengthening of "AU Holistic wellness ecosystem", by way of integrating Sports, Yoga &Meditation, and Mental Health Counseling centers, as a unique congruent Indian Knowledge Initiative on campus, that facilitates further promotion of physical, social, mental and spiritual wellness of youth for holistic personality development.

Chapter - 3

PROBLEM DEFENITION AND OBJECTIVES OF PROJECT

3.PROBLEM DEFENITION AND TASKS TAKEN UP

3.1 PROBLEM DEFINITION

- Drug discovery has to be accelerated with modern computational methods that play an important role in biomedicine. Drug design is a time taking process. Especially, target identification and target validation are both timing consuming processes.
- The solubility of the drug is to be known from its molecular properties. For this purpose, we are going to use Delaney dataset which provide ESOL (Estimated Solubility value) which give the numeric value of the solubility .
- The proposed model in the future scope GNN in which SMILES are converted to molecular graphs as model input and using Sequence based learning and Natural Language Processing can be used for checking the solubility of the drugs.

3.2 OBJECTIVES OF PROJECT

In the field of drug discovery and development, there exists a pressing need for efficient and accurate computational tools to expedite the process of drug target identification, interaction prediction, toxicity assessment, and water solubility prediction for chemical compounds. This problem statement outlines the development of a comprehensive computational AI-based approach to address these critical challenges in pharmaceutical research.

3.2.1. Drug Target Identification and Interaction Prediction:

- **Challenge:** The identification of suitable drug targets and the prediction of how chemical compounds interact with these targets are pivotal stages in drug development. Traditional methods are time-consuming and expensive.
- **Objective:** Develop an AI-based system that can predict potential drug targets and interactions between compounds and target proteins, leveraging data-driven approaches, structural analysis, and machine learning models.
- **Benefits:** Accelerate drug discovery by reducing the need for extensive experimental screening, saving time and resources while increasing the success rate of drug development.

3.2.2. Water Solubility Prediction:

- **Challenge:** Water solubility is a critical factor in a drug's effectiveness, and accurately predicting it for diverse compounds is a complex task.
- **Objective:** Develop an AI-based model that can predict the water solubility of chemical compounds based on their structural features and physicochemical properties.
- **Benefits:** Streamline the selection of compounds with favourable solubility profiles, leading to improved drug formulation and bioavailability.

Chapter – 4

LITERATURE SURVEY

4.LITERATURE SURVEY

4.1 ”Improved Lipophilicity and Aqueous Solubility Prediction with Composite Graph Neural Networks” Oliver Wieder 1,* , Mélaine Kuenemann 2 Sharon D. Bryant 3 and Thierry Langer 1 1 , Marcus Wieder 1 , Thomas Seidel 1 , Christophe Meyer 2.

This paper introduces a new type of graph-based neural network called D-GIN, which combines two sub-architectures to improve accuracy in predicting molecular properties important for drug discovery. By addressing limitations in current assessment methods for deep-learning models, the authors argue that combining different models can make graph neural networks more powerful and accurate.

The paper "Improved Lipophilicity and Aqueous Solubility Prediction with Composite Graph Neural Networks" uses a combined dataset for training and evaluation.The paper compares the D-GIN architecture with various baseline models and its individual sub-architectures (D-MPNN and GIN) on two main tasks: predicting lipophilicity (logD and logP) and aqueous solubility (logS).Overall, the results demonstrate that the D-GIN architecture offers a valuable tool for predicting lipophilicity and aqueous solubility with improved accuracy and generalizability compared to existing methods.

4.2.“Molecular Descriptors Property Prediction Using Transformer-Based Approach” Tuan Tran and ChinweEkenna.

This paper introduces a machine learning model that can predict molecular properties using a two-stage approach involving pre-training and fine-tuning. The model uses both labeled and unlabeled data represented as SMILES strings for molecules, and shows promising results in predicting anti-malaria drug candidates.

the authors utilize the MoleculeNet benchmark suite for evaluation. This suite encompasses several diverse datasets focused on various tasks related to molecular properties prediction.

It's crucial to note that MoleculeNet offers a standardized platform for evaluating performance and comparing different models across various tasks related to molecular properties. Using this benchmark allows for more reliable conclusions and facilitates comparisons with other approaches in the field.

4.3 Attention-Based Graph Neural Network for Molecular Solubility Prediction"
Waqar Ahmad, Hilal Tayara, Kil To Chong

This paper discusses using deep learning models to predict the solubility of different molecules in drug development, which can help reduce the time and cost of experimental testing. The best-performing model was able to predict the solubility of anticancer compounds with good accuracy, and the study suggests that further improvements can be made by enhancing the graph algorithms or including more molecular properties.The paper "Attention-Based Graph Neural Network for Molecular Solubility Prediction" by Waqar Ahmad et al. (2023) utilizes several benchmark datasets for evaluating their model's performance

1. **Delaney Dataset:** This widely used dataset contains experimental logS (aqueous solubility logarithm) values for 1,128 organic molecules.
2. **ESOL Dataset:** Another popular benchmark encompassing around 110,000 diverse organic molecules with logS values.
3. Overall, the results of this paper support the effectiveness of the attention-based GNN model for molecular solubility prediction, demonstrating improved performance, generalizability, and interpretability compared to existing methods.

4.4.Physiological variables in machine learning QSARs allow for both cross-chemical and cross-species predictionsJochen P. Zubrod, Nika Galic, Maxime Vaugeois, David A. Dreier

This paper explores using machine learning models to predict how different species are affected by chemicals based on their physiological properties, specifically focusing on dynamic energy budget (DEB) parameters. The study found that these models were successful in predicting species-specific endpoints, highlighting the importance of understanding how physiological processes influence species sensitivity in ecological risk assessment without the need for species-specific testing.the authors utilize multiple datasets to evaluate their machine learning approach.The authors compared their model with alternative approaches using various statistical metrics and visualizations. They discussed the limitations of the study, such as potential data biases and the need for further validation on broader datasets. They also proposed future directions for research,

including exploring different physiological variables and integrating the approach with other data sources.Overall, the results of this paper strongly support the use of physiological variables in machine learning QSAR models for improved accuracy, interpretability, and broader applicability in cross-chemical and cross-species predictions.

4.5 "SolPredictor: Predicting Solubility with Residual Gated Graph Neural Network"Waqar Ahmad, Hilal Tayara, HyunJoo Shim, Kil To Chong

This paper discusses how computational methods, specifically machine and deep learning, are being used to predict the solubility of molecules in drug discovery. The proposed SolPredictor model, based on residual graph neural network convolution, has shown high accuracy in predicting solubility, which can lead to cost and time savings in the drug development process.For the solubility prediction task, the authors mainly rely on two datasets within MoleculeNet.The RGNN architecture with residual connections and gating mechanisms contributes to the model's ability to capture complex relationships within molecular graphs for solubility prediction.SolPredictor, with its RGNN-based architecture and semi-supervised learning approach, presents a promising alternative for molecular solubility prediction. The results indicate competitive performance, improved generalization, and some level of interpretability

Chapter – 5
METHODOLOGY

5. METHODOLOGY

we have taken DELANEY dataset which is a very popular dataset in cheminformatics , consisting of molecular structures along with experimentally measured aqueous solubility values. Pre-processing this dataset typically involves several steps to prepare the molecular structures and associated solubility values for further analysis or modeling. First we have loaded the dataset into our environment and then we have performed the data cleaning steps which involves the handling of missing values , and removing of the duplicate values .And here for the next step we have taken the attribute 'SMILES'(Simplified Molecular Input Line Entry System) as our main data which contains the molecular structures of the chemical compounds in a string format. But in order to fed the data to the graph convolution network the string data must be converted to graph data. To convert this data into graph data we have used the parameter "featurizer" . The featurizer takes the smiles and process them in different steps . Here we have used the "ConvMolFeaturizer()" featurizer which is designed for converting molecular structures represented as SMILES strings into features suitable for input into graph convolutional neural network (GCN) model.

Even though there are some research on solubility where they used the molecular descriptors as the main parameter[5] for predicting the solubility . Molecular descriptors are the numerical or categorical representations of the chemical compounds which includes their properties , structure or composition . But here we've taken the SMILES as our main aspect in order to predict the solubility.

5.1 Graph Convolutional Network (GCN) Model:

Introduction to GCN:

Graph Convolutional Networks (GCNs) are a type of neural network specifically designed to process data structured as graphs. Graphs are mathematical representations composed of nodes (also called vertices) connected by edges (also called links). GCNs have gained significant popularity in recent years due to their ability to effectively learn and extract meaningful representations from graphstructured data. To understand GCNs, let's start with the concept of convolutional neural networks (CNNs). CNNs have been highly successful in tasks like image recognition, where data is structured as grids or arrays. CNNs exploit the local connectivity and shared weights to learn hierarchical

patterns in the data. However, directly applying CNNs to graph structured data is challenging since graphs lack a fixed grid-like structure. GCNs address this challenge by adapting the principles of convolutional layers from CNNs to graphs. The key idea behind GCNs is to propagate information between connected nodes in the graph. This information propagation process is analogous to the local neighborhood aggregation in CNNs. By iteratively aggregating and updating node features, GCNs can capture both local and global structural information of the graph. Let's break down the main components and steps involved in a typical GCN:

- **Graph Representation:** A graph consists of nodes and edges. In the context of GCNs, each node represents an entity or an element of interest, and edges represent relationships or connections between nodes. Nodes and edges can have associated features or attributes.
- **Graph Convolution:** The core operation in GCNs is the graph convolution operation. It aims to aggregate and combine information from a node's local neighborhood (i.e., its connected nodes) to update the node's representation. This process allows each node to learn from its neighbors.
- **Node Representation:** Each node in the graph has an initial feature representation. This representation can be an embedding or a feature vector associated with the node. During the graph convolution process, these initial representations are updated based on the information propagated from the neighboring nodes.
- **Propagation Rule:** The propagation rule determines how information is aggregated from neighboring nodes and combined to update a node's representation. One common propagation rule is to compute a weighted sum of the neighboring node features, where the weights can be learned during the training process.
- **Multiple Layers:** GCNs are typically organized into multiple layers. Each layer performs graph convolutions to refine node representations. Higher layers can capture more global information and dependencies by aggregating information from larger neighborhoods.
- **Output:** The final output of a GCN can vary depending on the task. For example, in a node classification task, the GCN might output a probability distribution over different classes for each node. In a graph classification task, the GCN might produce a single prediction for the entire graph. Overall, GCNs have been successfully applied to a wide range of tasks, including node classification, graph classification, link prediction, and

recommendation systems. By leveraging the connectivity and structural information present in graphs, GCNs enable effective learning and representation of complex relational data.

Node Features:

Assign attributes to the nodes within the molecular graph. These attributes may encompass atom types, atom charges, valence, electronegativity, and other pertinent chemical characteristics.

Edge Features :

Assign characteristics to the connections between nodes in the graph. These characteristics, known as edge features, could encompass attributes such as bond types (single, double, etc.), bond lengths, or other properties inherent to the molecules. fully connected layers working by connecting every neuron in one layer to every neuron in the next layer, facilitating complex mapping between input and output spaces. These attributes may encompass atom types, atom charges, valence, electronegativity, and other pertinent chemical characteristics.

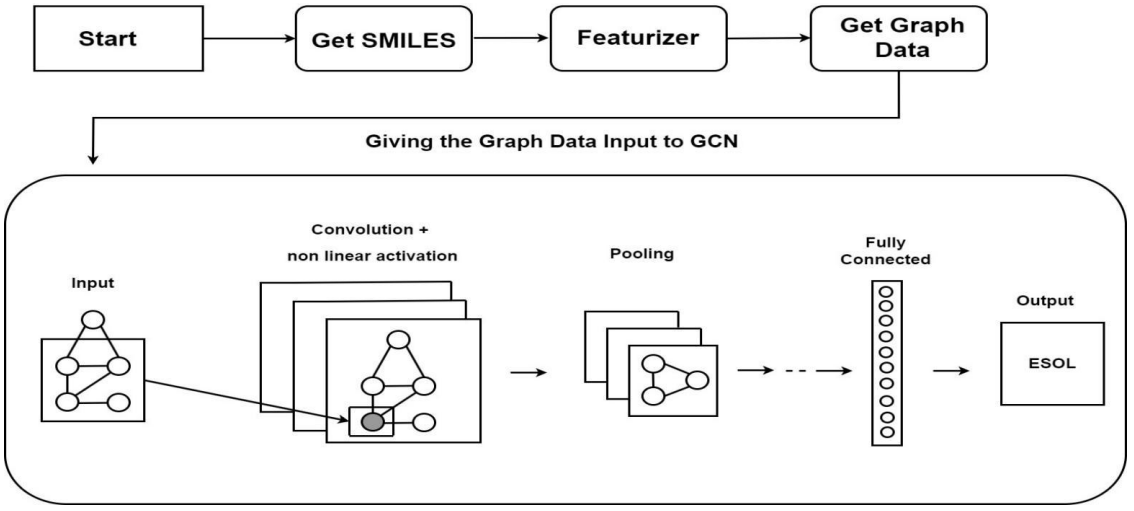

Fig.5.1. Architecture of GCN with QSPR

Graph Convolutional Network (GCN) model specifically for predicting water solubility. The model should include layers dedicated to performing graph convolutions, which extract features from the molecular graph.Convolutional operations are performed

directly on the graph structure. Unlike traditional CNNs where convolutions are applied over local receptive fields in a grid-like structure, in GCNs, convolutions are applied over the graph's adjacency matrix and feature matrices. After the convolution operation, a non-

linear activation function is typically applied element-wise to the resulting feature matrix. pooling operations are used to aggregate information from groups of nodes, effectively reducing the size of the graph or feature space while retaining important structural information. Fully connected layers working by connecting every neuron in one layer to every neuron in the next layer, facilitating complex mapping between input and output spaces. They allow the network to learn intricate patterns and relationships in the data through the adjustment of weights and biases during the training process.In order to prevent the model from overfitting the parameter tuning is used which is the process of optimizing hyperparameters , in GCN we used Grid Hyper-parameter tuning which is a type of parameter tuning that helps in optimizing the hyperparameters to enhance the GCN model's ability to accurately predict the solubility of chemical compounds based on their molecular structures.By performing this parameter tuning it helps the model from overfitting. Integrate an output layer into the GCN model that produces predictions for water solubility, typically approached as a regression task where the model predicts a continuous variable representing solubility**.** GCN model's ability to accurately predict the solubility of chemical compounds based on their molecular structures.

5.2. Working:

GCNs are a type of Graph Neural Network designed to work with graph-structured data, making them well-suited for this task. Here's a workflow for using GCNs to predict water solubility:

5.2.1. Data Collection and Preprocessing:

- **Data Collection:** Gather a dataset of SMILES compounds with known water solubility values. Ensure that the dataset includes both the SMILES representations and the corresponding solubility labels.
- **SMILES Parsing:** Convert SMILES notations into molecular graphs. You can use cheminformatics libraries like RDKit to assist with this step. Each atom in the molecule becomes a node, and each bond becomes an edge in the graph.

5.2.2. Graph Data Representation:

- **Node Features:** Assign features to the nodes in the molecular graph. Node features can include atom types, atom charges, valence, electronegativity, and any other relevant chemical properties.
- **Edge Features (Optional):** Assign features to the edges in the graph. Edge features might include bond types (single, double, etc.), bond length, or other molecular properties

5.2.3. Graph Convolutional Network (GCN) Model:

- **Model Architecture:** Design a GCN model tailored to the task of predicting water solubility. Your model should have layers that perform graph convolutions to extract features from the molecular graph.
- **Output Layer:** Add an output layer to the GCN model. The output layer should generate predictions for water solubility, typically as a regression task where the predicted value is a continuous variable.

5.2.4. Data Splitting:

- **Dataset Split:** Divide your dataset into training, validation, and testing sets. Ensure that all sets have a similar distribution of solubility labels. Common splits include 70% for training, 15% for validation, and 15% for testing.

5.2.5. Training:

- **Loss Function:** Define a loss function suitable for the regression task. Common choices include mean squared error (MSE) or mean absolute error (MAE).
- **Optimizer:** Choose an optimization algorithm, such as stochastic gradient descent (SGD) or Adam, to minimize the loss function. • Training Process: Train the GCN model on the training data while monitoring its performance on the validation set. Implement early stopping to prevent overfitting.

5.2.6. Model Evaluation:

- **Validation Metrics:** Evaluate the model's performance on the validation set using appropriate regression metrics, such as R-squared (coefficient of determination), RMSE (root mean squared error), or MAE (mean absolute error).

5.2.7. Model Testing and Inference:

- **Testing**: Assess the model's performance on the testing set to obtain a realistic estimate of its predictive accuracy.

- **Inference:** Use the trained GCN model to make predictions on new, unseen compounds represented as SMILES notations. Convert the SMILES into molecular graphs and feed them through the model for prediction.

5.2.8. Model Interpretability:

- Analyze the GCN model to understand which features or structural aspects of molecules contribute most to water solubility predictions. Interpretability can provide valuable insights for further drug design and optimization.

5.2.9. Model Refinement:

- If the model performance is not satisfactory, consider refining the model by adjusting hyperparameters, increasing data quality, or incorporating domainspecific knowledge into the feature engineering process. This workflow outlines the process of using Graph Convolutional Networks (GCNs) to predict the water solubility of compounds represented by SMILES notation. It leverages the graph structure of molecules to extract features and make regression predictions, making it suitable for applications in drug discovery and computational chemistry.

Chapter – 6
RESULTS

6. RESULTS

In our extensive exploration of machine learning models for the critical tasks of predicting both water solubility and toxicity in chemical compounds, we embarked on a journey through the realms of Weave Graph Neural Networks (GNNs), Multi-Task Classifiers, and Graph Convolutional Networks (GCNs). With a keen eye on precision, performance, and holistic chemical analysis, we aimed to uncover which of these methodologies would reign supreme in the multifaceted world of molecular properties prediction. The main aim of this work is to explore the performance of various state-of-art ML and DL techniques in addressing the aqueous solubility prediction problem. Graph Convolutional Neural Networks (GCN), have been a central theme in our investigation, being compared with several ML and DL algorithms reliant on molecular graph structures, across a dataset comprising over 1,128 compounds. In the course of this work, we computed and assessed the Root mean squared error (RMSE) of different existing ML and DL models and had developed a Graph Convolutional Neural Network model. We had achieved an RMSE value of 0.43 for GCN model which is less RMSE value when compared to the existing ML and DL models. Some of the ml algorithms here we included are RF, MLR,1D-CNN, XGBoost and among the ml models we observed RF has low RMSE value among DL we had included GAT, AGNN, SGC, GCN Ensemble from which GCN Ensemble has the least RSME value.

Table.6.1. Comparing ML and Dl models to GCN

Algorithms	**RMSE**
1D-CNN	0.971
XG Boost	0.848
RF	0.64
MLR	0.82
GAT	0.71
SGC	0.87
AGNN	0.76
GCN Ensemble	0.44
GCN	**0.43**

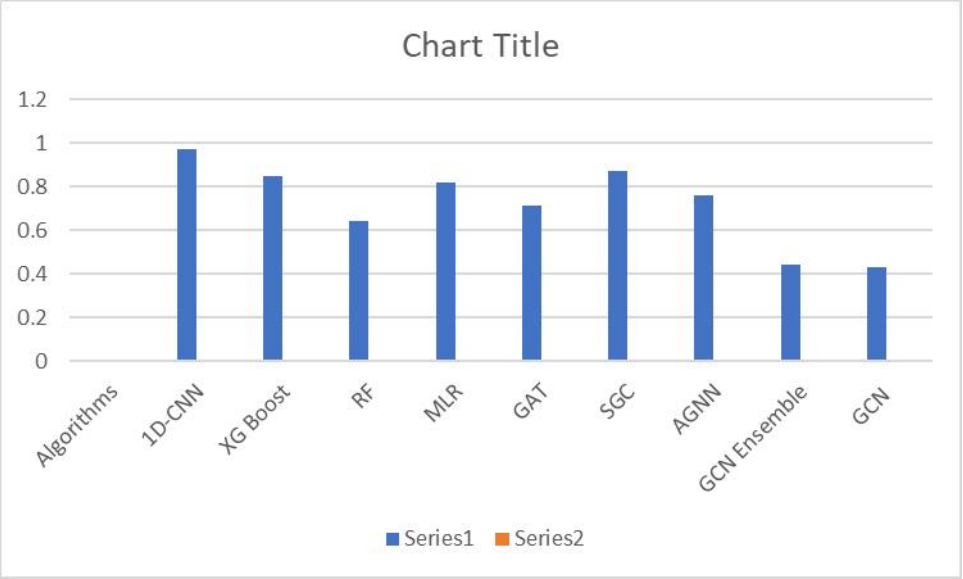

Fig.6.1. Comparison graph of existing models and GCN

Chapter – 7
CONCLUSION

7. CONCLUSION

In conclusion, we have successfully completed our full semester internship, accomplishing in the development of our project. This internship has been a significant chapter in our academic journey, allowing us to apply our theoretical knowledge to realworld scenarios and enhance our practical skills.

It is essential to understand that determining the "best" model for predicting water solubility depends on various factors, including the specific dataset, the nature of the problem, and the resources available. Graph Convolutional Networks (GCNs), Weave Graph Neural Networks (GNNs), and multi-task classifiers all have their strengths and weaknesses. Multi-task classifiers can be beneficial when predicting multiple properties simultaneously, which can lead to improvements in model performance and data efficiency. These models are especially useful when dealing with related tasks alongside water solubility prediction.

GCNs are well-suited for modeling molecular structures and have shown effectiveness in predicting various molecular properties. They can capture local and global structural information and are widely used in cheminformatics for tasks like predicting water solubility. However, their performance may vary depending on the quality and size of the dataset, the choice of hyperparameters, and other factors.

In conclusion, the "best" model for predicting water solubility depends on the specific requirements of the task at hand and the characteristics of the dataset being used. Researchers and practitioners should carefully evaluate different models, consider the nature of their data, and conduct comparative experiments to determine which model performs best for their specific use case. Furthermore, the combination of these models or ensembling techniques may also provide improved results by leveraging the strengths of each approach. It is crucial to approach the problem with a flexible mindset, experimenting with various techniques to determine the most effective solution.

The primary objective of this work has been to explore the aqueous solubility prediction problem and we had chosen graph convolutional networks and had build a model using GCN and had implemented GCN using QSPR approach and has achieved an RMSE value of 0.43 which is less when compared to the existing graph neural network models. Here we have used the grid based hyper parameter for fine tuning the model which had optimization. Grid Search uses a different combination of all the specified hyper

parameters and their values and calculates the performance for each combination and selects the best value for the hyperparameters. Hence using the grid based hyper parameter tuning with the QSPR approach had made the GCN model achieve a RMSE value of 0.43 which is less when compared to some other graph neural networks. n. Future work could explore potential enhancements to further optimize the GCN's performance.

Chapter – 8
FUTURE SCOPE

8. FUTURE SCOPE

Our findings resulting in proving the capabilities of GCNs could be useful for drug discovery and other applications where it is important to predict the water solubility of molecules. GCN models could be used to screen for new drug candidates that are likely to be soluble in water, which is important for oral bioavailability. GCNs are set to revolutionize drug discovery by accurately predicting molecular properties, toxicity, and drug-target interactions. In addition to their roles in social network analysis and NLP tasks, GCNs will prove pivotal in optimizing traffic and transportation systems, enhancing fraud detection in finance, and facilitating smart city planning. Furthermore, GCNs' ethical and responsible use will be paramount. As quantum computing advances, the integration of GCNs with quantum computing holds the potential to accelerate complex modelling and optimization tasks. In the context of Generative Adversarial Networks (GANs) & their quantization to reduce the time complexity, GCNs can be harnessed to generate new drugs by quickly learning from large chemical datasets to create molecular structures with desired properties, opening exciting possibilities in drug design and discovery. There are many new technologies coming up to solve the aqueous solubility prediction problem. Hence, Sequence based learning and Natural Language Processing can be used for checking the solubility of the drugs.

Chapter – 9
REFERENCES

9.REFERENCES

[1] Ahmad, W.; Tayara, H.; Shim, H.; Chong, K.T. SolPredictor: Predicting Solubility with Residual Gated Graph Neural Network. Int. J. Mol. Sci. 2024, 25,715. https://doi.org/10.3390/ijms25020715

[2] Ahmad, W., Tayara, H., & Chong, K. T. (2023). Attention-Based Graph Neural Network for Molecular Solubility Prediction. ACS Omega 2023, 8, 3236−3244.

[3] Tayyebi, A., Alshami, A.S., Rabiei, Z. *et al.* Prediction of organic compound aqueous solubility using machine learning: a comparison study of descriptor-based and fingerprints-based models. *J Cheminform* **15**, 99 (2023). https://doi.org/10.1186/s13321-023-00752-6.

[4] Meng, J., Chen, P., Wahib, M. *et al.* Boosting the predictive performance with aqueous solubility dataset curation. *Sci Data* **9**, 71 (2022). https://doi.org/10.1038/s41597-022-01154-3

[5] Zheng, T., Mitchell, J. B. O., & Dobson, S. (2023). Machine Learning for Solubility Prediction. Research Square. https://doi.org/10.21203/rs.3.rs-3544641/v1.

[6] Wu F, Zhang T, Souza A, Fifty C, Yu T, Weinberger KQ (2019) Simplifying graph convolutional networks.In: ICML 6861–6871

[7] Velickovic P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y (2018) Graph attention networks. https://arxiv.org/abs/1710.10903

[8] Thekumparampil KK, Wang C, Oh S, Li LJ (2018) Attentionbased graph neural network for semi-supervised learning. https:// arxiv.org/abs/1803.03735

[9] Deng, C., Liang, L., Xing, G. *et al.* Multi-channel GCN ensembled machine learning model for molecular aqueous solubility prediction on a clean dataset. *Mol Divers* **27**, 1023–1035 (2023). https://doi.org/10.1007/s11030-022-10465-x

[10] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system, 785ś794 (2016) https://doi.org/10.1145/2939672.2939785

[11] Nakano, M., Sugiyama, D.: Discriminating seismic events using 1d and 2d CNNs: applications to volcanic and tectonic datasets. Earth, Planets and Space **74**(1)(2022) https://doi.org/10.1186/s40623-022-01696-1

APPENDIX

APPENDIX

APPENDIX

```
import pandas as pd
import numpy as np
import networkx as nx
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
import tensorflow as tf
from sklearn.model_selection import train_test_split

# Sample data
csv_path = '/content/drive/MyDrive/Colab Notebooks/clintox.csv'  # Replace with the
actual path to your CSV file
df = pd.read_csv(csv_path)

# Drop rows with missing SMILES strings
df = df.dropna(subset=['smiles'])

# Drop 'FDA_APPROVED' column
df = df.drop(columns=['FDA_APPROVED'])
df
# Tokenize SMILES strings
tokenizer = Tokenizer(char_level=True)
tokenizer.fit_on_texts(df['smiles'])
df['smiles_sequences'] = tokenizer.texts_to_sequences(df['smiles'])

# Create a graph representation (simplified, assuming each character is a node)
graphs = []
for seq in df['smiles_sequences']:
    graph = nx.Graph()
    for i in range(len(seq) - 1):
```

```
        graph.add_edge(seq[i], seq[i + 1])
    graphs.append(graph)

# Get the maximum number of nodes in the graphs
max_nodes = max([len(graph.nodes) for graph in graphs])

# Pad adjacency matrices to a fixed size
padded_adj_matrices = [np.pad(nx.adjacency_matrix(graph).todense(),
                              ((0, max_nodes - len(graph.nodes)), (0, max_nodes -
len(graph.nodes))),
                              mode='constant') for graph in graphs]

# Convert data to TensorFlow tensors
X_data = np.array(padded_adj_matrices)
y_data = df['CT_TOX'].to_numpy()

# Split the dataset into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X_data, y_data, test_size=0.3,
random_state=42)

# Verify shapes
print("X_train shape:", X_train.shape)
print("y_train shape:", y_train.shape)
print("X_test shape:", X_test.shape)

import matplotlib.pyplot as plt

# Select a graph from the graphs list (you can change the index)
selected_graph = graphs[1]

# Visualize the graph
pos = nx.spring_layout(selected_graph)  # You can choose a different layout if needed
nx.draw(selected_graph, pos, with_labels=True, font_weight='bold', node_size=500,
```

```
node_color='skyblue', font_size=8)

# Show the plot
plt.show()
graphs[1]
print("y_test shape:", y_test.shape)

# Choose a random sample from the test set
# sample_index = np.random.randint(0, X_test.shape[0])
index=39
sample_input = X_test[index]
actual_output = y_test[index]  # Assuming the corresponding CT_TOX value is stored
in y_test

# Reshape the input for prediction
sample_input_reshaped = sample_input.reshape((1, flattened_shape))

# Predict using the trained model
predicted_output = model.predict(sample_input_reshaped)

# Reshape the predicted output to the original shape
predicted_output_reshaped = predicted_output.reshape((max_nodes, max_nodes))

# Create networkx graphs from adjacency matrices
original_graph = nx.Graph(sample_input)
reconstructed_graph = nx.Graph(predicted_output_reshaped)

# Plot the original graph
plt.figure(figsize=(10, 5))
plt.subplot(121)
plt.title('Original Graph')
nx.draw(original_graph, with_labels=True, font_weight='bold')
```

```
# Plot the reconstructed graph
plt.subplot(122)
plt.title('Reconstructed Graph')
nx.draw(reconstructed_graph, with_labels=True, font_weight='bold')

plt.show()

# Perform classification on the sample
classification_input = sample_input_reshaped.flatten()  # Flatten the input for classification
classification_input = np.expand_dims(classification_input, axis=0)  # Add batch dimension

# Perform classification using the trained model
classification_output = model.predict(classification_input)

# Convert the output to binary (1 or 0) based on a threshold
classification_threshold = 1
predicted_class = 1 if classification_output[0, 0] > classification_threshold else 0

print("Actual CT_TOX Value:", actual_output)
print("Predicted Class (1=Toxic, 0=Non-Toxic):", predicted_class)

import deepchem as dc

# Define the data loader
loader = dc.data.CSVLoader(
    tasks=['ESOL predicted log solubility in mols per litre'],# Task name (logS is a common task for the Delaney dataset)
    smiles_field='smiles', # Name of the field containing SMILES strings
    featurizer=dc.feat.ConvMolFeaturizer()  # Featurizer for converting SMILES to features
)
```

```
# Load the dataset from your folder
dataset = loader.create_dataset('delaney-process.csv')

# Split the dataset into train, validation, and test sets
splitters = dc.splits.RandomSplitter()
train_dataset, valid_dataset, test_dataset = splitters.train_valid_test_split(dataset)

# Define and train your model (for example, using a graph convolutional network)
model = dc.models.GraphConvModel(n_tasks=1, mode='regression')
model.fit(train_dataset, nb_epoch=50)

# Evaluate the model
rmse_metric = dc.metrics.Metric(dc.metrics.mean_squared_error, mode='regression')
train_scores = model.evaluate(train_dataset, [rmse_metric])
valid_scores = model.evaluate(valid_dataset, [rmse_metric])
test_scores = model.evaluate(test_dataset, [rmse_metric])

print("Train RMSE:", train_scores['mean_squared_error'])
print("Validation RMSE:", valid_scores['mean_squared_error'])
print("Test RMSE:", test_scores['mean_squared_error'])
```