CHAPTER 1

INTRODUCTION

This chapter covers the following subtopics: (1) Rationale, (2) Statement of the Problem, (3)Objective & Hypothesis, (4) Assumption, (5) Scope and Delimitations, and (6) Significance of the Study.

1.1 Rationale

The interaction between a drug and a target that alters the function or workflow of the target is known as drug-target interaction (DTI) [1]. A drug is any chemical substance that alters an organism's chemical composition when ingested. The term 'target' refers to any biological component (often a protein or nucleic acid) that interacts with the drug, leading to alterations in chemical conditions. DTI plays a crucial role in the drug development process, which may take up to 2.6 billion US dollars and at least 17 years to complete from the original hypothesis to official marketing [2]. The process is long, complex, costly, and has a low chance of success. In addition, because of their unknown interactions, the majority of known chemical substances have yet to be utilized as drugs [1]. Therefore, in recent years, there has been a great deal of interest in the process of understanding how drugs interact with their targets and how to predict drug-target interactions [2].

Direct laboratory experiments using techniques like high-throughput screening (HTS) can be used to perform the DTI prediction process [2]. However, the experimental method in the lab is time-consuming and expensive. Therefore, a new in-silico approach is needed to address these problems [1]. One approach that can be used is the computational method. Furthermore, the availability of large volumes of data on drug compounds with hundreds of potential targets makes computational methods crucial in DTI prediction [2]. Computational methods for DTI are categorized into three approaches: ligand-based, docking simulation-based, and chemogenomic-based. [3]. Ligand-based approaches and docking simulations are conventional approaches. Ligand-based approaches are founded on the principle that structurally identical molecules exhibit identical properties, so drug molecules should be able to bind to proteins that have identical molecules. However, this approach has drawbacks because the interaction predictions are limited to known drug molecules and proteins [4]. The second approach is based on docking by utilizing the 3D structure of the protein, but this approach has disadvantages such as the 3D structure of the protein which is still unknown, and the complexity and requires large computational power [4]. The chemogenomic approach is a solution to the shortcomings of the previous two approaches. This approach uses drug chemical space information and protein genomic space and unifies them in the same subspace to infer possible interactions [1]. Several methods

are often used in chemogenomic approaches, such as statistical methods [5, 6], machine learning algorithms [7, 8], and deep learning models [9, 10].

In recent years, the use of deep learning has become a frequently used method in DTI prediction. This is because deep learning architecture can identify hidden or complex patterns or data representations. As such, creating effective deep-learning models is essential for discovering hit compounds, which serve as potential drugs for therapeutic applications [11]. Generally, DTI prediction is commonly classified as a classification task [11–13]. This means only predicting whether a drug interacts with a specific target. However, the prediction results lack a crucial piece of information—the binding affinity value. This value represents the strength of interaction between a drug and its target pair and is measured as a continuous numerical value [14]. Predicting drug-target affinity (DTA) provides the advantage of estimating the interaction strength between a drug and its target, thereby narrowing the vast search space for potential compounds in drug discovery research [14].

Previous studies have been conducted, showing different results using various drug and target representations in drug-target affinity prediction using deep learning models. Several studies in DTA prediction often utilize drug and protein sequence representations. Convolutional Neural Networks (CNN) are a popular choice for feature extraction from these representations, with models like [14–19] achieving notable performance on benchmark datasets such as Davis and KIBA. Even though deep learning models have shown promising results for DTA prediction, many studies often only use simple concatenation between drug and protein in the interaction modeling. This can lead to some lost interaction information between drugs and proteins. To address this issue, various studies have integrated attention mechanisms for better representation and interaction modeling between drugs and proteins [20–22]. The addition of attention mechanisms has shown good performance in predicting drug-target binding affinity values for sequence-based models. However, there is still essentially a lack of representation for drug and target sequences as strings. This is due to the possibility of losing structural details about the drug and target when employing string representation, which may have an impact on binding affinity prediction [23]. Several studies have used graphs as drug representations. These studies employ graph neural network models (GNN) such as graph convolutional Network (GCN), graph attention network (GAT), and graph isomorphism network (GIN) with various modifications and additional drug and protein representation [23–26].

Although several studies have performed DTA prediction with drug representation as a graph, there are still some issues that need to be addressed. Firstly, many studies have used the GAT model for drug representation learning because of its attention mechanism. However, GAT models have a static attention mechanism. The weight, or "rank," given to other nodes in the graph when computing the representation of a query node does not change depending on the query node itself. In other words, GAT looks at every relationship between a query node and other nodes in the graph in the same way, without

considering the specific context of the query node. To overcome this limitation, GATv2 was introduced as a dynamic attention variant, providing more expressive representations by adapting attention weights to the query node's context [27]. Secondly, even though current graph-based methods already implemented GATv2 models for DTA prediction. These studies commonly represent drugs as a single graph structure. They overlook multi-scale structural information in drugs, including the features of individual amino acids, motifs, and various levels of structural elements such as atoms and molecular fragments. These interactions and correlations across different structural scales are essential in drug-target protein interactions. [28]. Additionally, substructures like motifs hold specific significance in drug molecules, such as NO_2 and carbon ring groups, which are susceptible to mutagenesis. [24]. Thus, motifs deserve more attention as additional drug representations. By integrating both the overall drug molecular graph and drug motifs graph, DTA models can achieve a more comprehensive representation of the drug, leading to improved prediction accuracy.

Thirdly, protein sequence representations are long, with each character describing an amino acid. Conventional models are unable to process contextual relationships within sequences, missing critical relationships between preceding and following amino acids. Bidirectional LSTM (BiLSTM) networks provide an alternative approach by considering both future and past contexts, enabling a more thorough understanding of the protein's structural and functional properties [29]. Lastly, recent interaction modeling often relies on straightforward concatenation of drug and protein representations. This simplified approach overlooks the complex relationships between drug graphs and protein sequences, potentially ignoring the essential interaction information that might influence binding affinity. Moreover, many current methods only focus on capturing interactions between two representations, drug graphs, and protein sequences without accounting for other potential representations (e.g., drug motifs). This limitation hinders the model's capability to fully capture the complex interactions between various drug and target features, possibly affecting binding affinity value. To address these issues, alternative approaches can leverage advanced attention mechanisms to incorporate additional representations and employ interaction models capable of processing multiple inputs in a more context-aware manner.

To address the limitations mentioned above, this study proposes an enhanced GATv2 model to obtain more informative features from the drug graph node and combine it with drug multi-scale features, enabling dynamic adaptation and selective fusion of features across different representation scales. For protein sequences, BiLSTM is utilized to capture long-term dependencies and contextual associations, leveraging its ability for sequence data. We also incorporate a new attention mechanism inspired by the AttentionDTA study [22], called a three-way multi-head attention mechanism. We modified the attention mechanism to handle additional input, which is the drug motif graph. this mechanism enables each representation to focus on critical regions of the other, effectively highlighting

important cross-interactions. This comprehensive framework aims to overcome existing limitations and improve the accuracy and interpretability of DTA predictions.

1.2 Statement of the Problem

In the DTA prediction field, common baseline models include GCN and GAT. GCN aggregates information from adjacent nodes to capture the structural properties of molecules, while GAT introduces an attention mechanism that assigns varying weights to adjacent nodes, focusing on the most relevant node features. For protein sequence embedding, baseline models include 1DCNN, which extract local sequential features, and LSTM, which capture long-term dependencies. These models are compared against the proposed BiLSTM, which incorporates bidirectional processing to capture both future and past contexts in the protein sequences, providing a more in-depth understanding of protein functionality. Meanwhile, several benchmark state-of-the-art methods include GraphDTA, MSGNN-DTA, and DGDTA. GraphDTA was a pioneer in applying graph-based representations to DTA prediction, laying the foundation for subsequent graph neural network applications. MSGNN-DTA utilizes multi-scale features, enabling the model to capture structural details at various molecular levels. DGDTA, on the other hand, incorporates the dynamic attention capabilities of GATv2, enhancing the representation of molecular graphs.

Despite the progress made in DTA prediction, there are still key challenges that hinder both accuracy and interpretability. Many existing methods employ the GAT model for drug feature learning which has static attention mechanisms, overlooks important contextual relationships in protein sequences, and relies on a single-scale representation (e.g., only a molecular graph or only a SMILES string sequences). Additionally, simple concatenation approaches used in these models often overlook the critical cross-interaction details between the drug and the target protein, limiting their predictive power. To address these limitations, this study introduces an improved method by integrating GATv2 for dynamic attention in drug molecular graphs, incorporating motif-level graphs for multi-scale drug representation, and leveraging BiLSTM for more comprehensive protein sequence encoding. A three-way multi-head attention mechanism is used for the interaction modeling between drug and protein features, capturing subtle yet crucial interaction cues that can improve predictive performance and Interpretability. Therefore, based on the problems discussed above, the research questions formed are:

1. How does the performance of an improved DTA prediction using a dynamic graph attention network (GATv2) and BiLSTM with multi-scale features and a three-way multi-head attention mechanism?

2. How do the parameters of the proposed method affect the model performance in DTA prediction?

3. How does the proposed method performance result in predicting DTA compared to baseline models, and in comparison to state-of-the-art models on benchmark datasets such as Davis and KIBA?

1.3 Objective and Hypotheses

1.3.1 Objectives

1. Developing an improved DTA prediction model using GATv2 and BiLSTM with multi-scale features and a three-way multi-head attention mechanism.
2. Evaluating the impact of the proposed method's parameters on DTA prediction performance.
3. Comparing the performance of the proposed method to baseline and state-of-the-art models on benchmark datasets such as Davis and KIBA.

1.3.2 Hypotheses

This study hypothesizes that combining GATv2 for drug encoding, multi-scale feature representation, BiLSTM-based protein encoding, and a three-way multi-head attention mechanism will enhance DTA performance. This hypothesis is supported by several premises. Recent studies have demonstrated that GATv2, with its dynamic attention mechanism, outperforms traditional static attention models by adapting the attention weights based on the specific context of the query node, leading to more expressive and context-aware representations of drug molecular graphs. Additionally, drug molecules have complex hierarchical structures, ranging from atomic to molecular levels, and previous research has indicated that multi-scale feature representations capture both the global structure and local motifs, thereby improving predictive accuracy. For protein sequence encoding, BiLSTM networks have proven effective in capturing long-range dependencies and contextual relationships within sequences, which is essential for understanding the interactions between proteins and drugs. Moreover, integrating attention mechanisms for the drug-target interaction has shown success in capturing subtle, yet crucial, cross-interactions between drug and target features, leading to more accurate affinity predictions. Based on these premises, it is expected that the proposed model will outperform baseline and existing benchmark methods, as demonstrated by improved evaluation metrics on widely recognized datasets such as Davis and KIBA.

1.4 Assumption

In this study, several key assumptions were made. First, the Davis and KIBA datasets are representative of real-world drug-target interactions and provide reliable binding affin-

ity measurements. Second, the motifs identified for drug substructures accurately capture crucial substructures relevant to binding between drug and target. Third, the BiLSTM model can effectively model contextual relationships within protein sequences, thereby enhancing the quality of protein encoding. Lastly, it is assumed that the model's hyperparameters, including those for GATv2, BiLSTM, and the attention heads, can be optimized within the available computational constraints to yield robust and reliable predictions.

1.5 Scope and Delimitation

1. **Principal Variables**

   (a) The independent variables include drug molecular graphs, drug motif graphs representing drug multi-scale structural features, and protein sequences encoded using BiLSTM.

   (b) The dependent variable is the predicted binding affinity, represented as a continuous value.

2. **Locale**
   The study utilizes two benchmark datasets, Davis and KIBA, which are publicly available and widely used in DTA research.

3. **Time frame**
   Research activities, including dataset preprocessing, model development, training, and evaluation, are conducted within the designated research period.

4. **Delimitation**
   This study is limited to sequence-based protein representations using BiLSTM; it does not incorporate structural protein data, such as 3D conformations. In addition, drug molecules are represented as 2D molecular graphs and motif graphs, excluding 3D conformational or quantum chemical properties. The focus remains on integrating multi-scale features and dynamic attention mechanisms to enhance DTA prediction accuracy. The scope is limited to computational methods and does not include real-scenario experimental validation of the predicted affinities for the drug and target.

1.6 Significance of the Study

This study makes a pivotal contribution to drug discovery research by improving the accuracy and interpretability of DTA predictions. By integrating GATv2 for dynamic graph attention, multi-scale drug motif graphs, and BiLSTM for comprehensive protein sequence encoding, the proposed model captures intricate relationships between drugs and their targets more effectively than existing methods. The incorporation of a three-way multi-head

attention mechanism further refines interaction modeling, allowing for the identification of critical substructures and amino acid residues that influence binding affinity. These improvements not only boost predictive performance but also deepen the understanding of drug and target interactions, thereby providing more information for decision-making during the early phases of drug development. Additionally, the study offers practical benefits by reducing the time and costs associated with experimental drug screening, enabling pharmaceutical researchers to prioritize promising candidates for further validation. Ultimately, this work sets the foundation for future methodological innovations in the cheminformatics and bioinformatics field, promoting more efficient and targeted approaches to discovering effective therapeutic agents.