

Democratic and Popular Algerian Republic
Ministry of Higher Education and Scientific Research

Yahia Fares University of Medea
Faculty of Sciences
Department of Mathematics and computer science

Final Master Project

Branch: Computer Science
Speciality: System Engineering and Web Technologies

Ligand-Based Virtual Screening Using Graph Neural Network and Learning Representation

Prepared and presented by:

ANTRI Chakib Arselane
HASNAOUI Djellal

Supervisor: **Hentabli Hamza**
Co-Supervisor: **Mazari Redha**

June 2023

Acknowledgment

First, we would like to thank ALLAH. for all the achievements that we have gained today. Next, we would like to extend our deepest appreciation to our supervisor Hentabli Hamza and our Co-supervisor, Mazari Redha, for their guidance, expertise, and encouragement in our research.

We would also like to thank the participants who generously shared their time and insights for this study.

We would like to express our gratitude to the staff and faculty of science for providing the necessary resources and facilities.

Additionally, we would like to acknowledge the contributions of our colleagues and friends who provided valuable feedback, suggestions, and moral support throughout the research process. Their input and discussions greatly enriched our understanding and helped shape the final version of this paper.

Lastly, we would like to thank our families for their unwavering support, understanding, and encouragement throughout this journey. Their love and encouragement kept us motivated during challenging times.

While we have made every effort to acknowledge all individuals who have contributed to this thesis, we apologize if anyone has been inadvertently omitted. We are truly grateful to everyone who has played a role, no matter how big or small, in the completion of this research paper.

Dedication

“This thesis is dedicated to my parents, beloved family. Their unwavering support and encouragement have been instrumental in my journey. Their belief in me has been a constant source of motivation, pushing me to overcome challenges and pursue my passions. I am grateful for their presence in my life, and I dedicate this work to them as a token of my deep appreciation.“

Antri Chakib

“This thesis is dedicated to my parents, beloved family,my childhood friend "Khadidja" and all my friends. Their unwavering support and encouragement have been instrumental in my journey. Their belief in me has been a constant source of motivation, pushing me to overcome challenges and pursue my passions. I am grateful for their presence in my life, and I dedicate this work to them as a token of my deep appreciation.“

Hasnaoui Djellal

Abstract

Information technology has become an essential component of the predict molecular properties process and virtual screening (VS) has emerged as a valuable computational technique for efficiently and cost-effectively screening chemical compounds. A crucial task in VS is the similarity search, which aims to estimate the similarity between molecules. This approach is based on the assumption that molecules with similar structures may exhibit similar biological activities. Several techniques have been developed to compare the biological similarity between a target compound and each compound in a database. While these methods perform well, particularly for molecules with homogeneous active structures, they often struggle when dealing with structurally diverse compounds. Ligand-based virtual screening (LBVS) is a critical computational approach in drug development that aids in the identification of potential lead compounds with high affinity and selectivity for a target protein. Traditional LBVS methods often rely on molecular descriptors and similarity measures, which may not adequately capture the complex relationships between chemical structures and biological activities. In recent years, graph neural networks (GNNs) have emerged as a powerful tool for learning representations of molecular graphs and predicting their properties. This paper presents an abstract on the application of GNNs in LBVS for drug development, highlighting their potential to overcome the limitations of conventional methods and enhance the efficiency and effectiveness of lead compound identification.

الملخص:

أصبحت تكنولوجيا المعلومات عنصرًا أساسيًا في عملية التنبؤ بخصائص الجزيئات وقد ظهر الفحص الافتراضي كتقنية حوسبة قيمة لفحص المركبات الكيميائية بكفاءة وبتكلفة منخفضة. ومن بين المهام الحاسمة في الفحص الافتراضي هي البحث عن التشابه، الذي يهدف إلى تقدير التشابه بين الجزيئات. تعتمد هذه النهج على افتراض أن الجزيئات ذات البنية المماثلة قد تظهر أنشطة حيوية مماثلة. تم تطوير العديد من التقنيات لمقارنة التشابه البيولوجي بين المركب المستهدف وكل مركب في قاعدة البيانات. بينما تؤدي هذه الأساليب بشكل جيد، لا سيما للجزيئات ذات الهياكل النشطة المتجانسة، إلا أنها غالبًا ما تواجه صعوبات عند التعامل مع المركبات ذات التنوع الهيكلي. يعد الفحص الافتراضي المبني على الرابطة نهجًا حاسوبيًا حاسمًا في تطوير الأدوية الذي يساعد في تحديد المركبات الرائدة المحتملة ذات التأثير العالي والانتقائية التقليدية على الأوصاف على بروتين مستهدف. غالبًا ما يعتمد أساليب الجزيئية ومقاييس التشابه، والتي قد لا تتمكن من التقاط العلاقات المعقدة بين الهياكل الكيميائية والأنشطة البيولوجية بشكل كاف. في السنوات الأخيرة، ظهرت الشبكات العصبية الرسومية كأداة قوية لتعلم التمثيلات الرسومية للجزيئات وتوقع خصائصها. يقدم هذا البحث ملخصًا عن تطبيق الشبكات العصبية الرسومية في الفحص الافتراضي المبني على الرابطة لتطوير الأدوية، مؤكدًا إمكانية تجاوز القيود المفروضة على الأساليب التقليدية وتعزيز كفاءة وفعالية تحديد المركبات الرائدة.

الكلمات المفتاحية: الفحص الافتراضي، الفحص الافتراضي المبني على الرابطة، الشبكات العصبية الرسومية، التمثيلات الرسومية للجزيئات.

Abbreviation	Description
2D	Two-Dimensional
3D	Three-Dimensional
3D-QSAR	Three-Dimensional Quantitative Structure-Activity Relationship
AI	Artificial Intelligence
ASCII	American Standard Code for Information Interchange
CAS RN	CAS Registry Number
CASTp	Computed Atlas of Surface Topography of Proteins
CADD	Computer Aided Drug Design
CAS Registry Number	CAS RN
ChEMBL	Chemical Biology Database
CID	Compound Identifier
CNNs	Convolutional neural networks
CT	Chemical table
CTab	Connection table
DS 1, DS 2 , DS 3	Dataset 1, Dataset 2, Dataset 3
GCN	Graph Convolutional Network
GNNs	Graph Neural Networks
GRU	Gated recurrent unit
HTS	High-throughput screening
LBDD	Ligand-based drug design
LBVS	Ligand-Based Virtual Screening
LBVSMPNN	Ligand-Based Virtual Screening Message Passing
Neural Network MACCS	Molecular ACCess System
MDDR	MDL Drug Data Report
ML	Machine Learning
MPNN	Message Passing Neural Network
NMR	Nuclear Magnetic Resonance
QSARs	Quantitative Structure-Activity Relationships
R-GNN	Recurrent Graph Neural Network
RDD	Rational Drug Design
RN	Registry Number
SARs	Structure-activity relationships
SBDD	Structure-Based Drug Design
SBVS	Structure-Based Virtual Screening
SMILES	Simplified Molecular Input Line Entry System

Contents

Acknowledgment ii

Dedication iii

Abstract iv

Abbreviations vii

Contents vii

List of Figures xi

List of Tables xiii

1 Introduction 1

1.1 Problem background 1

1.2 Problem statment 3

1.3 Research Questions 4

1.4 Research Objective 5

1.5 Research Scope 5

1.6 Thesis Organization 6

2 Literature Review 8

2.1 Introduction 8

2.2 Cheminformatics 9
2.3 Computer-aided drug design (CADD) 9
2.3.1 Rational drug design (RDD) 11
2.3.2 Structure-based drug design (SBDD) 11
2.3.3 Ligand-based drug design (LBDD) 13
2.3.3.1 Quantitative structure-activity relationship (QSAR) . . . 13
2.3.3.2 Pharmacophore modeling 14
2.3.4 Virtual Screening 14
2.4 Molecular Representations 15
2.4.1 String Representation 16
2.4.1.1 Registry system 16
2.4.1.2 Simplified molecular-input line-entry system SMILE . . . 18
2.4.2 Chemical Table Representations 19
2.4.2.1 Graph Representation 20
2.4.2.2 Molecular fingerprints 21
2.5 Molecular similarity measurement 23
2.6 Molecular Biological Activity 24
2.6.1 molecular structure-activity relationship 24
2.7 Artificial intelligence 26
2.8 Machine learning (ML) 26
2.9 Deep learning 28
2.10 Graph neural network GNN 28
2.10.0.1 Message-passing neural networks MPNN 29
2.10.0.2 Graph convolutional network GCN 31
2.10.0.3 Recurrent Graph Neural Network 33
2.10.1 Transformer Encoder 34
2.11 Related Work 35
2.12 Conclusion 36

3 Research Methodology 37

3.1 Introduction . . . 37

3.2 Research Framework . . . 37

3.2.1 Phase 1 . . . 39

3.2.2 Phase 2 . . . 39

3.2.3 Phase 3 . . . 39

3.2.4 Phase 4 . . . 40

3.3 Dataset . . . 40

3.3.1 DS1 . . . 41

3.3.2 DS2 . . . 42

3.3.3 DS3 . . . 43

3.4 Design and steps . . . 44

3.4.1 Loading and Preprocessing Data . . . 45

3.4.2 SMILE to Graph . . . 48

3.4.2.1 molecule from smile . . . 48

3.4.2.2 Graph from molecule . . . 48

3.4.3 Splitting the Data . . . 51

3.4.4 Message Passing . . . 51

3.4.5 Ttransformer Encoder . . . 52

3.4.6 MPNN Model . . . 52

3.4.7 Training the model . . . 53

3.5 Performance Evaluation . . . 54

3.5.1 Ligand-based virtual screening . . . 54

3.6 Conclusion . . . 55

4 Results and Discussion 56

4.1 Introduction . . . 56

4.2 Results and Discussion . . . 56

4.2.1 MPNN Model 56
4.2.2 Ligand-based virtual screening 58
4.3 Conclusion 71

Conclusion **72**

Bibliography **74**

List of Figures

2.1 Computer-aided drug design [1] 10
2.2 Overview of the different classes of molecular representations [2] 16
2.3 Molecular string representations [3] 17
2.4 Aspirin SMILES representation [4] 19
2.5 Chemical table representation MDL molfile [5] 20
2.6 acetic acid and adjacency matrix Graph Representation [5] 21
2.7 node features matrix and edge features matrix Graph Representation [5] . 21
2.8 representation of a hypothetical fingerprint [6] 22
2.9 Main structure-activity relationships found in [7] study. 25
2.10 Types of Artificial intelligence [8] 26
2.11 Schematic of a message passing neural network (MPNN)[9] 30
2.12 General architecture of GCN [10] 33
2.13 Transformer Encoder sub-layers 34
2.14 Multi-head self-attention mechanism [11] 35

3.1 Flow Chart of the Framework 38
3.2 Chart that represents the Design steps 44
3.3 Molecule Object 49
3.4 MPNN model Plot 54

4.1 MPNN Model accuracy chart 57
4.2 MPNN Model loss chart 58

4.3 The comparison among methods for the average recall percentage at the top 1% in MDDR-DS1 66
4.4 The comparison among methods for the average recall percentage at the top 5% in MDDR-DS1 67
4.5 The comparison among methods for the average recall percentage at the top 1% in MDDR-DS2 68
4.6 The comparison among methods for the average recall percentage at the top 5% in MDDR-DS2 69
4.7 The comparison among methods for the average recall percentage at the top 1% in MDDR-DS3 70
4.8 The comparison among methods for the average recall percentage at the top 5% in MDDR-DS3 71

List of Tables

2.1 Similarity measurement table. 23

3.1 The MDDR-DS1 structure activity classes 41
3.2 The MDDR-DS2 structure activity classes 42
3.3 The MDDR-DS3 structure activity classes 43
3.4 DS1 Dataframe 46
3.5 DS2 Dataframe 46
3.6 DS3 Dataframe 47
3.7 MDDR Classes DataFrame 47

4.1 Retrieval results of MDDR-DS1 for Activity classes 59
4.2 Retrieval results of MDDR-DS2 for Activity classes 59
4.3 Retrieval results of MDDR-DS3 for Activity classes 60
4.4 Retrieval results of top 1% for MDDR-DS1 dataset 61
4.5 Retrieval results of top 5% for MDDR-DS1 dataset 62
4.6 Retrieval results of top 1% for MDDR-DS2 dataset 63
4.7 Retrieval results of top 5% for MDDR-DS2 dataset 64
4.8 Retrieval results of top 1% for MDDR-DS3 dataset 65
4.9 Retrieval results of top 5% for MDDR-DS3 dataset 65

Chapter 1

Introduction

1.1 Problem background

Before the advent of machine learning, drug discovery relied heavily on high-throughput screening (HTS) to identify potential drug candidates. HTS involves screening large libraries of compounds against specific biological targets to identify those with the desired properties. However, this process can be time-consuming, and expensive, and often leads to false positives, where compounds that appear promising in screening do not have the desired effect in clinical trials. Therefore, the drug discovery process often took years and required substantial investment in time and resources.

In recent years, the emergence of machine learning and artificial intelligence has transformed the drug discovery process by enabling the rapid identification of potential drug candidates with higher accuracy and efficiency. Machine learning algorithms have the ability to rapidly analyze vast quantities of information, such as chemical structures, gene sequences, and clinical trial data, in order to identify potential drug candidates and make predictions about their effectiveness and safety.

Machine learning has revolutionized the field of drug discovery in recent years by

enabling researchers to analyze vast amounts of data more efficiently and accurately than ever before.

The first applications of machine learning in drug discovery can be traced back to the early In the 1990s when researchers began using neural networks to model the structure-activity relationships (SARs) of small molecules. Over time, machine learning algorithms became more sophisticated, and researchers began to incorporate more data types into their models like graphs. Although it is well known that graphs are ubiquitous in the real world, it is very challenging to utilize deep learning methods to analyze graph data [12]. using deep learning techniques to analyze data in the form of graphs poses a significant challenge. This issue is not easy to address due to the obstacles presented by the following factors: Firstly, graphs have irregular structures, unlike images, audio, and text, which have grid-like structures. This irregularity makes it challenging to extend basic mathematical operations to graphs, such as defining convolution and pooling operations - fundamental operations in convolutional neural networks (CNNs). Secondly, graphs exhibit heterogeneity and diversity, adding to the complexity of the problem. A graph itself can be complicated, containing diverse types and properties. These diverse types, properties, and tasks require different model architectures to tackle specific problems [12]. Thirdly graphs of massive scale. And lastly Integrating cross-disciplinary knowledge. Graphs frequently have links to other fields, including biology, and chemistry.

Graphs have become increasingly important in drug discovery, with many studies focused on developing and applying graph-based machine-learning techniques to improve drug discovery processes. The development of GNNs represented a significant breakthrough in drug discovery. Graph Neural Networks (GNNs) have shown promising results in various areas of drug discovery, such as virtual screening, the previous attempts using GNNs have had some limitations or challenges that researchers are working to overcome. some challenges still need to be addressed, including the limited interpretability of GNNs, the difficulty of dealing with large- scale graphs, and the need for better methods to

incorporate heterogeneous data types into GNNs.

Researchers are actively working to overcome these challenges and further improve the performance of GNNs in drug discovery applications.

1.2 Problem statment

Although chemists have applied machine learning to many problems over the years, predicting the properties of molecules and materials using machine learning (and especially deep learning) is still in its infancy [13]. The other main line of research is the optimization of the model architecture, whether the model is applied to descriptors or fingerprints [14][15] or is directly applied to SMILES [16] strings [14] or the underlying graph of the molecule [17].

Graph-based learning methods have been widely developed. Several studies have demonstrated that graph-based neural network models can learn pharmacological features and generate de novo molecules with the desired properties [18]. Some methods are based on convolutional networks [19][20][21], and others use the attention mechanism [22] to featurize both local and global regions of a graph. [23][24][25] showed that graph convolution combined with attention mechanisms can improve the predictive performance than when only the graph convolution was used.

Like GCN methods, the Message-passing neural network model learns task-specific molecule features from the graph structure and avoids feature engineering in the preprocessing stage. This type of method also presents an approach for the secure sharing of chemical data, i.e. it is possible to disseminate trained models for activity predictions without the risk of reverse-engineering IP-sensitive structural information [26][27].

The activity and properties of a molecule depend on its chemical structure, which

can vary in a large number of ways. Additionally, the interactions between a molecule and its target are often highly specific and can be influenced by a wide range of factors, such as steric hindrance, electrostatics, and hydrogen bonding. To address this problem, computational methods are used to predict the biological activity and physical properties of molecules. Combination of Ligand-based virtual screening and Graph neural network to learn the molecular representation of a known active compound and then using this representation to screen a database of compounds using ligand-based virtual screening.

Today we started a new project that combines message-passing neural network technologies and Ligand-based virtual screening, Our goal is to leverage the power of GNNs to predict molecular properties and use ligand-based virtual screening to identify potential new drug candidates with desired properties, ultimately accelerating the drug discovery process.

1.3 Research Questions

The aim of the research is overall to address an existing gap in the knowledge, to devise and test a solution to an existing problem, and so This study aims to answer the following question:

- Could graph neural networks learn better molecular representation to predict properties?
- Is Message passing a flexible framework to implement so that it can be applied to any prediction task with Graph representation?
- Is it possible to achieve improved results by combining the GNNs models and the Ligand-based virtual screening approach?
- Can we trust the results of our research?
- How far our result can be actionable in real life?

1.4 Research Objective

The research objective of this study is to develop a more accurate method for predicting molecular properties using a combination of Graphs neural networks (GNNs) and ligand-based virtual screening. Specifically, our goals are to:

- pre-process and prepare a dataset of molecular structures and their associated properties.
- To develop and optimize a GNN model for predicting molecular properties.
- To implement a ligand-based virtual screening approach to identify potential drug candidates from a large compound library.
- To evaluate the performance of the GNN model and the ligand-based virtual screening approach individually and in combination.
- To compare the results of the combined approach to those obtained from each individual approach and assess the added value of the integration of the two methods.
- To provide insights into the relationship between molecular structures, properties, and ligand- binding interactions to inform drug design and discovery efforts.

The ultimate goal of this research is to contribute to the development of more efficient and accurate methods for predicting molecular properties and identifying potential drug candidates.

1.5 Research Scope

The research scoop involves developing a novel approach for predicting molecular properties using a combination of Graph neural Networks (GNNs) and ligand-based virtual screening. The proposed method leverages the power of GNNs to learn representations

of molecular graphs and uses these representations to predict various molecular properties.

Additionally, the method employs ligand-based virtual screening to further refine the predictions by considering the similarity of the molecular structures with known active compounds.

The approach is expected to improve the accuracy of molecular property prediction and facilitate the discovery of new drug candidates. The effectiveness of the proposed method is demonstrated through extensive experiments on various datasets, which shows significant improvements in predictive performance compared to existing state-of-the-art methods.

1.6 Thesis Organization

The thesis is structured into five chapters. Chapter one serves as an introduction and provides a brief overview of the project, including the problem background, statement, and objectives.

Moving to chapter two, a literature review is conducted, covering the fields of Cheminformatics, Computer-aided drug design, and Molecular Representations. This includes background knowledge of the key terms related to similarity searching and deep learning algorithms.

In chapter three, the research methodology is detailed, with a focus on the techniques used for predicting molecular properties through the implementation of deep learning algorithms. This section also delves into the required hardware and software for the project. Chapter four presents the findings of the deep learning algorithm application and offers a comprehensive discussion of the results.

Finally, chapter five provides a conclusion of the work conducted throughout the thesis.

Chapter 2

Literature Review

2.1 Introduction

The application of deep learning methods to the prediction of the biological activity and physical properties of molecules in drug discovery programs has resulted in one of the most significant advances in the development of quantitative structure-activity relationships (QSARs), Rather than using expert-derived chemical features, which are typically used to build predictive models [28].

for that reason, machine learning became one of the most important and rapidly evolving topics in computer-aided drug discovery[29], Pattern recognition algorithms are used in machine learning approaches to determine mathematical relationships between empirical observations of small molecules and extrapolate them to predict chemical, biological, and physical properties of novel compounds[30].one of the main objectives of computer-aided drug discover helping researchers understand and exploit relationships between chemical structures and their biological activities or SAR[31]

2.2 Cheminformatics

Chemoinformatics is a field that combines chemistry, computer science, and mathematics to develop methods for saving, retrieving, analyzing, and visualizing chemical and biological data. The area has evolved as a result of the fast growth of chemical and biological data in recent years, and the necessity for effective tools to organize and evaluate this data, this development is causing a large amount of data to be generated and thus used by chemists [32]

Chemoinformatics plays a vital role in drug discovery, where it helps to identify and optimize promising drug candidates based on their chemical and biological features. Additionally, it has applications in other domains such as materials science, environmental science, and bioinformatics. It has generally evolved into a crucial tool for chemists and biologists to make sense of complex chemical and biological data and to quicken the pace of discovery in these fields.

many Methods have been created for the early stages of drug design that are employed by all major pharmaceutical corporations, all domains of chemistry can benefit from chemoinformatics methods [33]

2.3 Computer-aided drug design (CADD)

Computer-aided drug design or CADD is a set of methods used for the development of new drugs by using computer-based models, algorithms, and simulations for the prediction of molecules with biological targets, Nowadays, this methodology is being utilized as a supplementary tool to experimental biochemical investigations and the determination of 3-D structure[34].

To design a medicinal product for treating a disease or relieving a symptom, a clear understanding of the disease pathway and relevant processes is crucial for selecting a therapeutic target.[35] In general, modeling approaches are categorized into structure-based and ligand-based methods Figure.2.1, structure-based drug design or SBDD uses 3D structures of the target (enzyme/receptor) while ligand-based methods rely on the known chemical and biological properties of the set of molecules with similar structures and activities.

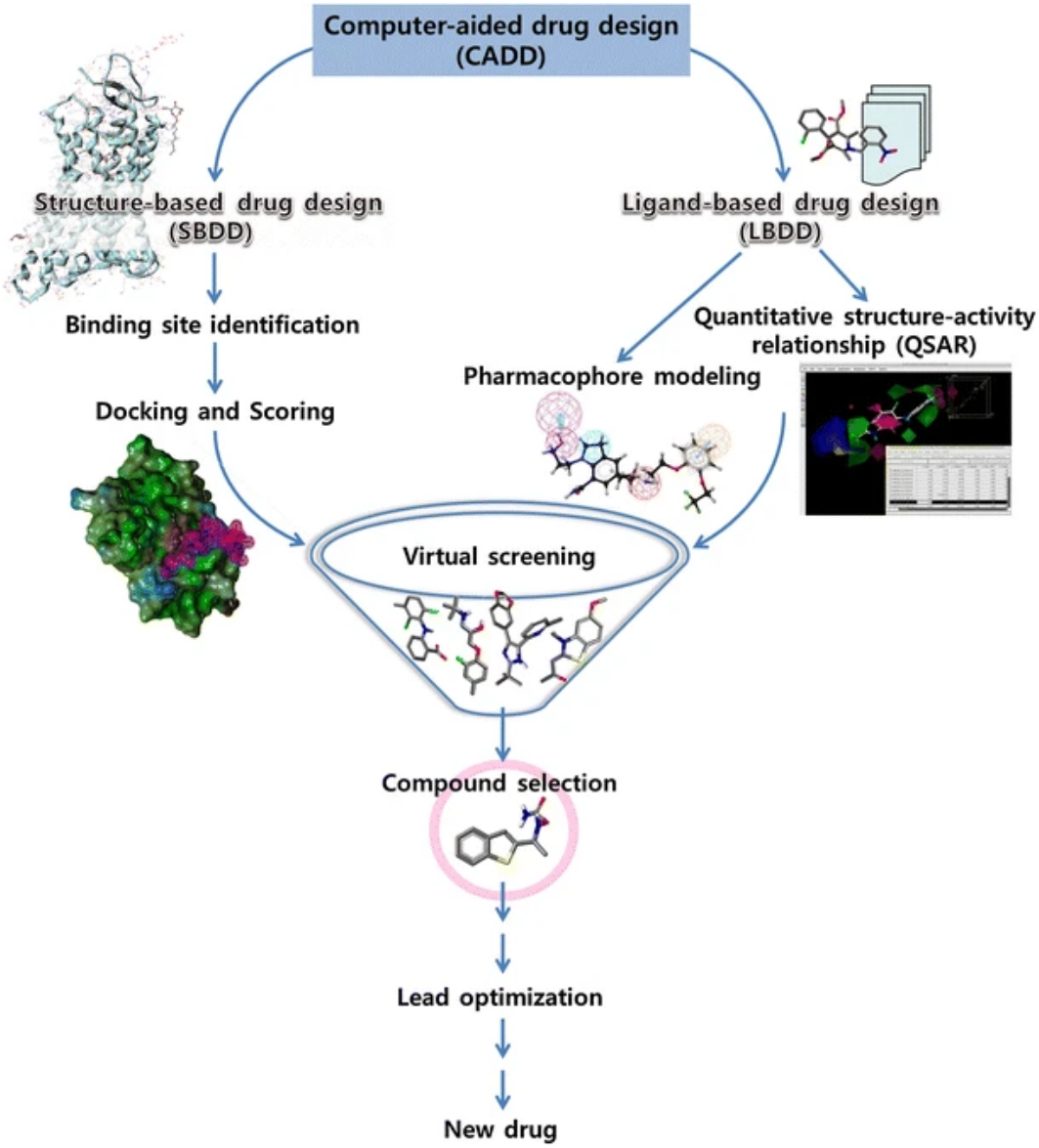


Figure 2.1: Computer-aided drug design [1]

Overall, CADD has the capacity to significantly accelerate the drug discovery process, however, it is important to understand that these methods are only a part of drug discovery.

2.3.1 Rational drug design (RDD)

Modern drug design revolves around the principles of rationality and methodology, with the pharmaceutical industry seeking greater control in proposing new structures for drugs, despite past successes in drug discovery by chance or through the isolation of natural products. Rational drug design (RDD) utilizes various techniques based on a semantic approach to improve pharmacological properties and decrease toxicity in drug candidates.

RDD is classified into two categories [1][36]:

- The first category focuses on developing small molecules with desired properties for known targets, such as proteins or nucleic acids, with known functional roles and 3D structural information.
- The second category involves developing small molecules with predefined properties for targets, whether the cellular functions and structural information are known or unknown.

The first category is more commonly used by the pharmaceutical industry, while the second is utilized more for academic purposes, but collaboration between the two allows for effective communication between both segments.

2.3.2 Structure-based drug design (SBDD)

The 3D structure of a target is used in the SBDD drug discovery method to create tiny compounds that bind specifically and are highly affinitive to the active site of a protein. This strategy is based on the idea that a protein's structure and active site are crucial

to its biological function and that tiny compounds can be created to mimic the form and chemical composition of the protein's natural ligand in order to bind it and control its activity. SBDD typically involves a series of steps :

- Target Identification: Finding a target protein associated with a specific disease or ailment is the first step. Numerous methods, including genomics, proteomics, and biochemical assays, can be used to do this.
- Structural Characterization: The next step is to create a high-resolution 3D structure of the target protein using techniques like X-ray crystallography or NMR spectroscopy. The protein's active site and interactions with ligands are detected.
- Ligand Binding Site Identification: Finding the protein's ligand binding site, which is typically a pocket or cavity on the protein's surface where ligands can bind, is the next step after determining the protein's structure. Different computational techniques, such as CASTp or SiteMap, can be used to do this.
- Virtual Screening: In virtual screening, large libraries of compounds are screened computationally against the protein's binding site to identify potential hits or leads. This can be done using molecular docking, molecular dynamics simulations, or other
- Hit Optimization: The next step after finding possible hits is to maximize their affinity and selectivity for the target protein. This can be accomplished utilizing a variety of computational and experimental techniques, including high-throughput screening, fragment-based design, and structure-based design.
- Lead Optimization: The next step is to optimize a lead compound's pharmacological characteristics, such as potency, selectivity, and toxicity, when it has been shown to have the desired attributes. Structure-based design, synthesis, and biological testing may all be used in this process.
- Preclinical and Clinical Evaluation: The lead compound is examined in preclinical and clinical research to determine its safety, effectiveness, and pharmacokinetic

characteristics in the last phase. Before the medication is licensed for use in humans, this process might take several years and involves multiple animal models and clinical trials.

2.3.3 Ligand-based drug design (LBDD)

ligand-based drug design has a relatively long history, dating back to the 20th century when chemists first started investigating the structure of proteins and enzymes. the field really began to take off in the 1970s, when advances in computational and molecular biology methods allowed researchers to start designing molecules that could interact with specific targets in a more rational and systematic way. The development of the anti-inflammatory medication naproxen in 1976 was the earliest instance of ligand-based drug design [37].

LBDD is used to develop new drugs where the focus is on designing a molecule that will interact specifically with a target protein or enzyme. it starts with the identification of the target's protein or enzyme that is involved in a disease process, and then designing a molecule (or a ligand) that will interact with it to either activate or inhibit its function.

2.3.3.1 Quantitative structure-activity relationship (QSAR)

In ligand-based drug design, one of the key approaches is to identify potential drug candidates by analyzing the physicochemical and topological descriptors of a set of similar molecules. These descriptors may include electronic, hydrophobic, and steric properties, among others [38]. A multivariate regression analysis is then performed between these descriptors and the biological activity of the molecules. This generates an equation that represents a model of the system being investigated.

One popular method for generating these models is the three-dimensional quantitative structure-activity relationship (3D-QSAR) approach [39]. This approach involves aligning

the three-dimensional structures of compounds and placing a probe, such as a positive carbon, around a grid of points. The steric and electrostatic energy of each point is then calculated and used as a physicochemical descriptor in the model. The biological activity data is used as the dependent variable.

The points of interaction that are favorable or unfavorable to biological activity are then converted into a three-dimensional map. This map shows regions that are favored by large or small groups, as well as regions that are favored by positive or negative groups. By analyzing these maps, researchers can gain insights into the optimal molecular features necessary for strong binding to the target protein or enzyme.

2.3.3.2 Pharmacophore modeling

The original concept of the pharmacophore was developed by Paul Ehrlich during the late 1800s [40].pharmacophore has become a valuable tool in drug discovery. A pharmacophore feature can be attributed to any atom or group in a molecule that exhibits distinctive properties related to molecular recognition. These features can be classified as hydrogen bond donors or acceptors, cationic, anionic, aromatic, or hydrophobic, or a combination of these [41].

By comparing different molecules at the pharmacophore level, one can create "pharmacophore fingerprints." Sometimes, when only a few pharmacophore features are considered in a three-dimensional model, the pharmacophore is referred to as a "query."

2.3.4 Virtual Screening

The virtual screening method is a computational technique that involves a comparative analysis of various lead compounds to identify new drug candidates with higher affinity for a biological target. Docking, the most commonly used technique, involves the calculation of interaction energies between a molecule and a biological target to predict the inhibitory activity of the drug.

Molecular dynamics simulations are often employed to calculate interaction-free energies and inhibition constants[42], which can be experimentally verified. Artificial intelligence techniques play a crucial role in the systematic analysis of large datasets[43]. In addition, virtual tissue and organ models offer an elegant solution to anticipate physiological and neural phenomena, enabling a more accurate prediction of drug efficacy and toxicity.

2.4 Molecular Representations

molecular representations are formats that are readable by both humans and machines, these formats are used in order to portray the structure and proprieties of molecules in order for them to be used as data in various machine models for computer-aided drug design, there are many ways to represent molecular data because each representation has its own advantages and limitations, and different representations may be better suited for different types of analyses and applications An overview of the different classes of molecular representation can be seen in Figure.2.2. each representation is useful for different purposes and contexts. each representation emphasizes different aspects of the molecule's structure and can be useful for different types of analyses and applications. a chemist may use three-dimensional models to visualize the structural arrangement of atoms in the molecule, which can help detect and predict it is activity or interaction with other molecules, then again a computer-based cheminformatics analysis may require a simplified, one-dimensional representation of the molecule's structure to use it as data in computer models

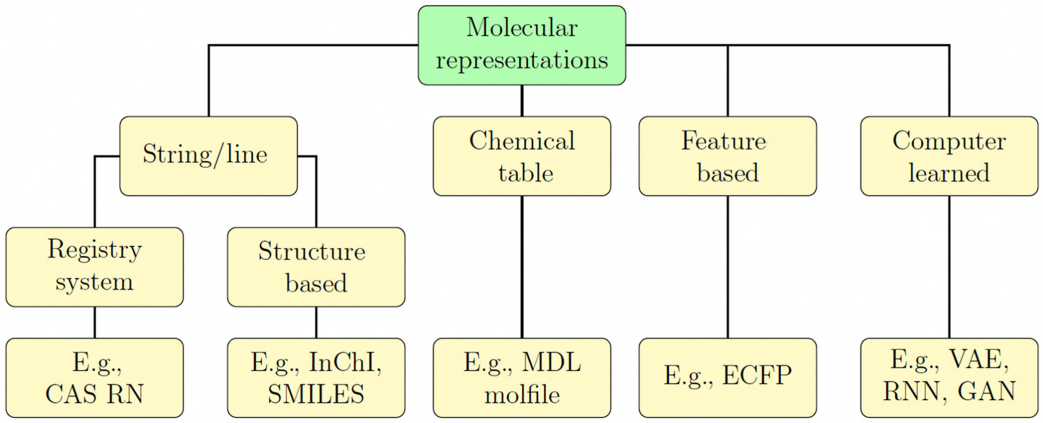


Figure 2.2: Overview of the different classes of molecular representations [2]

another reason for the diversity of molecular representations is the historical development of chemistry, When computers were first commercialized, strings of alphanumerical characters were preferred, as these required less memory to store and less computational power to process, but as computers developed and memory/processing power became less expensive, less compact representations (that are more flexible and less ambiguous) became more widespread [2]

2.4.1 String Representation

String representations generally consist of characters from the American Standard Code for Information Interchange (ASCII) character encoding standard [2]. string representation, Smiles, has powered cheminformatics since the late 1980s. it describes the molecule using a series of characters or symbols that represent the atoms and bonds within the molecule Figure.2.3

2.4.1.1 Registry system

Registry system representation uses distinct identifiers to represent chemical structures; these identifiers are assigned by a central authority like the Chemical Abstracts Service or the PubChem Compound Database and are frequently based on a combination of

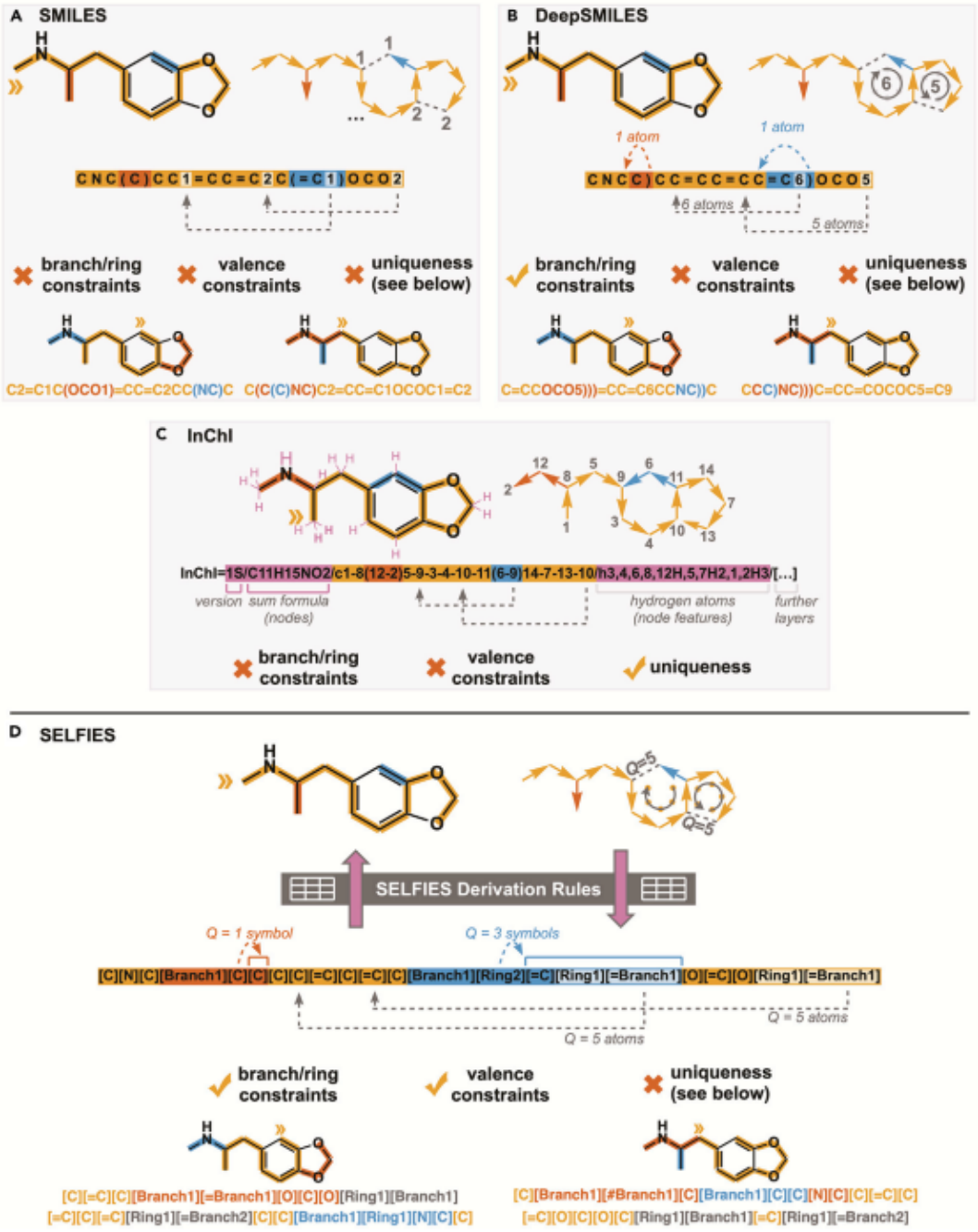


Figure 2.3: Molecular string representations [3]

molecular properties like molecular weight. This representation makes sure that every molecule has a distinct identifier that can be used to retrieve information about the molecule from a database. Examples include the CAS Registry Number(RN), PubChem CID, ChemSpider, ChEMBL, and others

2.4.1.2 Simplified molecular-input line-entry system SMILE

smiles uses a simple syntax to encode a molecule's atoms, bonds, and connectivity information as a linear string of characters. atoms are represented by their element symbols, and bonds are represented by various characters such as '-' (single bond), '=' (double bond), '' (triple bond), and ':' (aromatic bond). The connectivity between atoms is represented by the order in which the atoms and bonds are listed in the SMILES string.

the SMILES string for Aspirin Figure.2.4 is CC(=O)OC1=CC=CC=C1C(=O)O. Although ambiguity may occur in that a compound may be represented in more than one way using SMILES (generic SMILES), we use a normalization algorithm to ensure that one single SMILES representation is derived from one compound (this approach is also known as unique SMILES, canonical SMILES, or absolute SMILES) [4]

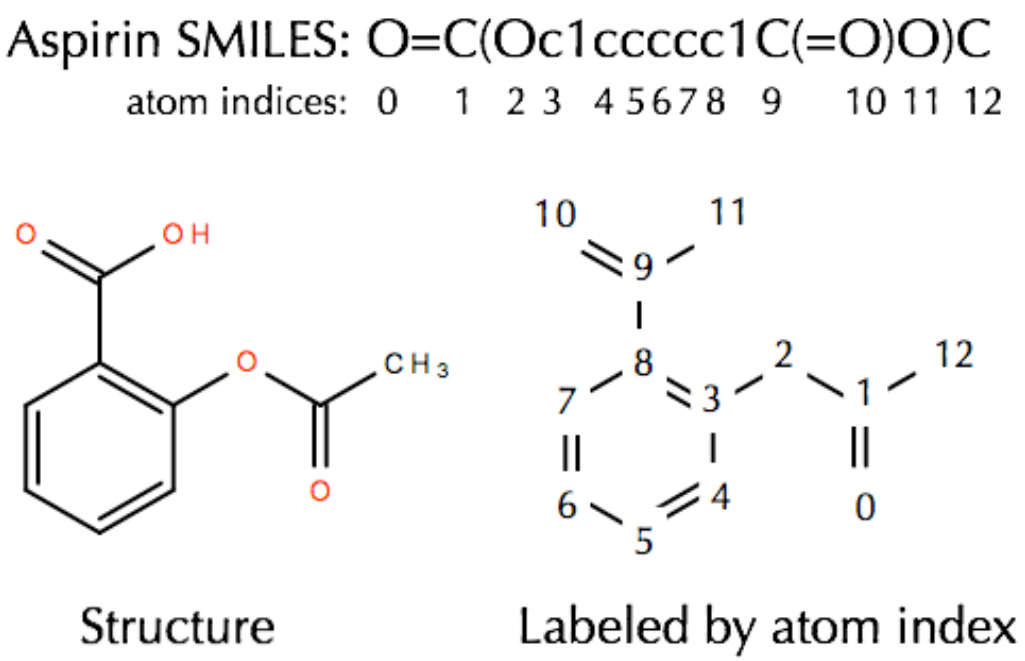


Figure 2.4: Aspirin SMILES representation [4]

2.4.2 Chemical Table Representations

In a connection table (CTab), each atom's x, y, and z coordinates are listed along with their bonds to other atoms in a chemical table (CT). This makes it simple to create 2D/3D graphic representations from a CT, and they are frequently used to represent molecules in databases and software.[2]

The most commonly used chemical table is the MDL molfile Figure.2.5, there are 2 versions of MDL V2000 and v3000M, it consist of three main section :

- Header
- Block
- CTab

and an end line which must read M END

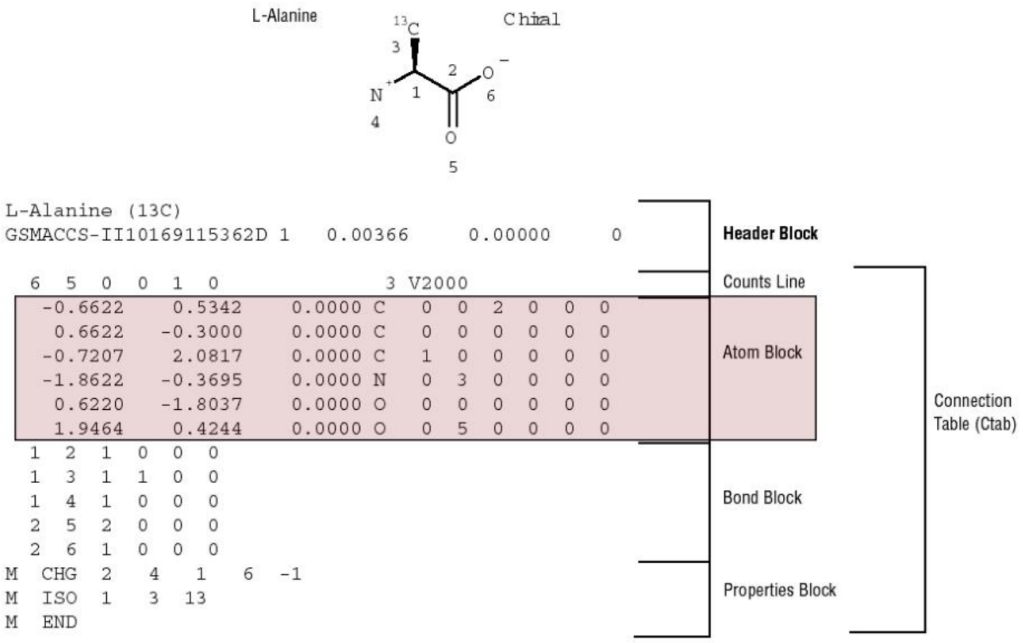


Figure 2.5: Chemical table representation MDL molfile [5]

2.4.2.1 Graph Representation

graph representation is to represent chemical structures as a graph, it represents atoms as nodes, and bonds between atoms are represented by edges. The 2D and 3D representations of graphs can easily be visualized by many software packages, including ChemDraw [44], Mercury [45], Avogadro [46], VESTA [47], PyMOL [48], and VMD [49] (the latter 5 are suitable for small- and macro-molecules, and either free or open-source).

Example graph representation Figure.2.6 and Figure.2.7 for acetic acid. a Graph representation of acetic acid with nodes numbered from one to four. b Example adjacency matrix, A, for an acetic acid graph with the corresponding node ordering on the left. c Example node features matrix, X, which one-hot encodes a few selected properties. d Example edge features matrix, E, where each edge feature vector is a one-hot encoding of single, double, or triple bonds. "Implicit Hs" stands for the number of implicit hydrogens on a given node [5]

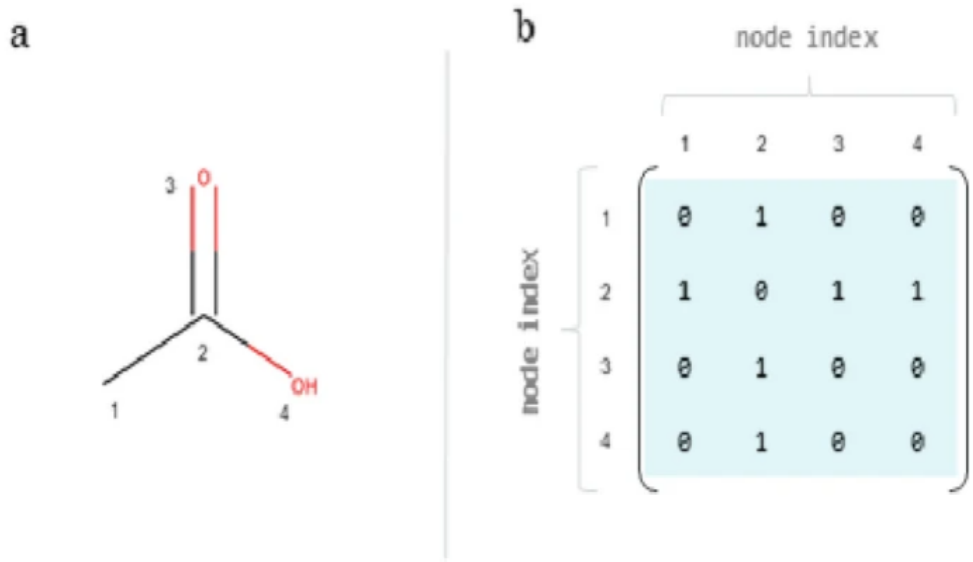


Figure 2.6: acetic acid and adjacency matrix Graph Representation [5]



Figure 2.7: node features matrix and edge features matrix Graph Representation [5]

2.4.2.2 Molecular fingerprints

The basis of molecular fingerprints is the encoding of a molecule's structural data into fixed-length binary or integer (mainly binary) vectors. They depict molecules as a collection of numerical descriptors that are simple for computer algorithms to process. There are several types of molecular fingerprints :

- Circular fingerprints: These are based on a circular substructure of the molecule,

and are created by hashing the substructure into a fixed-length bit string.

- Extended-connectivity fingerprints: These are based on the connectivity of the atoms in the molecule, and are created by recursively generating a set of subgraphs around each atom in the molecule.

- MACCS keys: These are based on a set of predefined substructures, and are created by encoding the presence or absence of each substructure into a fixed-length bit string.

Molecular fingerprints are useful for a wide range of applications, including virtual screening, similarity searching, and machine learning models for predicting various properties of molecules.

A representation of a hypothetical 10-bit topological fingerprint Figure.2.8, in this case a linear path-based fingerprint with fragments up to a length of 5. All fragments found from the starting atom (circled) are shown [6]

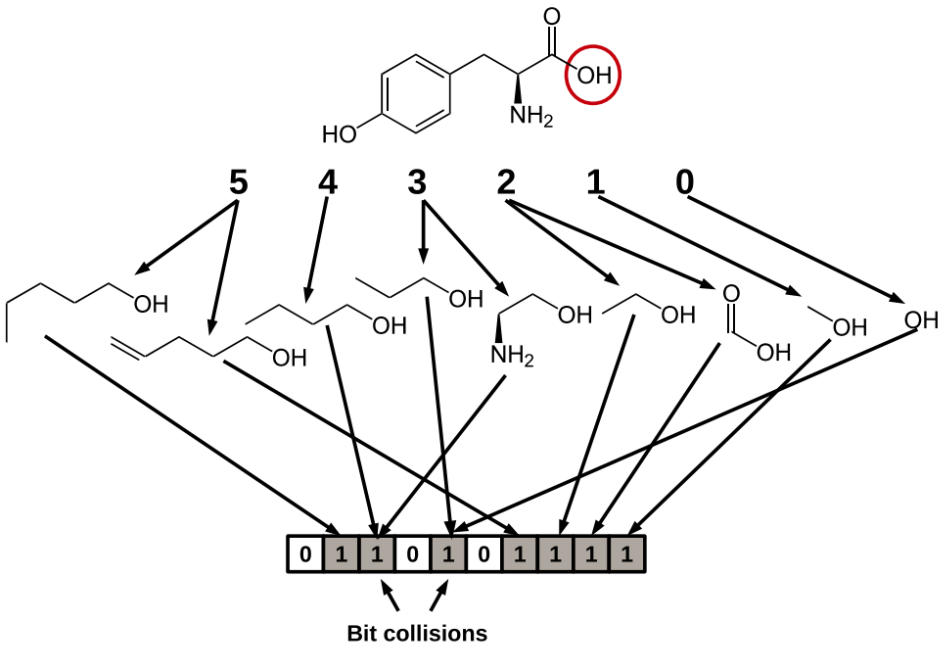


Figure 2.8: representation of a hypothetical fingerprint [6]

2.5 Molecular similarity measurement

The method of measuring the degree of resemblance based on the structural characteristics between two molecules or a set of molecules is known as molecular similarity measurement. similarities typically are taken to be real numbers that lie on the unit interval [0,1]. However, since similarity is an inherently "fuzzy" concept, it may be appropriate to take a fuzzier view [50]. Chemical informatics and computational chemistry frequently compare and categorize molecules according to their chemical and biological characteristics.

No	Coefficient	Formula
1	Jaccard/Tanimoto	$\frac{\sum_{i=1}^{N} W_{iA}W_{iB}}{\sum_{i=1}^{N}(W_{iA})^2 + \sum_{i=1}^{N} W_{iB}^2 - \sum_{i=1}^{N} W_{iA}W_{iB}}$
2	Dice	$Sim_{AB} = \frac{2\sum_{i=1}^{N} W_{iA}W_{iB}}{\sum_{i=1}^{N}(W_{iA})^2 + \sum_{i=1}^{N}(W_{iB})^2}$
3	Russell/Rao	$RR = \frac{a}{n}$
4	Simple matching	$SMC = \frac{M_{00} + M_{11}}{M_{00} + M_{01} + M_{10} + M_{11}}$
5	Hamann	$H = \frac{(N_{11} + N_{00}) - (N_{10} + N_{01})}{N}$
6	Forbes	$F = \frac{aN}{(a+b)(a+c)}$
7	Manhattan	$d_M(X,Y) = \sum_{i=1}^{n} \lvert x_i - y_i \rvert$
8	Ochiai/Cosine	$O(X,Y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$
9	Fossum	$Fossum\, coefficient = \frac{n(a-0.5)^2}{(a+b)(a+c)}$
10	Szymkiewicz–Simpson	$S(A,B) = \frac{\lvert A \cap B \rvert}{min(\lvert A \rvert, \lvert B \rvert)}$
11	Dennis	$DNC(a,b) = \frac{(a \cap b) \times [(\neg a) \cap (\neg b)] - [(\neg a) \cap b] \times [a \cap (\neg b)]}{\sqrt{n \times [(a \cap b) + [(\neg a) \cap b]] \times [(a \cap b) + [a \cap (\neg b)]}}$

Table 2.1: Similarity measurement table.

2.6 Molecular Biological Activity

In the context of molecular biology, the term "activity" can refer to a wide range of processes, such as enzymatic activity, gene expression, signal transduction, and metabolic pathways. For example, the activity of an enzyme is determined by its ability to catalyze a specific biochemical reaction, while the activity of a gene is determined by its ability to produce a functional protein. The biological activity of a molecular entity is defined as $A = cf$, where A is the activity, c the amount-of-substance concentration, and f is a parameter designated as "inherent activity." [51]

Inherent activity refers to the activity of a substance or molecule-based solely on its chemical structure and properties,

Determining the inherent activity of a target molecule is an important step in drug discovery and development, as it can help identify molecules that have the potential to produce a desired biological effect. This can be done using various experimental as computational methods that predict the activity of a molecule based on its chemical structure.

2.6.1 molecular structure-activity relationship

The relationship between a compound's molecular structure and its properties is a fundamental principle in chemistry. The molecular structure of a compound refers to the arrangement of atoms and bonds within its molecule. This structure can have a profound effect on the chemical and physical properties of the compound.

For example, the functional groups in a molecule can determine its reactivity, solubility, and biological activity. The presence of double or triple bonds can affect the compound's polarity, boiling point, and melting point. Additionally, the stereochemistry of a compound can play a crucial role in its biological activity, as it can determine the molecule's ability to fit into specific biological receptors.

to make the molecular structure-activity relationship more clear, [7] study found that

the aryl ring and the amine moiety of the compounds were important for their selectivity for the β3-adrenergic receptor and their anti-obesity and anti-diabetic activity. The authors used their findings to design new aryl propanol-amine agonists with improved selectivity and activity Figure.2.9.

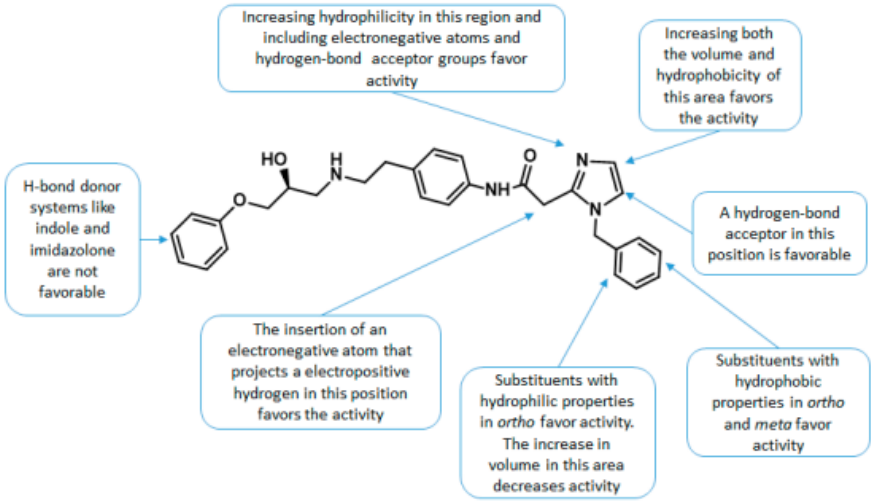


Figure 2.9: Main structure-activity relationships found in [7] study.

2.7 Artificial intelligence

Artificial intelligence (AI) has become a main topic in almost every industry and in every modern technological landscape. AI is the science of designing machines that can perform tasks that typically require human intelligence and the ability to process a large amount of data and calculations. such as understanding natural language, and recognizing images.

AI is a broad field that includes several sub-fields Figure.2.10, including machine learning, natural language processing, and robotics [8]. machine learning is a subset of AI that involves the use of algorithms to identify patterns in data and make predictions or decisions based on those patterns

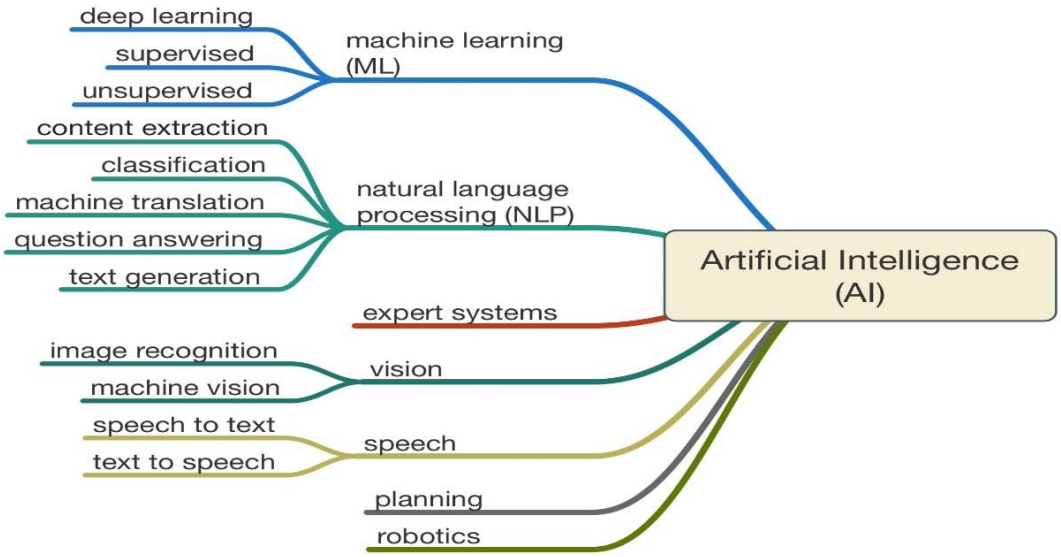


Figure 2.10: Types of Artificial intelligence [8]

2.8 Machine learning (ML)

Machine learning is a type of artificial intelligence that allows computers to learn from data without being programmed explicitly machine learning enables computers to identify patterns and insights with astonishing precision and speed resulting in numerous applications across various spheres of our lives machine learning will continue its explosive growth by becoming even more advanced and crucial than ever before.

The integration of machine learning algorithms in the field of chemistry has revolutionized drug discovery and development with the vast amount of data available, understanding how certain chemical compounds will behave can be a difficult task for scientists however machine learning algorithms can help predict the properties and activities of chemical compounds [52]. aiding in drug discovery and development by analyzing large datasets through pattern recognition techniques these algorithms can identify relationships between molecular structures and their activities which allows researchers to more accurately predict which compounds are most likely to succeed in clinical trials before they even enter them moreover machine learning algorithms provide a more systematic approach that eliminates much of the supposition traditionally involved in drug discovery while reducing production costs by identifying viable candidates faster than traditional methods. would allow as such this technology is becoming an essential tool for pharmaceutical companies seeking to create innovative drugs that benefit patients worldwide.

Databases can be used to train machine learning models which enables researchers to make rapid and accurate predictions about various chemical properties. machine learning techniques can be broadly classified as supervised or unsupervised learning, supervised learning involves training a model on labeled data that is already categorized into different classes while unsupervised learning involves identifying patterns within unlabeled data without any prior knowledge of the categories involved [30]. by utilizing cheminformatics databases for machine learning purposes scientists can leverage vast amounts of stored data to develop predictive models that are much faster and more reliable than traditional methods.

2.9 Deep learning

Deep learning is a subfield of machine learning, this subfield relies on a neural network to simulate human-like decision-making through pattern recognition and data analysis. at the core of deep learning are artificial neural networks that are designed to process and analyze data similar to how the human brain works. these networks consist of layers of interconnected nodes, each performing a specific task. The applications of deep learning are virtually limitless, as evidenced by their ability to perform tasks such as image and speech recognition, natural language processing, and autonomous driving systems [53]

The advancements in deep learning are visible in the fields of Cheminformatics, Deep learning algorithms can analyze vast amounts of chemical data and help scientists identify new compounds with potential therapeutic properties [54], Cheminformatics and deep learning have the potential to transform drug discovery by predicting drug-target interactions and accelerating the drug development process [52].

2.10 Graph neural network GNN

recently there has been a significant surge in the use of graph neural networks. This is due to their immense effectiveness in dealing with complex and massive amounts of data that are structured as graphs. These types of networks derive their name from their ability to operate on graph structures seamlessly. They work by considering the relationships between entities within the graph and utilize this information for making predictions and taking decisions. Graph neural networks are still a relatively new field of research, with many challenges remaining to be addressed [55],
graphs became a tool for presenting complex information in a condensed, visual form. however, this comes with its own set of challenges [56]. graphs can be visually overwhelming and difficult to translate accurately. the problem is in the visual complexity that arises from trying to represent multiple variables on one graph. When different types

of data are put together it can create confusion for those who are not well-versed in interpreting graphs or statistics. Additionally, there is no standardized format for creating them which compounds the challenge of interpreting them correctly. Graphs come in various forms such as line charts, scatterplots, bar charts and more making it important for individuals to understand how each type works when representing certain kinds of data. However, in spite of their initial daunting appearance, graphs cannot be dismissed because they require some effort to decipher as they have significant value when used correctly. Graphs have the ability to present vast amounts of information clearly and succinctly; researchers especially appreciate these abilities. If interpreted correctly over time, we could learn much from them about trends which would help us make informed decisions based on past experiences instead of blindly jumping into uncertainties without any prior knowledge. GNNs have the potential to revolutionize chemical discovery by enabling the efficient prediction of molecular properties and behavior. This is because existing methods for predicting chemical reaction success rely mainly on chemical properties, such as electronegativity and bond length, which limit their accuracy and scope [57]

Despite the considerable advancement achieved in enhancing the precision and complexity of graph neural network models in recent times, there are still plenty of tasks that need to be tackled for these approaches to reach their maximum potential. such as enhancing model interpretability and scalability across different domains.

2.10.0.1 Message-passing neural networks MPNN

Its general framework for supervised learning on graphs called MPNNs simply abstracts the commonalities between several of the most promising existing neural models for graph-structured data, in order to make it easier to understand the relationships between them and come up with novel variations [58].

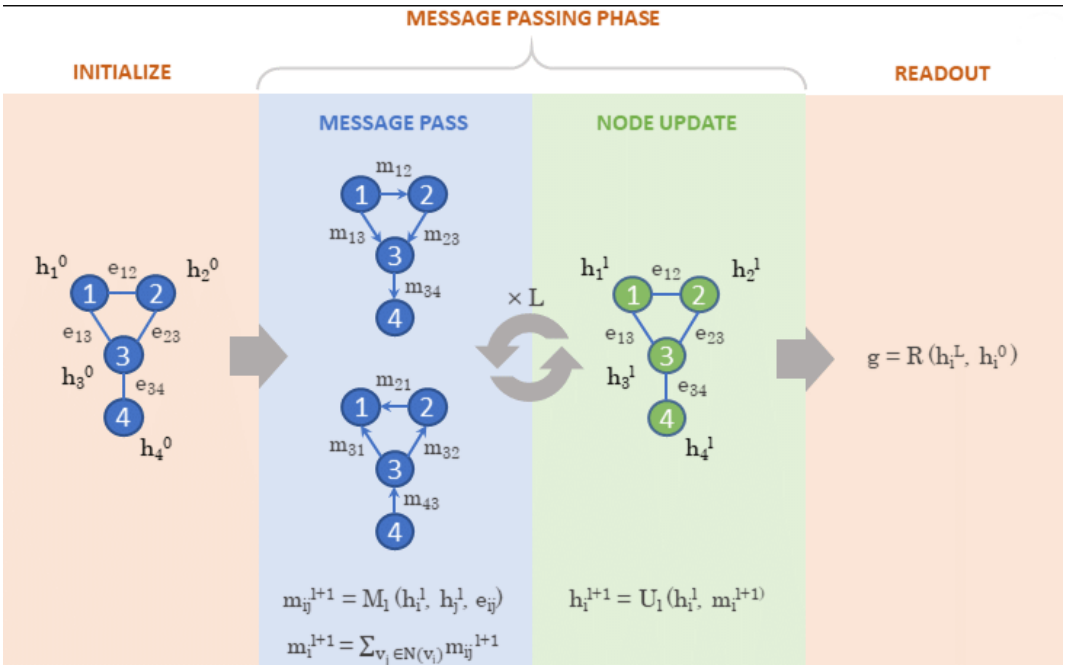


Figure 2.11: Schematic of a message passing neural network (MPNN)[9]

MPNNs operate by iteratively passing messages between nodes in a graph and updating the node representations at each step based on the received messages and their current states Figure.2.11[9].

To put it simply, MPNNs function in the following way: first, an initial set of states is created, one for each node v in the graph. Then In a message passing phase, the feature of node v is updated based on the message m of node v, which contains feature aggregation of the node v and some neighboring nodes w and edges e [59], the propagation rules in MPNNs can use “arbitrary" neighborhood aggregation functions [9]. This allows each node to become aware of its immediate surroundings. This process is repeated, with each step allowing nodes to become aware of a broader neighborhood. In the final, the readout function is needed to incorporate all node features into one feature. In this phase, all vectors from each node are summed or pooled so that the output is invariant to the initial node permutation [59].

Formally, MPNNs consist of three major operations: message passing, node update, and readout. Implementing a message-passing neural network involves iteratively updat-

ing a hidden state $h_v \in RD$ of each node v, which is accomplished through the following equations:

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

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

this message passing algorithm is repeated for a specified t step, on each step t, hidden states h_v^t and messages m_v^t associated with each vertex v are updated using message function M_t and vertex update function U_t according to where $N(v)$ is the set of neighbors of v in graph G, and h_v^0 is some function of the initial atom features [60]. After that, the process reaches the final readout phase.

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

The output $\hat{y}$ may be either a scalar or a vector, depending on whether the MPNN is designed to predict a single property or multiple properties (in a multitask setting) [60].

2.10.0.2 Graph convolutional network GCN

Graphs naturally arise in many real-world applications [61], GCN has shown great promise in efficiently processing graph-structured data. GCNs have already found practical applications in a wide range of domains, including social network analysis and drug discovery. These networks are designed to operate on graphs by leveraging learnable filters that can be convolved over the graph's nodes and edges, allowing them to capture both local and global information from the structure. The incorporation of this feature makes it possible for GCNs to identify and reveal hidden patterns that go unnoticed with traditional machine learning techniques or statistical methods.

GCN is one of the most powerful deep learning methods for non-Euclidean structure data and achieve impressive performance in many fields [62], There unique architecture allows them to provide greater accuracy than other models when analyzing complex datasets

such as molecular structures or large-scale social networks. Unlike other models, they can handle missing or incomplete data points without any drop in performance. By adapting these networks to generalize their learning skills for various graph-based tasks, researchers are making significant contributions toward finding solutions to society's most challenging problems.

As more researchers continue to explore new applications for Graph Convolutional Networks across different fields, they will undoubtedly make further strides toward advancing artificial intelligence capabilities within our contemporary world. With their potential impact on diverse areas like medicine and finance, it is evident that Graph Convolutional Networks represent a promising avenue for future research exploration into this exciting technology's capabilities.

a general architecture of GCN is shown in Figure.2.12 each node pulls the embedding from all its neighbors, calculates their sum, and passes them along with its embedding to the recurrent unit, which will produce a new embedding. This new embedding contains the information of the node plus the information of all the neighbors. In the next time step, it will also contain the information of its second-order neighbors. And so on and so on. The process continues until every node knows about all the other nodes in the graph. Each one of the embeddings has now information from all the other nodes

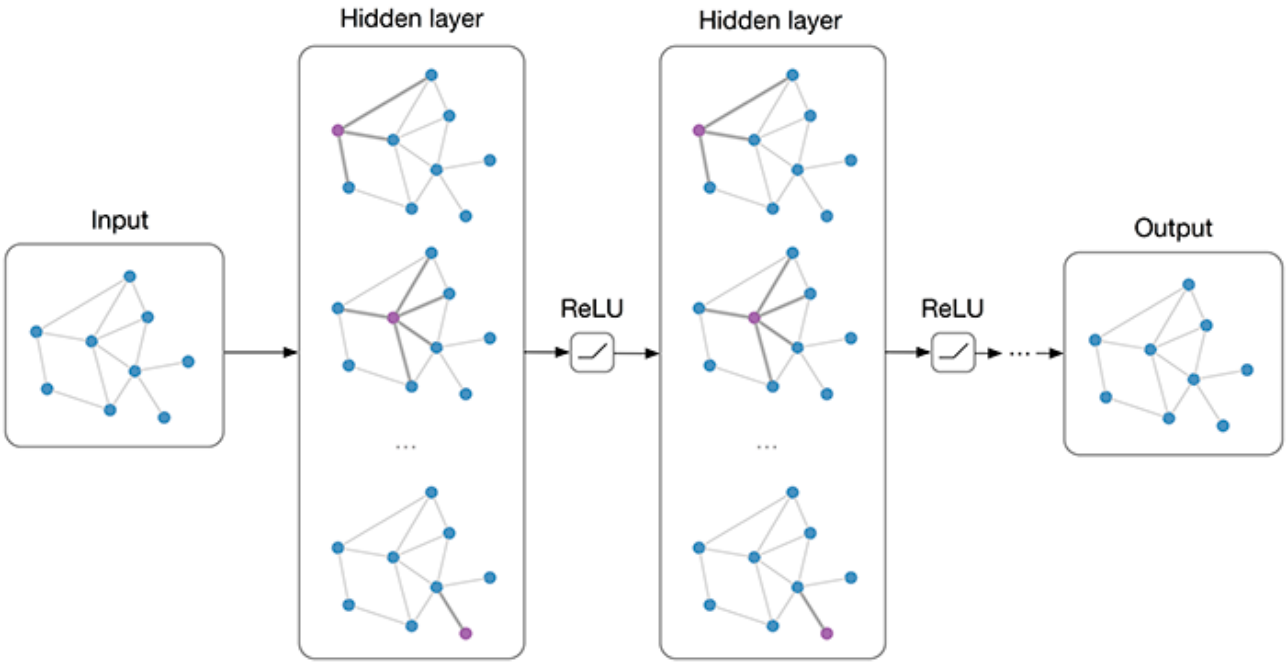


Figure 2.12: General architecture of GCN [10]

2.10.0.3 Recurrent Graph Neural Network

A Recurrent Graph Neural Network (R-GNN) is a type of neural network that can process structured data such as graphs and sequences. It is an extension of the traditional Recurrent Neural Network (RNN) architecture that is designed to handle graph-structured data.

R-GNNs operate on graph data by applying the same set of weights to every node in the graph, allowing them to encode the graph structure as well as the node features. And considers past information via a recurrent connection [63], In Recurrent Graph Neural Networks (R-GNNs), recurrent connections are used to allow the network to process graph-structured data over time, which brings the benefit of end-to-end training [63], similar to traditional RNNs.

2.10.1 Transformer Encoder

The Transformer Encoder is a neural network architecture that was introduced in [64]. It is a sequence-to-sequence model that uses self-attention to encode the input sequence and generate an output sequence such as [65].

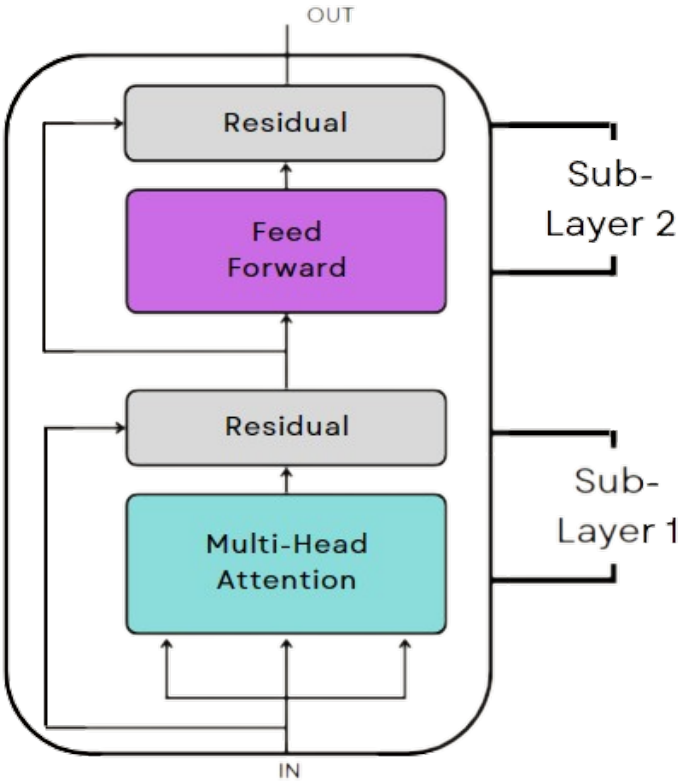


Figure 2.13: Transformer Encoder sub-layers

The Transformer Encoder consists of a stack of identical layers, each of which has two sub-layers [66] Figure.2.13: a multi-head self-attention mechanism, and a position-wise fully connected feedforward network. Each of these sub-layers has a residual connection around it, followed by layer normalization.

The multi-head self-attention mechanism Figure.2.14 allows the model to attend to different parts of the input sequence at different positions, while the position-wise feedforward network Figure.2.13 applies a non-linear transformation to each position separately.

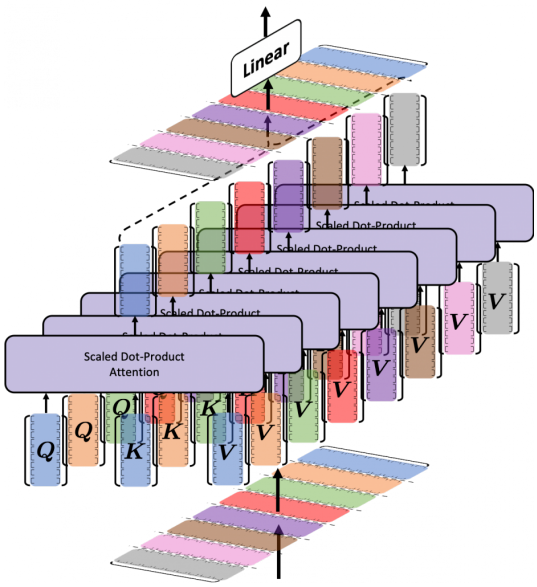


Figure 2.14: Multi-head self-attention mechanism [11]

Prior to [67], there were some attempts to use transformers in the context of graph-structured data [68], the Transformer Encoder produces node embeddings as output, where each embedding corresponds to a node in the input graph. This output differs from the output in natural language processing tasks, as it is not a sequence of hidden representations. The number of node embeddings produced is equal to the number of nodes in the input graph, and each embedding has a fixed size that depends on the number of hidden units in the model.

2.11 Related Work

- **Similarity-Based Virtual Screen Using Enhanced Siamese Multi-Layer Perceptron [69]:** The primary objective of this research is to enhance the performance of similarity searching, particularly for molecules that exhibit structural heterogeneity. To achieve this, the study proposes the utilization of a Siamese network, which has proven effective in handling complex data samples across various domains. The Siamese multi-layer perceptron architecture will be improved by incorporating two similarity distance layers along with a fused layer.

- **Deep Convolutional Neural Network to improve the performances of screening process in LBVS[70]:** This paper introduces a novel approach called DCNNLB (Deep Convolutional Neural Network for Ligand-Based Virtual Screening) to enhance the performance of the Ligand-Based Virtual Screening (LBVS) process. The main contribution of this research lies in the development of a model based on Deep Convolutional Neural Network (DCNN) specifically designed for LBVS.

- **Hybrid-Enhanced Siamese Similarity Models in Ligand-Based Virtual Screen[71]:**The paper proposes the use of hybrid-enhanced Siamese similarity models in the context of Ligand-Based Virtual Screening (LBVS). The main objective is to improve the performance of LBVS by leveraging the power of Siamese networks and incorporating enhancements to the model architecture.

  The hybrid-enhanced Siamese similarity models combine the Siamese network architecture with additional techniques or modifications to enhance the screening process. These modifications could include the integration of feature engineering methods, the incorporation of additional layers or modules in the network, or the utilization of ensemble learning techniques.

  The paper aims to demonstrate the effectiveness of these hybrid-enhanced Siamese similarity models in LBVS.

2.12 Conclusion

In summary, this literature review chapter serves as a comprehensive foundation for the upcoming chapters, providing a thorough understanding of the key concepts and techniques in cheminformatics, CADD, molecular representations, molecular similarity measurement, molecular biological activity, and deep learning in the context of drug discovery. The insights gained from this chapter will inform and support discussions and analysis throughout the remainder of the study.

Chapter 3

Research Methodology

3.1 Introduction

in This chapter, we provide a detailed description of the research framework and methodology employed in this study. The chapter begins with an overview of the four phases of the research process, highlighting the steps taken to achieve the study's objectives. It then describes the dataset used in the study. The chapter also outlines the design and steps taken in the data preprocessing phase, including the treatment of data, the conversion of SMILE strings to graphs, and the splitting of data. we also discuss the message-passing mechanism used in the study to capture the relationships between atoms in molecules. and applying message-passing neural networks for model training.

then we talk about the performance evaluation using ligand-based virtual screening, Overall, this chapter provides a detailed description of the proposed methodology.

3.2 Research Framework

in this section we will be taking a look at the process that we followed in order to realize this work, The study methods will be conducted according to the illustrated workflow process represented in Figure.3.1

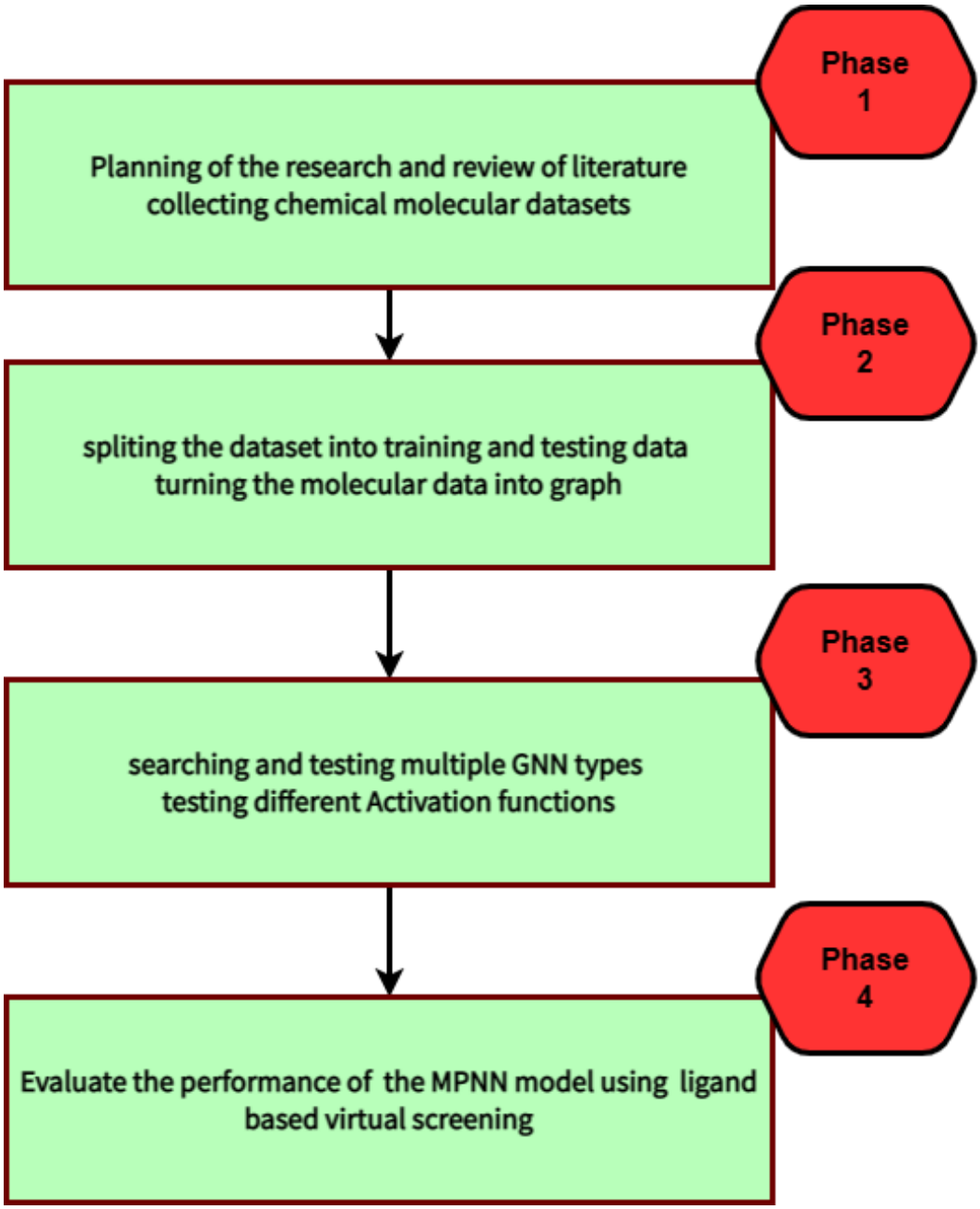


Figure 3.1: Flow Chart of the Framework

3.2.1 Phase 1

To plan our research, we first defined our research questions and objectives. We found the problems that we wanted to address and the goals we hoped to achieve through our study. Then we conducted a literature review by merging the information we found from various sources that we had collected and found the gaps in the existing literature that our study would address. This helped us in developing a research framework that would make a meaningful contribution to the field.

after that, we collected the molecular data set that we needed by identifying and gathering appropriate datasets from various sources that include information on chemical compounds, their molecular structures, and their proprieties.

3.2.2 Phase 2

in this phase we split the data into training and testing data for our model, after that we Defined features To encode features for atoms and bonds to preprocess the data and generated complete graphs from SMILES for our GNN model to learn from.

3.2.3 Phase 3

in this phase, the key task was for us to explore and experiment with different models and activation functions to find the best possible solution for the problem at hand. This involves searching through various GNN architectures, such as Graph Convolutional Networks GCN[61], Message Passing Neural MPNN[72], and Recurrent Graph Neural Network RGNN[73], Each model has its own unique properties and strengths that need to be evaluated against the specific task requirements. Additionally, it is essential to test different activation functions for each architecture, such as ReLU, SoftMax, and Sigmoid, to determine which one provides the best performance for the chosen model

3.2.4 Phase 4

in this last phase, we test the capacity of the MPNN model using ligand-based virtual screening

3.3 Dataset

there are many data sets that could test or train our model but the one that gave promising results was still The MDDR (MDL Drug Data Report) [74] data set which is a comprehensive database of over 102,000 chemical compounds with hundreds of different activities, some of which are related to therapeutic areas such as antihypertensive, while others are related to non-therapeutic areas such as herbicides [75]. The MDDR dataset is also used by many literature to validate ligand-based virtual screening approaches

The MDDR database contains three data sets: DS1, DS2, and DS3[76]

3.3.1 DS1

The MDDR-DS1 is a database that consists of thousands of chemical compounds with 8294 different activitie molecules. It involves eleven activity classes, some of which are involved in activities that are homogenous in structure, whereas others are involved in activities that are heterogeneous in structure (structurally diverse) s shown in Table 3.1

Activity Index	Activity Class	Active Molecules	Pairwise Similarity
31420	Renin Inhibitors	1130	0.290
37110	Thormbin Inhibitors	803	0.198
715523	HIV Protease Inhibitors	750	0.180
31432	Angiotensin II AT1 antagonists	943	0.229
42731	Substance P antagonists	1246	0.149
06233	Substance P antagonists	752	0.140
06245	5HT Reuptake Inhibitors	359	0.122
07701	D2 antagonists	395	0.138
06235	5HT1A antaonists	827	0.133
78374	Protein Kinase C inhibitors	453	0.120
78331	Cyclooxygenase Inhibitors	636	0.108

Table 3.1: The MDDR-DS1 structure activity classes

3.3.2 DS2

The MDDR-DS2 dataset comprises 10 homogeneous classes of activity, it contains 5033 active molecules divided over each class as shown in Table 3.2

Activity Index	Activity Class	Active Molecules	Pairwise Similarity
07707	Adenosine A1 agonists	207	0.229
07708	Adenosine A2 agonists	156	0.305
31420	Renin Inhibitors 1	1300	0.290
42710	CCK agonists	111	0.361
64100	Monocyclic-lactams	1346	0.336
64200	Cephalosporins	113	0.322
64220	Carbacephems	1051	0.269
64500	Carbapenems	126	0.260
643500	Tribactams	388	0.305
64350	Vitamin D analogous	455	0.386

Table 3.2: The MDDR-DS2 structure activity classes

3.3.3 DS3

The MDDR-DS3 dataset contains 10 homogeneous classes of activity, it has 8568 active molecules, and it also has the largest number of active molecules in one class in the Lipoxygenase Inhibitor's Activity class s shown in Table 3.3

Activity Index	Activity Class	Active Molecules	Pairwise Similarity
09249	Muscarinie M1 agonists	900	0.111
12455	NMDA receptor antagonists	1400	0.098
12464	Nitrice oxide synthase Inhibitors	505	0.102
31281	Dopamine-hydroxylase Inhibitors	106	0.125
43210	Aldose reductase Inhibitors	957	0.119
71522	Reverse transcriptase Inhibitors	700	0.103
75721	Aromatase Inhibitors	636	0.110
78331	Cyclooxygenase Inhibitors	636	0.108
78348	Phospholipase A2 Inhibitors	617	0.123
78351	Lipoxygenase Inhibitors	2111	0.113

Table 3.3: The MDDR-DS3 structure activity classes

3.4 Design and steps

in this section we will be presenting the process we took in order to create our model from treating the data to training our Model, the following chart Figure.3.2 will give a general idea on the major approaches that we took.

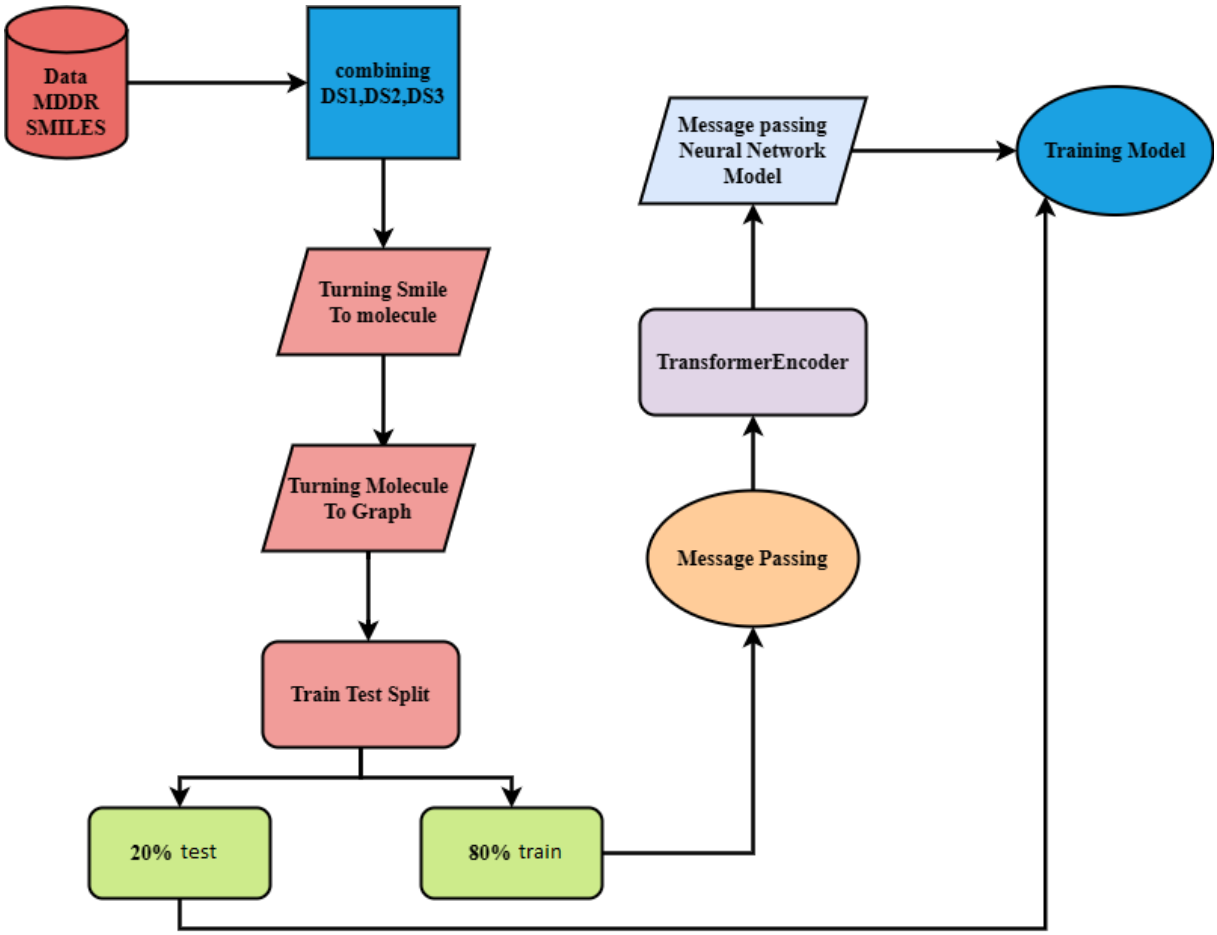


Figure 3.2: Chart that represents the Design steps

3.4.1 Loading and Preprocessing Data

the first step we took while treating the data is to combine the 3 datasets of DS1, DS2, and DS3 this is a very necessary part in prepossessing for our model to be a multiclass model with each class representing one of the classes in the DS1, DS2, and DS3. with DS1 having 11 class and both DS2 and DS3 having 10 classes this will leave us with 31 classes.

the datasets are stored in a text file with each row representing a molecular SMILE structure which will leave us with 3 text files one for each DS. The text file's content is represented in a Dataframe which will show the order of the classes and the decoys. The following tables represent the Dataframes of each DS Table 3.4 Table 3.5 Table 3.6 .

in each Dataframe the first row has one column that represents the number of classes in the DS the following rows have three columns, the first two represent the beginning and the end respectively of the class in the DS Text file the third column represents the number of molecules in that class, the rest of the lines in the text files represent decoys.

After extracting the data from the DS1, DS2, and DS3 files we combined the molecular SMILES structure of all the DS and sorted them in one Dataframe Table 3.7, this will make our model a multiclass model with 31 classes and 21945 to learn from.

11		
1	1130	1130
1131	1880	750
1881	2683	803
2684	3626	943
3627	4872	1246
4873	5624	752
5625	5983	359
5984	6378	395
6379	7205	827
7206	7658	453
7659	8294	636

Table 3.4: DS1 Dataframe

10		
1	207	207
208	363	156
364	1493	1130
1494	1604	111
1605	2905	1301
2906	3063	158
3064	4114	1051
4115	4240	126
4241	4628	388
4629	5083	455

Table 3.5: DS2 Dataframe

10		
1	900	900
901	2300	1400
2301	2805	505
2806	2911	106
2912	3868	957
3869	4568	700
4569	5204	636
5205	5840	636
5841	6457	617
6458	8568	2111

Table 3.6: DS3 Dataframe

Number	Class	SMILE
0	0	CC[C@H](C)[C@H](NC(=O)C[C@H](O)[C@H]...
1	0	CC(C)C[C@@H](NC(=O)C\C=C[C@@H](Cc1c...
.	.	.
1130	1	COc1ccc(CN[C@H]([C@H](O)[C@H](Cc2ccccc2)...
.	.	.
4001	4	Cl.OC(CCc1cc(cc(c1)C(F)(F)F)C(F)(F)F)C2(C...
.	.	.
10015	15	[Na+].CN(C)C(=O)OCC1=C(N2C(SC1)C(NC(=O)...
.	.	.
21944	30	CC(C)C[C@H](O)[C@H](O)[C@H](CC1CCCCC1...

Table 3.7: MDDR Classes DataFrame

3.4.2 SMILE to Graph

The data that we will train our Graph neural network with must be in the form of graphs, in order to realize that our data must go through two functions, molecule from smiles and graph from molecule.

3.4.2.1 molecule from smile

This function that takes a SMILES string as input and returns a molecule object, The function first attempts to create a molecule object from the SMILES string without sanitizing it, using the Chem.MolFromSmiles function from the RDKit library. Sanitization is the process of correcting any inconsistencies or errors in the molecule representation. If sanitization is successful, the molecule object is returned.

If sanitization fails, the function catches the error and tries again, but this time without the sanitization step that caused the error. The Chem.SanitizeMol function is called with the catchErrors parameter set to True to enable error catching. The sanitizeOps parameter is used to specify which sanitization operations to perform. The ^ operator is the bitwise XOR operator, which is used to remove the flag that caused the error from the set of sanitization operations to be performed.

Finally, the Chem.AssignStereochemistry function is called with the cleanIt and force parameters set to True to assign stereochemistry to the molecule.

This function will take a input as a SMILE for example (CC(=O)Oc1ccccc1C(O)=O) and return it as a molecule object Figure.3.3.

3.4.2.2 Graph from molecule

before using the molecule from smile function we must first define three other classes that we will be needing. Featurizer,AtomFeaturizer and BondFeaturizer.

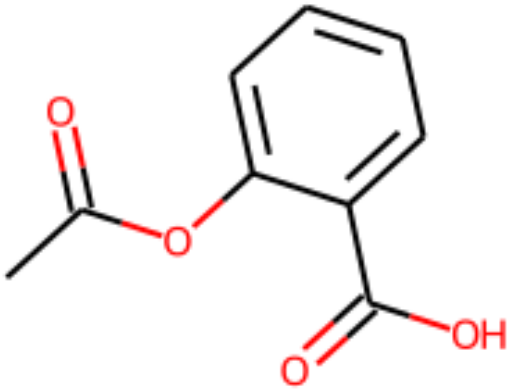


Figure 3.3: Molecule Object

Featurizer The Featurizer class can be used to create feature encodings for inputs based on allowable sets of features.

The __init__ method takes in a dictionary allowable_sets, where the keys represent the names of feature sets, and the values are lists of allowable feature values for each set. For each feature set, the method creates a mapping from the allowable feature values to a range of indices, and stores the mapping in an instance variable features_mapping. The total number of indices is also stored in dim.

The encode method takes in an input inputs, and creates a feature encoding for it as a one-hot vector of length dim. For each feature set, the method retrieves the corresponding feature value from the input using a method with the same name as the feature set, and looks up the corresponding index in the features_mapping dictionary. If the feature value is not in the allowable set, the method skips that feature set. Otherwise, it sets the corresponding index in the output vector to 1.0.

AtomFeaturizer The AtomFeaturizer class defines four methods that correspond to four different feature sets for atoms: symbol, n_valence, n_hydrogens, and hybridization. Each method takes an atom object as input and returns a value corresponding to that feature set. These methods are used in the encode method of the parent class to generate a feature vector for an atom.

BondFeaturizer The BondFeaturizer class defines three methods that correspond to three different feature sets for bonds: bond_type, conjugated, and encode. The bond_type and conjugated methods take a bond object as input and return a value corresponding to that feature set. The encode method is overridden from the parent class to add an additional feature to the feature vector, representing whether the bond is None or not.

Molecule_To_Graph Now we can use the three function in the molecule to graph function The function first initializes three empty lists atom_features, bond_features, and pair_indices. These lists will be used to store the features of each atom, each bond, and each pair of atoms respectively.

The function then iterates over each atom in the molecule using the GetAtoms() method of the molecule object. For each atom, it encodes the atom's features using the atom_featurizer.encode(), The resulting feature vector is appended to the atom_features list.

Next, the function adds self-loops to the graph by appending the index of the current atom twice to the pair_indices list, and encoding a null bond using the bond_featurizer.encode(None) function. This ensures that each atom is connected to itself in the graph.

Finally, for each neighbor of the current atom, the function gets the bond between the current atom and its neighbor using the Get_Bond_BetweenAtoms() method of the molecule object. It then appends the indices of the current atom and its neighbor to the pair_indices list, and encodes the bond using the bond_featurizer.encode() function. The resulting feature vector is appended to the bond_features list.

The function returns the three lists as numpy arrays: atom_features, bond_features, and pair_indices.

this function will take as a input a molecule object Fig(3.3) and returns a Graph representation of the molecule in the form of numpy arrays . example for the Fig(3.3) molecule object:

Graph (including self-loops):

- atom features (13, 29)
- bond features (39, 7)
- pair indices (39, 2)

3.4.3 Splitting the Data

we shuffle the data in Tabel(**??**) then we split it into a training set and a validation set. The first 80% of the shuffled indices are used to select the rows of the input data that will be used for training, while the remaining 20% are used for validation. The SMILES strings in the selected rows of the input data are then converted into molecular graphs for both the training and validation sets. Finally, the classification target variable in the selected rows of the input data is assigned to the corresponding training and validation sets.

3.4.4 Message Passing

The message-passing step itself consists of two parts:

EdgeNetwork defines a layer that takes the atom features, bond features, and pair indices as input, applies a linear transformation to the bond features, aggregates information from the neighbors of each atom, and returns the updated atom features.

this function passes messages from 1-hop neighbors w_i of v to v, based on the edge features between them (e_vw_i), resulting in an updated node (state) v'. w_i denotes the i th neighbor of v.

Gated recurrent unit(GRU) The gated recurrent unit (GRU), which takes as input the most recent node state and updates it based on previous node states. In other words, the most recent node state serves as the input to the GRU, while the previous node states

are incorporated within the memory state of the GRU. This allows information to travel from one node state (e.g., v) to another (e.g., v").

Message passing defines a layer that performs a number of steps of message passing by applying the EdgeNetwork layer and updating the node state via a step of GRU for each step. after that, it determines the number of units used for the update step of the GRU, which can be padded with zeros if necessary to match the dimensionality of the atom features.

3.4.5 Ttransformer Encoder

In the Transformer Encoder class we define two layers of our neural network model.

The PartitionPadding which is designed to handle input data that represents atom features of molecules. In some cases, molecules have varying numbers of atoms, and this layer partitions the atom features based on a molecule indicator and pads them to have the same length within each partition. This process ensures that the input data is compatible with subsequent operations in the neural network.

the TransformerEncoderReadout layer takes the inputs and passes them through the partition_padding layer to partition and pad the data. It then creates a padding mask by checking for non-zero values in the padded data. This mask is used to mask out padded elements during the attention calculation.

3.4.6 MPNN Model

to construct a neural network model for Message Passing Neural Networks (MPNNs). The model takes atom features, bond features, pair indices, and molecule indicators as inputs.

The function first creates input layers for atom features, bond features, pair indices, and molecule indicators.

Next, the input data is passed through a MessagePassing layer, which performs message-passing iterations, allowing atoms to exchange information with their neighboring atoms through bonds. This layer updates the atom features based on the bond features and pair indices.

The output of the message passing layer is then passed through a TransformerEncoder-Readout layer. This layer applies transformer-based encoding and readout operations to capture global information from the atom features. It aggregates the updated features and produces a transformed output.

The transformed output is further processed by a fully connected layer with ReLU activation, Finally, a dense layer with softmax activation is applied to the intermediate output to obtain the final output probabilities for each class.

The function returns two models: model and model2. model takes the atom features, bond features, pair indices, and molecule indicators as inputs and produces the final output probabilities. model2 takes the same inputs but produces the intermediate output obtained after the fully connected layer, which we will be needing in our LBVS Evaluation. Figure.3.4 will give a good representation of the model.

3.4.7 Training the model

To assess the model's performance during training and monitor its ability to generalize, the test dataset is utilized for validation purposes. This dataset serves as a benchmark for evaluating the model's accuracy and detecting potential overfitting.

The training process is divided into a fixed number of iterations known as epochs. In this scenario, the model undergoes 50 epochs, meaning that the entire training dataset is processed 50 times. During each epoch, the model's internal parameters, also known as weights, are adjusted based on calculated loss values, aiming to optimize the model's

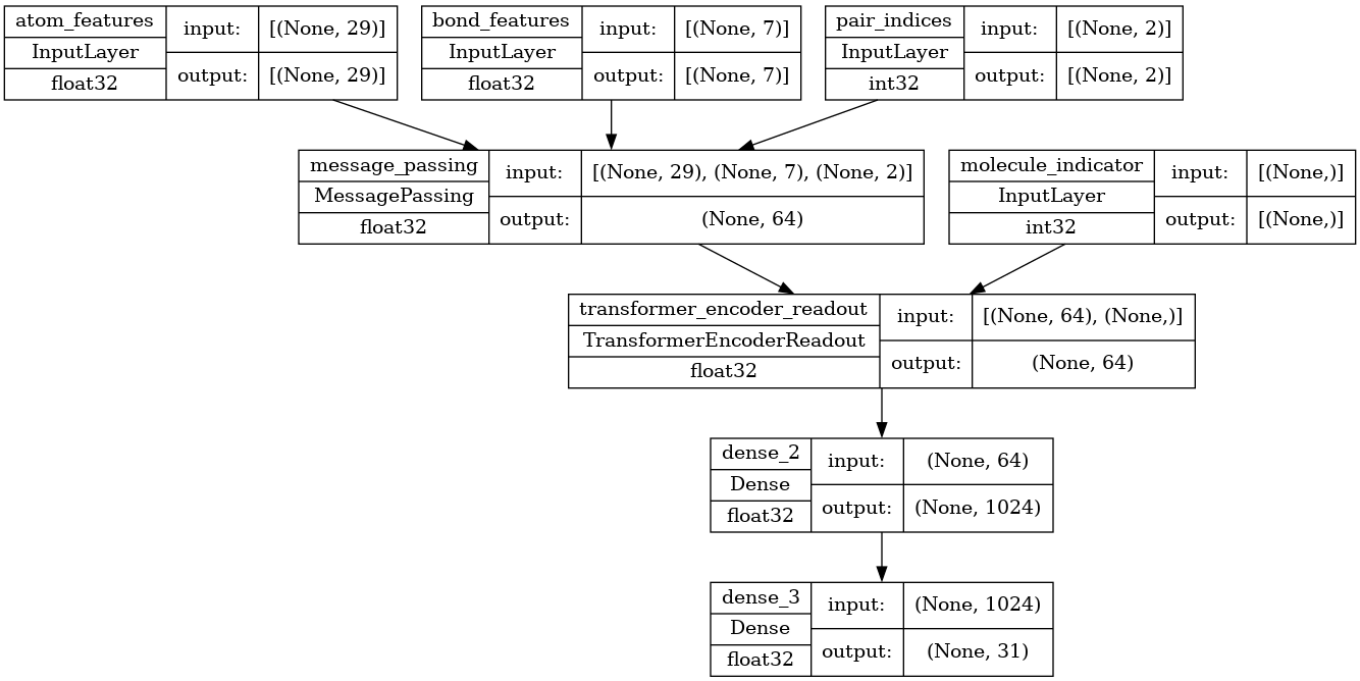


Figure 3.4: MPNN model Plot

performance.

The verbose argument is set to 2, which enables the display of detailed information throughout the training process. This information includes metrics such as loss, which represents the discrepancy between predicted and actual values and assists in monitoring the model's progression over individual data batches within each epoch.

3.5 Performance Evaluation

3.5.1 Ligand-based virtual screening

Ligand-based virtual screening is a computational technique used to predict if a new molecule is likely to bind to a target protein based on the knowledge of the ligands that bind successfully to the target. This technique takes advantage of the information of a known active ligand to identify the most promising peptide hits. Ligand-based virtual screening can be used to reduce the chemical space of large databases to a manageable size where chosen ligands can be experimentally tested.

we evaluated the MPNN model by testing how well it does with LBVS in different activity classes, we select a query of 10 active molecules from each activity class, and we calculate the similarity between each molecule from the query and every molecule from our total mddr dataset using the cosine similarity as shown in Figure.2.1,

we ranked the similarity results in descending order where we only took the first 1% and 5% of the results. to put this work into numbers we used the recall or true positive rate which calculates the percentage of data samples correctly labeled.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{3.1}$$

3.6 Conclusion

In conclusion, this chapter on research methodology provided a comprehensive overview of the approach and techniques employed in the study, in which we discussed the dataset used in the research how we loaded, preprocess it and use it in our model. then we talked about the design and steps we took in our model as well as the techniques we used to calculate the similarity and recall, and finally we talked about the LBVS technique where we test how effective our model was.

Chapter 4

Results and Discussion

4.1 Introduction

in This chapter, we will talk about the findings of our research on the application of GNN-based models for LBVS. We evaluated the performance of our approach on MDDR-DS1, MDDR-DS2, and MDDR-DS3 datasets and demonstrated the effectiveness of transfer learning in improving the model's accuracy on smaller datasets. Our study also identified some limitations and suggested potential directions for future research.

Overall, our proposed MPNN framework has the potential to enhance the efficiency and accuracy of LBVS, and our research outcomes can inform the development of more effective drug discovery pipelines. Our findings hold important implications for developing more efficient and accurate LBVS methods, which can accelerate the drug discovery process and lead to the discovery of new therapeutics.

4.2 Results and Discussion

4.2.1 MPNN Model

The MPNN model demonstrated strong performance in the multiclass classification task, achieving an overall accuracy[77] of **98%** Figure.4.1. This indicates that the model was

able to effectively classify molecules into their respective activity categories, considering the diverse set of activities present in the dataset.

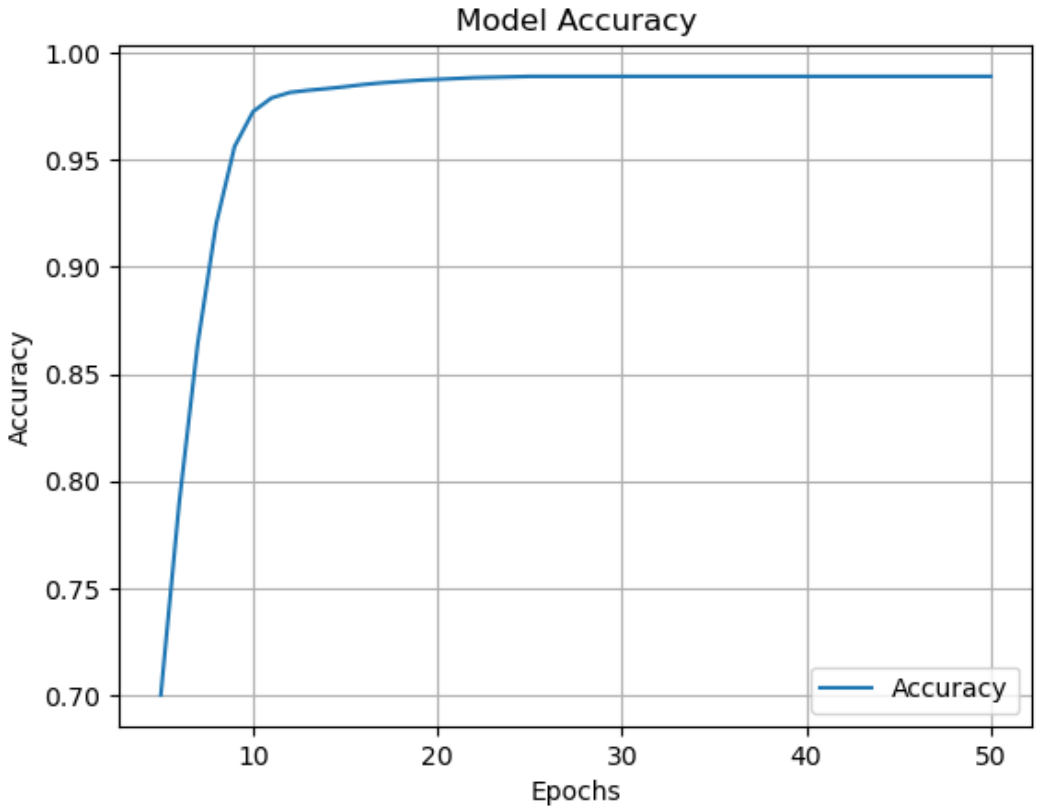


Figure 4.1: MPNN Model accuracy chart

The loss[77] value of **0.184** Figure.4.2 indicates the performance of our MPNN model during training. Loss is a measure of the discrepancy between the predicted output of the model and the true labels. In this case, a loss value of **0.184** suggests that the model was able to effectively minimize the difference between its predictions and the actual activity labels in the dataset.

It's important to note that the interpretation of loss values can vary depending on the specific problem and dataset. In some cases, a loss value of **0.184** might be considered low, while in others it might be considered high. To better understand the significance of the loss value in the context of your specific task, it would be helpful to compare it with the baseline performance and evaluate the performance of other models or approaches.

Overall, achieving a high overall accuracy of **98%** and a relatively low loss value

of **0.184** suggests that our MPNN model performed well in classifying molecules into their respective activity categories, considering the diverse set of activities present in the dataset.

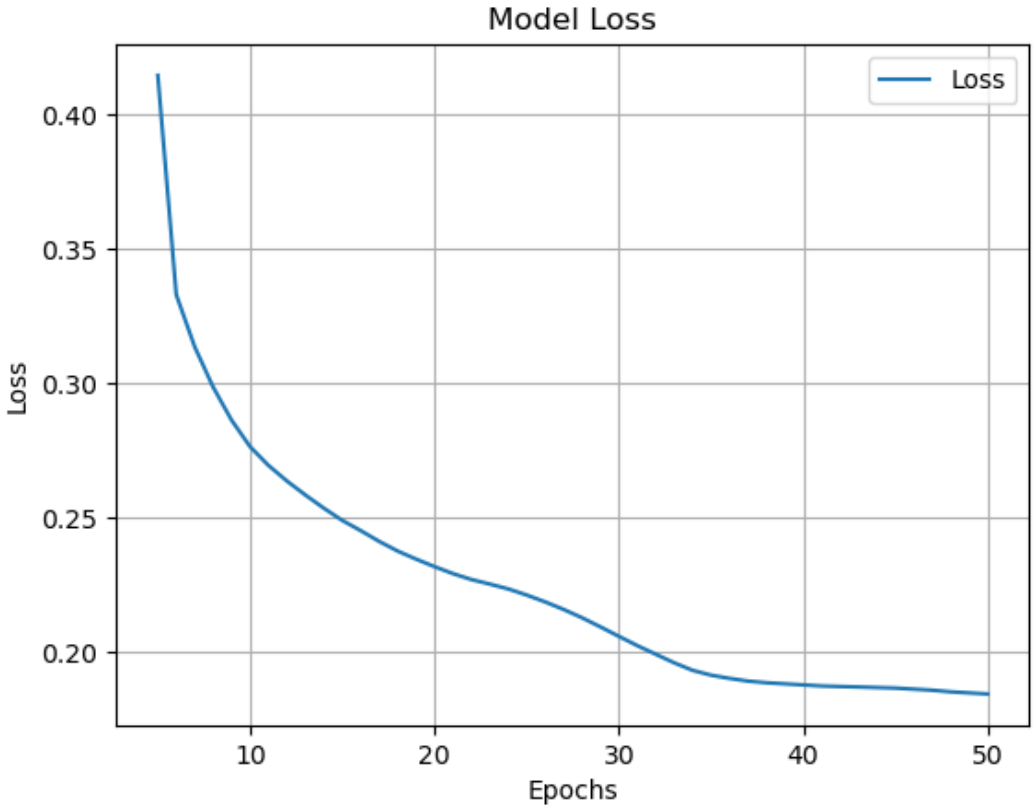


Figure 4.2: MPNN Model loss chart

4.2.2 Ligand-based virtual screening

To assess the effectiveness of our MPNN model, we utilized the entire set of activity classes from the MDDR datasets. DS1 encompasses 8294 molecules across eleven activity classes, while DS2 consists of 5083 molecules distributed over ten activity classes. The final dataset contains 8568 molecules across ten activity classes. We evaluated and validated our model on these datasets,

We utilized our fully saved network model to feed the datasets used for virtual screening. Our primary objective is to explore and assess whether this new representation can

enhance the performance of Ligand-Based Virtual Screening (LBVS).

Our retrieval results of top 1% and 5% for MDDR-DS1, MDDR-DS2 and MDDR-DS3 data sets for each activity class represented on Table 4.1, 4.2 and 4.3 respectively .

Table 4.1: Retrieval results of MDDR-DS1 for Activity classes

Activity Index	Activity Class	Active Molecules	1%	5%
31420	Renin Inhibitors	1130	0.888	0.902
71523	HIV Protease Inhibitors	750	0.666	0.880
37110	Thormbin Inhibitors	803	0.443	0.732
31432	Angiotensin II AT1 antagonists	943	0.490	0.979
42731	Substance P antagonists	1246	0.380	0.674
06233	Substance P antagonists	752	0.328	0.721
06245	5HT Reuptake Inhibitors	359	0.145	0.378
07701	D2 antagonists	395	0.176	0.449
06235	5HT1A antaonists	827	0.138	0.422
78374	Protein Kinase C inhibitors	483	0.304	0.427
78331	Cyclooxygenase Inhibitors	636	0.234	0.550

Table 4.2: Retrieval results of MDDR-DS2 for Activity classes

Activity Index	Activity Class	Active Molecules	1%	5%
07707	Adenosine A1 agonists	207	0.734	0.963
07708	Adenosine A2 agonists	156	0.885	0.991
31420	Renin Inhibitors 1	1130	0.894	0.996
42710	CCK agonists	111	0.940	0.977
64100	Monocyclic-lactams	1301	0.885	0.996
64200	Cephalosporins	158	0.649	0.884
64220	Carbacephems	1051	0.759	0.998
64500	Carbapenems	126	0.985	1.000
643500	Tribactams	388	0.880	0.997
64350	Vitamin D analogous	455	0.988	0.997

Table 4.3: Retrieval results of MDDR-DS3 for Activity classes

Activity Index	Activity Class	Active Molecules	1%	5%
09249	Muscarinie M1 agonists	900	0.435	0.727
12455	NMDA receptor antagonists	1400	0.286	0.391
12464	Nitrice oxide synthase Inhibitors	505	0.438	0.741
31281	Dopamine-hydroxylase Inhibitors	106	0.517	0.775
43210	Aldose reductase Inhibitors	957	0.445	0.710
71522	Reverse transcriptase Inhibitors	700	0.164	0.365
75721	Aromatase Inhibitors	636	0.316	0.515
78331	Cyclooxygenase Inhibitors	636	0.257	0.601
78348	Phospholipase A2 Inhibitors	617	0.089	0.266
78351	Lipoxygenase Inhibitors	2111	0.402	0.475

the experimental outcomes on the MDDRDS1, MDDR-DS2 and MDDR-DS3 data sets are provided in Tables 4.4-4.9, using cut-offs 1% and 5%. Our proposed LBVSMPNN method which is based on the Message-passing Neural Network model recorded in these tables compared to the Hybrid Siamese Similarity Models [71], DCNNLB method [70] and enhanced Siamese MLP [69] similarity model, Each row in the tables lists recall values for the top 1% and 5% of the activity class, and in each row, the best recall rate shaded in each row. The mean rows in the tables represent the average for all activity classes, whereas the shaded cells row represents the total number of shaded cells for each technique throughout the whole range of activity classes.

The first column of the table represents the activity classes of the dataset. This is followed by six columns that represent previous studies MPL, H-Dmax2, H-Dmax3, H-F-Sum, H-F-Max, and DCNNLB, and this is followed by one column that where our result LBVSMPNN.

When comparing the MDDR-DS1 recall results for the top 1% and 5% in Tables 4.4 and 4.5, the suggested LBVSMPNN technique was clearly superior to the other studies MPL, H-Dmax2, H-Dmax3, and DCNNLB in terms of the mean, and better than all other methods in term of the number of shaded cells. The suggested technique has the Third highest mean value (38.15) in Table 4.4, followed by H-Dmax2, H-Dmax3, DCNNLB, and MPL methods. In the suggested approach, the number of shaded cells is (4). The suggested approach has the fourth highest mean value (64.71) in Table 4.5, followed by H-Dmax2, H-Dmax3, and lastly MPL methods. In the suggested technique, shaded cells have a value of (3) the second-best value.

Index	MPL	H-Dmax 2	H-Dmax 3	H-F-Sum	H-F-Max	DCNNLB	LBVSMPNN
31420	82.1416	84.22	83.88	86.94	88.28	80,80	88.82
71523	49.4118	60.53	61.48	59.02	66.45	59,92	66.67
37110	45.5639	52.94	50.28	46.42	58.13	39,65	44.36
31432	45.5957	68.33	68.18	78.8	69.72	45,12	49.00
42731	32.0546	42.19	42.81	34.02	55.45	48,60	38.09
06233	22.970	33.36	35.64	32.66	47.93	35,25	32.87
06245	5.36313	10.9	10.85	19.02	16.25	21,64	14.58
07701	12.0918	14.89	14.86	29.86	22.03	19,95	17.66
06235	10.7767	16.32	14.29	26.97	23.58	21,55	13.82
78374	21.9196	24.67	24.31	25.53	25.89	29,62	30.45
78331	9.70199	8.69	8.44	16.12	14	13,63	23.40
mean	30.69	37.91	37.73	41.39	44.33	37.79	38.15
Shaded cells	0	0	0	3	3	1	4

Table 4.4: Retrieval results of top 1% for MDDR-DS1 dataset

Index	MPL	H-Dmax 2	H-Dmax 3	H-F-Sum	H-F-Max	DCNNLB	LBVSMPNN
31420	87.3628	87.55	87.58	90.43	94.06	90,12	90.21
71523	75.8289	79.51	80.65	83.44	86.44	86,31	88.07
37110	71.4536	76.55	75.56	81.71	84.54	75,95	73.24
31432	84.1489	93.81	92.26	95.23	95.02	97,17	97.92
42731	51.3644	61.2	62.53	69.74	77.27	72,32	67.41
06233	49.443	62.67	64.21	75.23	80.2	67,22	72.15
06245	16.0894	27.94	28.54	34.79	39.35	52,09	37.82
07701	29.7449	38.43	39.82	46.68	49.65	52,51	44.97
06235	28.7379	39.93	37.3	51.65	53.21	50,41	42.21
78374	36.7857	49.44	47.78	55.27	57.82	46,16	42.76
78331	24.1391	27.23	26.77	35.06	39.69	37,64	55.08
mean	50.463	58.57	58.45	65.38	68.48	66,17	64.71
Shaded cells	0	0	0	0	6	2	3

Table 4.5: Retrieval results of top 5% for MDDR-DS1 dataset

Furthermore, the MDDR-DS2 recall values obtained at the top 1%, as shown in Table 4.6, demonstrate that the suggested LBVSMPNN technique outperforms other techniques in terms of the shaded cells, the LBVSMPNN approach produced the best retrieval recall results, and the suggested method's mean value is extremely close to that of prior research. However, by comparison, The MDDR-DS2 recall values obtained at the top 5% in Table 4.7 clearly show that the suggested LBVSMPNN approach outperforms all the methods (MPL, H-Dmax2, H-Dmax3, H-F-Sum, H-F-Max and DCNNLB). In terms of the mean and number of shaded cells, the suggested approach has the best mean value (98.023) and the highest number of shaded cells (8), which that result better than other studies in all terms.

Index	MPL	H-Dmax 2	H-Dmax 3	H-F-Sum	H-F-Max	DCNNLB	LBVSMPNN
7707	86.4706	88.2	91.8	83.46	83.61	73,86	73.49
7708	97.3077	93.16	94.9	92.32	93.1	83,85	88.51
31420	71.7699	80.02	80.64	78.93	77.35	65,97	89.40
42710	82.9091	85.82	84.73	84.45	84.91	92,07	94.00
64100	94.2769	96.47	95.66	91.73	93.22	87,32	88.53
64200	35.5696	65.87	77.35	55.94	60.39	76,65	64.96
64220	88.5333	82.62	90.53	92.29	92.34	82,39	75.95
64500	62.8571	68.56	72.4	72.56	73.2	87,94	98.56
64350	91.8557	85.27	88.34	84.99	85.06	75,41	88.06
75755	90.5727	90.53	90.99	90.9	90.99	98,46	98.89
mean	80.21226	83.65	86.73	82.76	83.42	82,39	86.035
Shaded cells	2	1	2	0	1	0	4

Table 4.6: Retrieval results of top 1% for MDDR-DS2 dataset

Index	MPL	H-Dmax 2	H-Dmax 3	H-F-Sum	H-F-Max	DCNNLB	LBVSMPNN
7707	94.2157	89.9	94.83	88.29	87.17	90,53	96.35
7708	98.7179	93.74	95.61	93.48	93.94	95,51	99.16
31420	92.9381	95.82	95.72	95.49	94.89	94,03	99.66
42710	88	87.82	85.45	86.91	85.64	98,29	97.72
64100	96.6615	98.58	98.39	95.92	98,62	87,32	99.67
64200	38.6076	74	85.61	63.48	70.58	91,90	88.40
64220	93.2381	89.93	94.04	93.72	94.28	99,64	99.82
64500	71.2698	71.04	73.84	73.76	74.16	98,89	100
64350	95.3608	92.26	90.7	90.6	90.65	99,28	99.79
75755	93.8767	90.84	90.99	90.95	90.99	99,67	99.71
mean	86.28862	88.39	90.52	87.29	87.87	96,64	98.028
Shaded cells	0	0	0	0	0	2	8

Table 4.7: Retrieval results of top 5% for MDDR-DS2 dataset

In addition, the MDDR-DS3 recall values recorded at the top 1% and 5% in Tables 4.8 and 4.9, respectively, in Table 4.8, the proposed method, in terms of the number of shaded cells, our method outperforms all other, and close than the best mean value of DCNNLB study and followed by H-F-Max, H-F-Sum, MPL, H-Dmax2 and finally, H-Dmax3 method. By comparison, in Table 4.9, the suggested LBVSMPNN method was obviously superior to the H-F-Sum and H-Dmax3, followed by H-Dmax2, and MPL in term of mean value.

Index	MPL	H-Dmax 2	H-Dmax 3	H-F-Sum	H-F-Max	DCNNLB	LBVSMPNN
09249	39.7556	38.53334	41.16666	45.12	42.19	74.71	43.58
12455	9.8	16.24286	18.37856	23.09	23.26	28.08	28.67
12464	31.84	26.79208	31.94058	37.31	41.33	20.95	43.82
31281	61.8	65.52382	64.7619	66.76	68.38	36.89	51.71
43210	17.5789	28.78536	28.37696	36.05	37.35	41.53	44.50
71522	6.42857	9.08571	10.985706	14.7	15.43	52.53	16.46
75721	57.5667	50.48818	51.73228	54.43	56.06	26.38	31.62
78331	41.3	12.56694	12.125994	14.96	16.65	27.56	25.77
78348	12.2	15.999974	13.512192	14.33	15.79	17.42	8.94
78351	14.3024	11.715648	14.8673	16.68	16.42	11.52	40.20
mean	29.257217	27.57	28.78	32.34	33.29	33,76	33.527
Shaded cells	2	0	0	0	1	3	4

Table 4.8: Retrieval results of top 1% for MDDR-DS3 dataset

Index	MPL	H-Dmax 2	H-Dmax 3	H-F-Sum	H-F-Max	DCNNLB	LBVSMPNN
09249	61.1556	62.47778	71.52222	75.03	67.78	98.30	72.71
12455	27.1429	34.05	41.58572	46.95	47.07	58.84	39.18
12464	53.72	47.70298	56.37622	60.95	80.2	47.74	74.16
31281	75.8	74.09524	73.14286	72.67	88.57	85.47	77.52
43210	36.2105	52.09424	56.50262	62.08	51.15	73.75	71.06
71522	15.9143	23.62858	30.95714	37.27	31.36	86.54	36.56
75721	78.2333	76.6614	81.44882	83.46	98.66	50.00	51.57
78331	78.2	31.52754	35.32284	36.93	42.36	52.08	60.15
78348	25.4667	35.51222	33.43086	38.41	47.8	50.18	26.62
78351	34.2667	30.95734	40.02844	41.2	37.89	16.49	47.59
mean	48.611	46.87	52.03	55.5	59.28	61,94	55.712
Shaded cells	1	0	0	0	2	5	2

Table 4.9: Retrieval results of top 5% for MDDR-DS3 dataset

Figures 4.3–4.8 show the contrast between methods for the average recall percentage of successful compound retrieval at the top of the 1% and 5% in MDDR-DS1, MDDR-DS2, MDDR-DS3 respectively.

Figure 4.3 compares methods for the average recall percentage of successful compound retrieval at the top 1% in MDDR-DS1. By comparison, our LBVSMPNN gives the best retrieval recall results in 31420, 71523, 78374, and 78331 Activity classes with a percentage of 88.82%, 66.67%, 30.45%, 23.40% respectively.

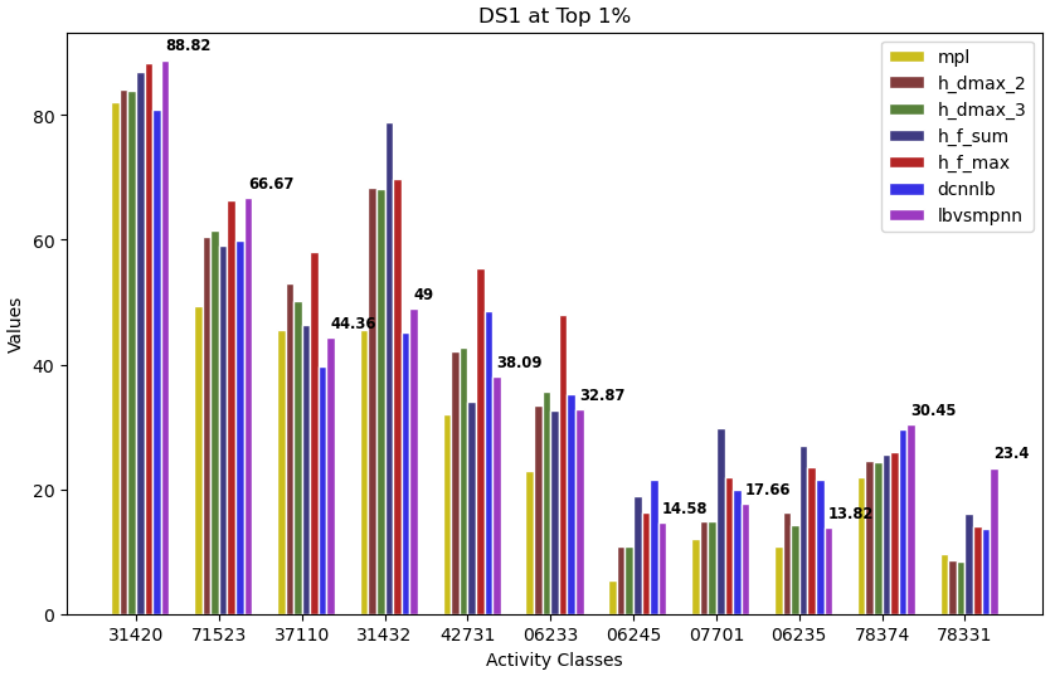


Figure 4.3: The comparison among methods for the average recall percentage at the top 1% in MDDR-DS1

Figure 4.4 presents a comparison of different approaches in terms of the average recall percentage achieved for successful compound retrieval in the top 5% of results in MDDR-DS1.4.4 shows outperform from our proposed LBVSMPNN in 71523, 31432 and 78331 Activity classes with a percentage of 88.7%,97.92%,55.08% respectively.

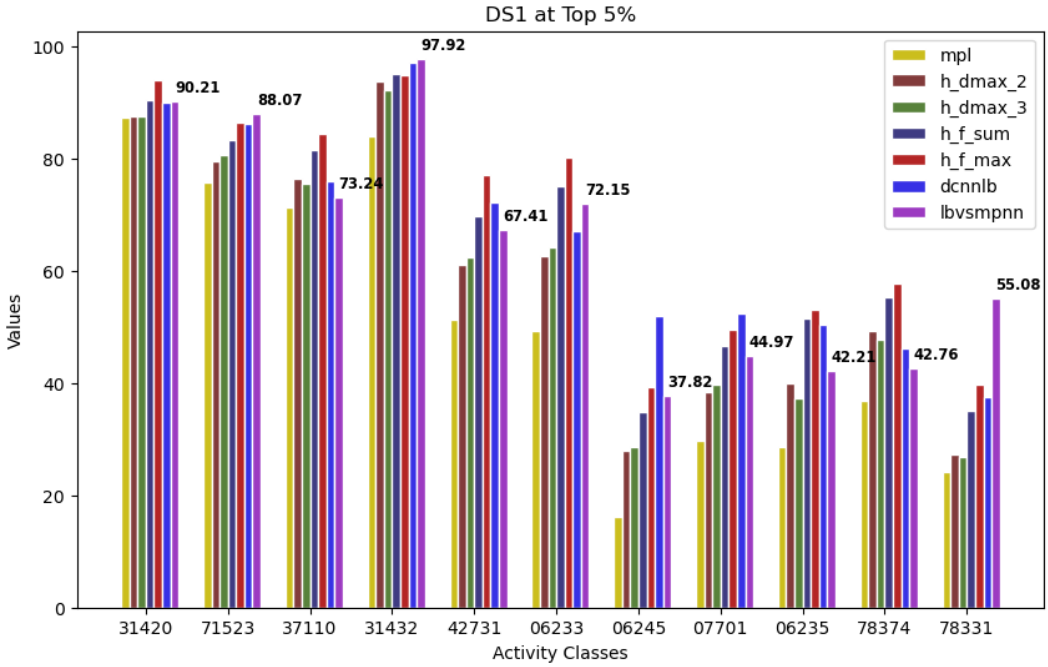


Figure 4.4: The comparison among methods for the average recall percentage at the top 5% in MDDR-DS1

In Figure 4.5, the average recall percentage achieved for successful compound retrieval in the top 1% of results in MDDR-DS2 is compared among different approaches. Our proposed LBVSMPNN method demonstrates superior performance in 31420, 42710,64500, and 75755 activity classes, achieving percentages of 89.40%, 94.00%, 98.56%, and 98.89%, respectively.

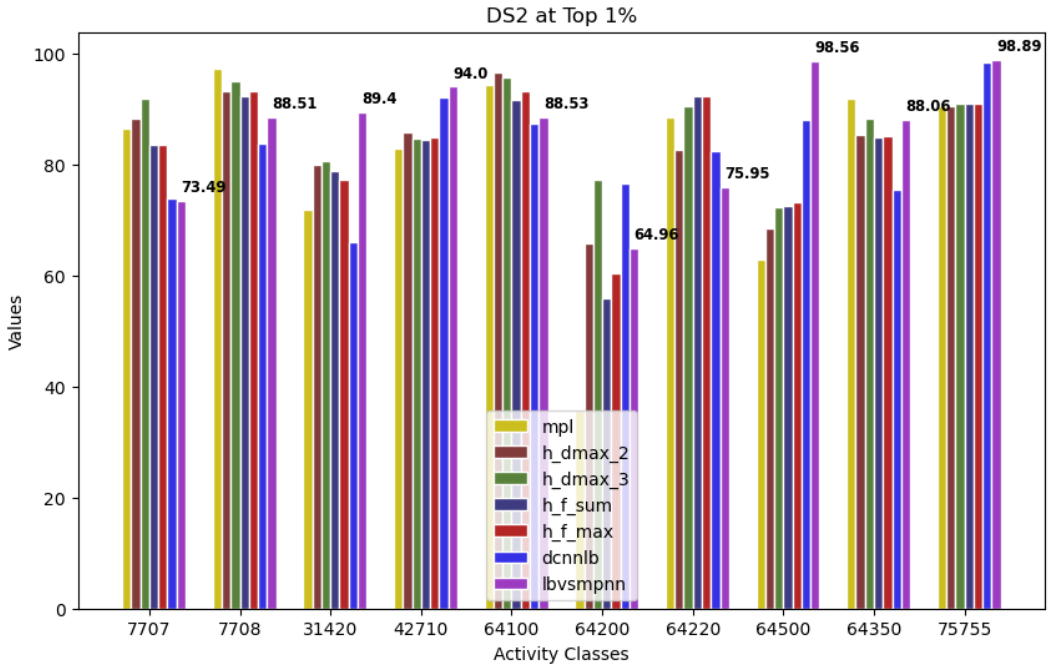


Figure 4.5: The comparison among methods for the average recall percentage at the top 1% in MDDR-DS2

In Figure 4.6, a comparison is depicted among various methods regarding the average recall percentage attained for successful compound retrieval within the top 5% of results in MDDR-DS2. The results demonstrate that our proposed LBVSMPNN outperforms other approaches in activity classes 7707, 7708,31420,64100,64220,64500,64350 and 75755 achieving recall percentages of 96.35%, 99.16%, 99.66%,99.67%,99.82%,100%,99.79%, and 99.71% respectively.

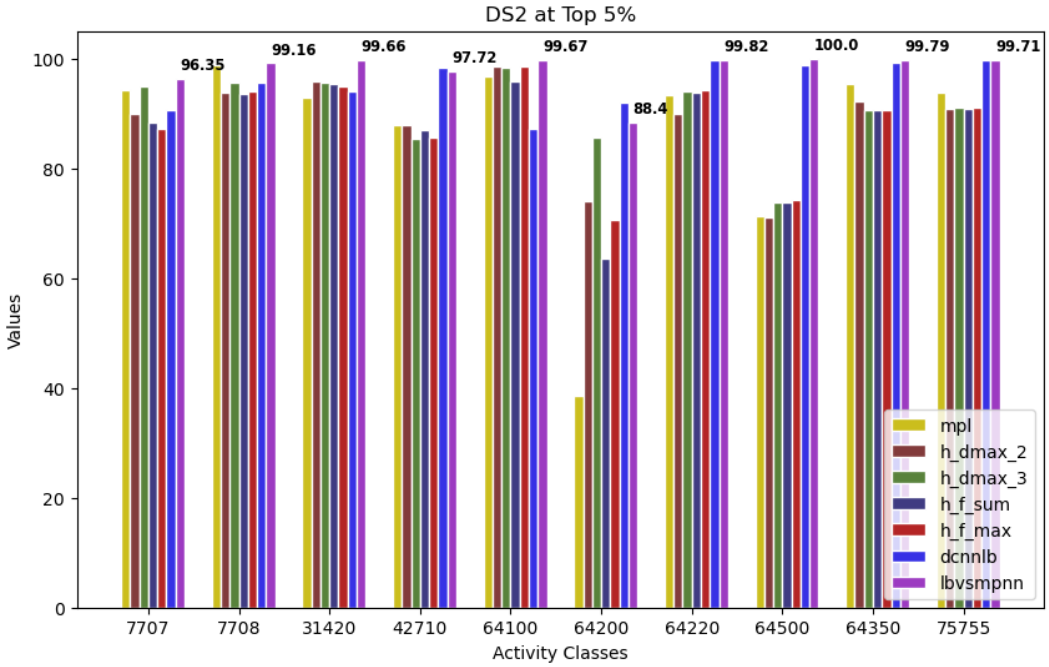


Figure 4.6: The comparison among methods for the average recall percentage at the top 5% in MDDR-DS2

Figure 4.7 illustrates a comparison of various methods based on the average recall percentage achieved for successful compound retrieval within the top 1% of results in MDDR-DS3. The results highlight that our proposed LBVSMPNN surpasses other approaches in activity classes 12455, 12464,43210, and 78351, with recall percentages of 28.67%, 43.82%, 44.50%, and 40.20% respectively.

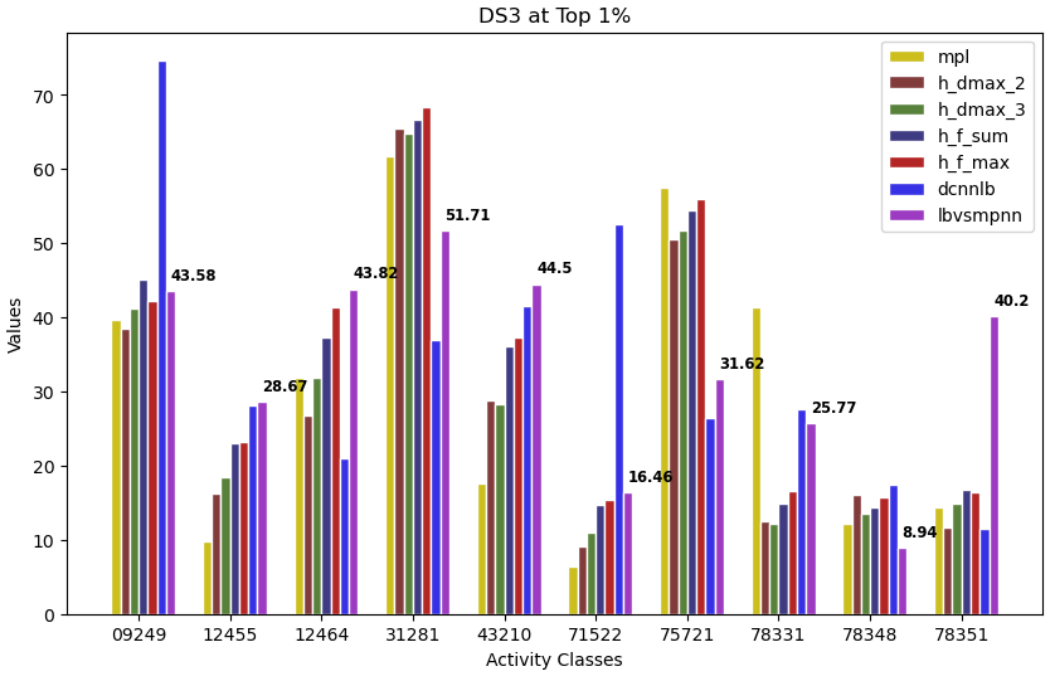


Figure 4.7: The comparison among methods for the average recall percentage at the top 1% in MDDR-DS3

The record shown on Figure 4.8 present a comparison of various methods based on the average recall percentage achieved for successful compound retrieval within the top 5% of results in MDDR-DS3. the results show that our proposed LBVSMPNN outperforms other approaches in activity classes 12464 and 78351 with recall percentages of 74.16% and 47.59%respectively.

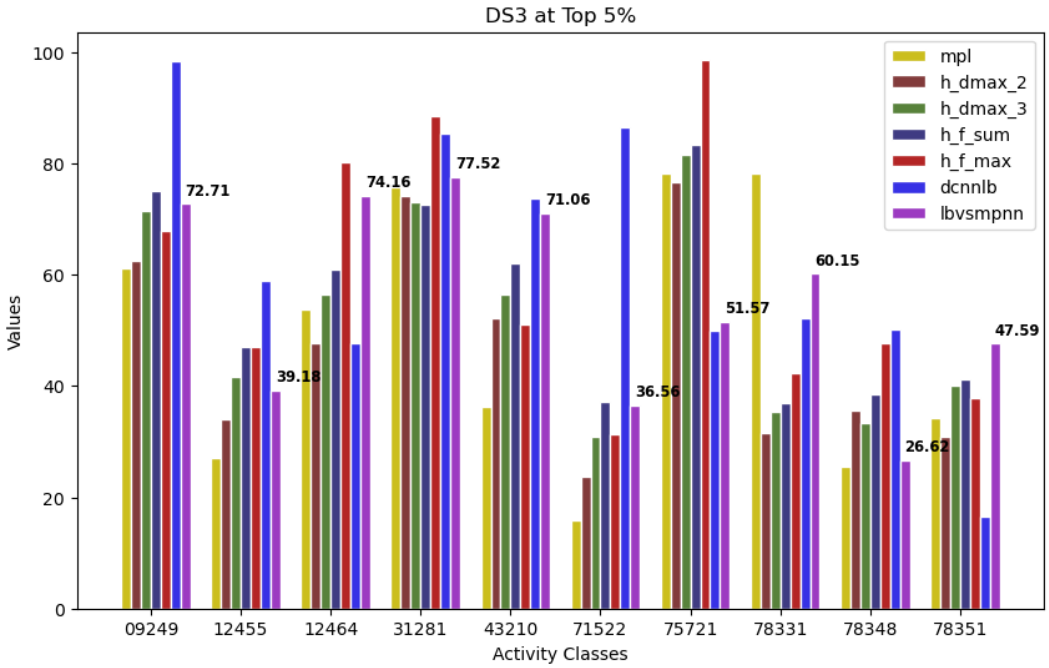


Figure 4.8: The comparison among methods for the average recall percentage at the top 5% in MDDR-DS3

4.3 Conclusion

In conclusion, the results from this study demonstrate the effectiveness of MPNNs with LBVS in accurately predicting biological activities. The strong performance of our model on the DS1, DS2, and DS3 datasets, when compared to existing literature, underscores its potential as a reliable and valuable tool in drug discovery research. These results contribute to the growing body of evidence supporting the utility of MPNNs in the field and lay the foundation for further advancements in computational drug design.

Conclusion

This thesis has successfully demonstrated the effectiveness of the MPNN model in computer-aided drug design, specifically in ligand-based virtual screening. The model has shown improved performance compared to previous studies, offering a promising approach for accelerating the discovery and development of new drugs.

However, several challenges and limitations should be acknowledged. The first challenge is the availability and quality of data, as obtaining comprehensive and reliable datasets for training and validation can be demanding. Additionally, while the MPNN model exhibits strong performance, it may still encounter difficulties in generalizing to novel chemical compounds and target proteins. Another limitation is the computational resources required for training and deploying MPNN models, which can be substantial and hinder their widespread use. Furthermore, the interpretability of MPNN models remains a challenge, as understanding the molecular interactions and decision-making processes behind the model's predictions is complex.

To address these challenges and limitations, our future research can focus on several directions. Firstly, efforts should be made to enhance data collection and curation, including the inclusion of diverse chemical compounds and target proteins, as well as experimental validation of the results. Secondly, exploring transfer learning techniques to improve the generalization of MPNN models across different drug discovery scenarios and expanding their applicability is crucial. Additionally, developing more computationally efficient training strategies, such as model compression and parallelization, can help mitigate the resource requirements of MPNN models. Moreover, research should be conducted

to improve the interpretability and explainability of MPNN models, enabling researchers to gain deeper insights into the model's predictions. Integrating multiple modalities, such as structure-based methods and omics data, is another avenue for future work, allowing for a more comprehensive and accurate drug discovery pipeline.

Bibliography

[1] Stephani Joy Y Macalino, Vijayakumar Gosu, Sunhye Hong, and Sun Choi. Role of computer-aided drug design in modern drug discovery. *Archives of pharmacal research*, 38:1686–1701, 2015.

[2] Daniel S Wigh, Jonathan M Goodman, and Alexei A Lapkin. A review of molecular representation in the age of machine learning. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5):e1603, 2022.

[3] Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Frei, Nathan C Frey, Pascal Friederich, Théophile Gaudin, Alberto Alexander Gayle, Kevin Maik Jablonka, et al. Selfies and the future of molecular string representations. *Patterns*, 3(10):100588, 2022.

[4] Maya Hirohara, Yutaka Saito, Yuki Koda, Kengo Sato, and Yasubumi Sakakibara. Convolutional neural network based on smiles representation of compounds for detecting chemical motif. *BMC bioinformatics*, 19:83–94, 2018.

[5] Laurianne David, Amol Thakkar, Rocío Mercado, and Ola Engkvist. Molecular representations in ai-driven drug discovery: a review and practical guide. *Journal of Cheminformatics*, 12(1):1–22, 2020.

[6] Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, 2015.

[7] Mauricio Lorca, César Morales-Verdejo, Diego Vásquez-Velásquez, Jaqueline Andrades-Lagos, Johanna Campanini-Salinas, Jorge Soto-Delgado, Gonzalo Recabarren-Gajardo, and Julio Mella. Structure-activity relationships based on 3d-qsar comfa/comsia and design of aryloxypropanol-amine agonists with selectivity for the human 3-adrenergic receptor and anti-obesity and anti-diabetic profiles. *Molecules*, 23(5):1191, 2018.

[8] Hossein Hassani, Emmanuel Sirimal Silva, Stephane Unger, Maedeh TajMazinani, and Stephen Mac Feely. Artificial intelligence (ai) or intelligence augmentation (ia): what is the future? *Ai*, 1(2):8, 2020.

[9] Rocío Mercado, Tobias Rastemo, Edvard Lindelöf, Günter Klambauer, Ola Engkvist, Hongming Chen, and Esben Jannik Bjerrum. Graph networks for molecular design. *ChemRxiv*, 2020.

[10] Jimmy Shen. Graph neural networks. `https://jimmy-shen.medium.com/graph-neural-networks-5df67d78f262`, Year of publication. Accessed: July 7, 2023.

[11] Unknown. Multi-head attention mechanism. `https://data-science-blog.com/en/blog/2021/04/07/multi-head-attention-mechanism/`, 2021. Accessed: July 7, 2023.

[12] Lingfei Wu, Peng Cui, Jian Pei, Liang Zhao, and Xiaojie Guo. Graph neural networks: foundation, frontiers and applications. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4840–4841, 2022.

[13] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

[14] Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Marvin Steijaert, Jörg K

Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chemical science*, 9(24):5441–5451, 2018.

[15] Alpha A Lee, Qingyi Yang, Asser Bassyouni, Christopher R Butler, Xinjun Hou, Stephen Jenkinson, and David A Price. Ligand biological activity predicted by cleaning positive and negative chemical correlations. *Proceedings of the National Academy of Sciences*, 116(9):3373–3378, 2019.

[16] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[17] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28, 2015.

[18] John Boaz Lee, Ryan Rossi, and Xiangnan Kong. Graph classification using structural attention. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1666–1674, 2018.

[19] James Atwood and Don Towsley. Diffusion-convolutional neural networks. *Advances in neural information processing systems*, 29, 2016.

[20] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 30:595–608, 2016.

[21] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. In *International conference on machine learning*, pages 2014–2023. PMLR, 2016.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[23] Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. Low data drug discovery with one-shot learning. *ACS central science*, 3(4):283–293, 2017.

[24] Jingjia Huang, Zhangheng Li, Nannan Li, Shan Liu, and Ge Li. Attpool: Towards hierarchical feature representation in graph convolutional networks via attention mechanism. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6480–6489, 2019.

[25] Bonggun Shin, Sungsoo Park, Keunsoo Kang, and Joyce C Ho. Self-attention based molecule representation for predicting drug-target interaction. In *Machine Learning for Healthcare Conference*, pages 230–248. PMLR, 2019.

[26] Cristian Bologa, Tharun Kumar Allu, Marius Olah, Michael A Kappler, and Tudor I Oprea. Descriptor collision and confusion: Toward the design of descriptors to mask chemical structures. *Journal of computer-aided molecular design*, 19:625–635, 2005.

[27] Igor V Tetko, Ruben Abagyan, and Tudor I Oprea. Surrogate data–a secure way to share corporate data. *Journal of computer-aided molecular design*, 19:749–764, 2005.

[28] W Patrick Walters and Regina Barzilay. Applications of deep learning in molecule generation and molecular property prediction. *Accounts of chemical research*, 54(2):263–270, 2020.

[29] Alexandre Varnek and Igor Baskin. Machine learning methods for property prediction in chemoinformatics: Quo vadis? *Journal of chemical information and modeling*, 52(6):1413–1437, 2012.

[30] Yu-Chen Lo, Stefano E Rensi, Wen Torng, and Russ B Altman. Machine learning in chemoinformatics and drug discovery. *Drug discovery today*, 23(8):1538–1546, 2018.

[31] Syed M Ali, Michael Z Hoemann, Jeffrey Aubé, Gunda I Georg, Lester A Mitscher, and Lalith R Jayasinghe. Butitaxel analogues: Synthesis and structure- activity relationships. *Journal of medicinal chemistry*, 40(2):236–241, 1997.

[32] B Firdaus Begam and J Satheesh Kumar. A study on cheminformatics and its applications on modern drug discovery. *Procedia engineering*, 38:1264–1275, 2012.

[33] Johann Gasteiger. Chemoinformatics: Achievements and challenges, a personal view. *Molecules*, 21(2):151, 2016.

[34] Johan Åqvist, Carmen Medina, and Jan-Erik Samuelsson. A new method for predicting binding affinity in computer-aided drug design. *Protein Engineering, Design and Selection*, 7(3):385–391, 1994.

[35] Carlton A Taft, Vinicius Barreto Da Silva, et al. Current topics in computer-aided drug design. *Journal of pharmaceutical sciences*, 97(3):1089–1098, 2008.

[36] Wei Tian, Chang Chen, Xue Lei, Jieling Zhao, and Jie Liang. Castp 3.0: computed atlas of surface topography of proteins. *Nucleic acids research*, 46(W1):W363–W367, 2018.

[37] Hsi-Ming Lee, Sebastian G Ciancio, Gülay Tüter, Maria E Ryan, Eugene Komaroff, and Lorne M Golub. Subantimicrobial dose doxycycline efficacy as a matrix metalloproteinase inhibitor in chronic periodontitis patients is enhanced when combined with a non-steroidal anti-inflammatory drug. *Journal of periodontology*, 75(3):453–463, 2004.

[38] Hung-Jin Huang, Kuei-Jen Lee, Hsin Wei Yu, Chien-Yu Chen, Chih-Ho Hsu, Hsin-Yi Chen, Fuu-Jen Tsai, and Calvin Yu-Chian Chen. Structure-based and ligand-based drug design for her 2 receptor. *Journal of Biomolecular Structure and Dynamics*, 28(1):23–37, 2010.

[39] Jitender Verma, Vijay M Khedkar, and Evans C Coutinho. 3d-qsar in drug design-a review. *Current topics in medicinal chemistry*, 10(1):95–115, 2010.

[40] Xiaoyu Qing, Xiao Yin Lee, Joren De Raeymaecker, Jeremy RH Tame, Kam YJ Zhang, Marc De Maeyer, and Arnout Voet. Pharmacophore modeling: advances, limitations, and current utility in drug discovery. *Journal of Receptor, Ligand and Channel Research*, 7:81–92, 2014.

[41] P Gund. Evolution of the pharmacophore concept in pharmaceutical research. *Pharmacophore perception, development, and use in drug design*, pages 5–11, 2000.

[42] Talia B Kimber, Yonghui Chen, and Andrea Volkamer. Deep learning in virtual screening: recent applications and developments. *International Journal of Molecular Sciences*, 22(9):4435, 2021.

[43] Yusuf O Adeshina, Eric J Deeds, and John Karanicolas. Machine learning classification can reduce false positives in structure-based virtual screening. *Proceedings of the National Academy of Sciences*, 117(31):18477–18488, 2020.

[44] Rebekah A Adams, Gabrielle Leon, Natalia M Miller, Saira P Reyes, Chantal H Thantrong, Alina M Thokkadam, Annabel S Lemma, Darshan M Sivaloganathan, Xuanqing Wan, and Mark P Brynildsen. Rifamycin antibiotics and the mechanisms of their failure. *The Journal of antibiotics*, 74(11):786–798, 2021.

[45] Clare F Macrae, Ioana Sovago, Simon J Cottrell, Peter TA Galek, Patrick McCabe, Elna Pidcock, Michael Platings, Greg P Shields, Joanna S Stevens, Matthew Towler, et al. Mercury 4.0: From visualization to analysis, design and prediction. *Journal of applied crystallography*, 53(1):226–235, 2020.

[46] Marcus D Hanwell, Donald E Curtis, David C Lonie, Tim Vandermeersch, Eva Zurek, and Geoffrey R Hutchison. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *Journal of cheminformatics*, 4(1):1–17, 2012.

[47] Koichi Momma and Fujio Izumi. Vesta 3 for three-dimensional visualization of crystal, volumetric and morphology data. *Journal of applied crystallography*, 44(6):1272–1276, 2011.

[48] Warren L DeLano et al. Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr*, 40(1):82–92, 2002.

[49] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

[50] Gerald M Maggiora and Veerabahu Shanmugasundaram. Molecular similarity measures. *Chemoinformatics and computational chemical biology*, pages 39–100, 2011.

[51] Craig Jackson, M. Esnouf, Donald Winzor, and David Duewer. Defining and measuring biological activity: Applying the principles of metrology. *Accreditation and Quality Assurance*, 12:283–294, 06 2007.

[52] Ola Spjuth. Novel applications of machine learning in cheminformatics, 2018.

[53] Jiulin Song and Yansheng Chen. A study on the application and the advancement of deep neural network algorithm. In *Journal of Physics: Conference Series*, volume 2146, page 012001. IOP Publishing, 2022.

[54] Niclas Ståhl, Göran Falkman, Alexander Karlsson, Gunnar Mathiason, and Jonas Boström. Deep convolutional neural networks for the prediction of molecular properties: challenges and opportunities connected to the data. *Journal of Integrative Bioinformatics*, 16(1):20180065, 2018.

[55] Vijay Prakash Dwivedi, Chaitanya K Joshi, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking graph neural networks. 2020.

[56] Andy D Perkins. Addressing challenges in a graph-based analysis of high-throughput biological data. 2008.

[57] Mandana Saebi, Bozhao Nan, John Herr, Jessica Wahlers, Olaf Wiest, and Nitesh Chawla. Graph neural networks for predicting chemical reaction performance. 2021.

[58] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR, 06–11 Aug 2017.

[59] Jeonghee Jo, Bumju Kwak, Hyun-Soo Choi, and Sungroh Yoon. The message passing neural networks for chemical property prediction on smiles. *Methods*, 179:65–72, 2020. Interpretable machine learning in bioinformatics.

[60] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019. PMID: 31361484.

[61] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(1):1–23, 2019.

[62] Yuchen Zhou, Hongtao Huo, Zhiwen Hou, and Fanliang Bu. A deep graph convolutional neural network architecture for graph classification. *Plos one*, 18(3):e0279604, 2023.

[63] E. Brissman, J. Johnander, M. Danelljan, G. Bhat, F. S. Khan, M. Felsberg, and A. C. Berg. Recurrent graph neural networks for video instance segmentation. *International Journal of Computer Vision*, 131(2):471–495, 2023.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon,

U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[65] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J. Kim. Graph transformer networks. In *Neural Information Processing Systems*, 2019.

[66] Md. Arid Hasan, Firoj Alam, Shammur Absar Chowdhury, and Naira Khan. Neural machine translation for the bangla-english language pair. In *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, pages 1–6, 2019.

[67] Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. *ArXiv*, abs/2012.09699, 2020.

[68] Grégoire Mialon, Dexiong Chen, Margot Selosse, and Julien Mairal. Graphit: Encoding graph structure in transformers. *ArXiv*, abs/2106.05667, 2021.

[69] Mohammed Khaled Altalib and Naomie Salim. Similarity-based virtual screen using enhanced siamese multi-layer perceptron. *Molecules (Basel, Switzerland)*, 26(21):6669, 2021.

[70] Fouaz Berrhail, Hacene Belhadef, and Mohammed Haddad. Deep convolutional neural network to improve the performances of screening process in lbvs. *Expert Systems with Applications*, 203:117287, 2022.

[71] Mohammed Khaled Altalib and Naomie Salim. Hybrid-enhanced siamese similarity models in ligand-based virtual screen. *Biomolecules*, 12(11):1719, 2022.

[72] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Message passing neural networks. *Machine learning meets quantum physics*, pages 199–214, 2020.

[73] Vassilis N Ioannidis, Antonio G Marques, and Georgios B Giannakis. A recurrent graph neural network for multi-relational data. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8157–8161. IEEE, 2019.

[74] BIOVIA - Dassault Systèmes®. `https://www.3ds.com/products-services/biovia/products/data-science/pipeline-pilot/`. Accessed: June 3, 2023.

[75] Robert P Sheridan and Joseph Shpungin. Calculating similarities between biological activities in the mdl drug data report database. *Journal of chemical information and computer sciences*, 44(2):727–740, 2004.

[76] Maged Nasser, Naomie Salim, Faisal Saeed, Shadi Basurra, Idris Rabiu, Hentabli Hamza, and Muaadh A Alsoufi. Feature reduction for molecular similarity searching based on autoencoder deep learning. *Biomolecules*, 12(4):508, 2022.

[77] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.