CHAPTER 1

INTRODUCTION

This chapter highlights the fundamental challenge of identifying novel anticancer drugs that target the Cyclooxygenase-2 (COX-II) enzyme. It adds context to our research by highlighting the critical need for more efficient and precise techniques to predicting the bioactivity of potential drug candidates, which is a key bottleneck in traditional pharmaceutical development. The goal of this introduction is to present a computational strategy that uses Graph Convolutional Networks (GCNs) to overcome the limitations of earlier approaches. This chapter is broken into sections to provide a detailed case. The Rationale (Section 1.1) explains the medical necessity and scientific foundation for targeting COX-II. Sections 1.2 and 1.3 discuss scientific principles and research objectives. The chapter covers the problem statement, goals, and hypotheses (Sections 1.4 and 1.5) to lead the research. Finally, the chapter summarizes the study's assumptions, scope, and significance (Sections 1.6-1.8). In order to provide a more efficient method for computational drug discovery, this chapter describes the framework for developing and validating the GCN model, which will be covered in later chapters.

1.1 Rationale

Uncontrolled proliferation of abnormal cells is an indicator of cancer [1]. Globally, cancer continues to be the leading cause of death. Around 20 million new cases of cancer and 9.7 million deaths from the disease are anticipated globally in 2022, according to the World Health Organization (WHO) and the International Agency for Research on Cancer (IARC) [2]. It is estimated that there will be more than 35 million new cases by 2050 [3]. Recent projections indicate that by 2025, there will be 600,000 cancer-related deaths and two million new cases in the US [4]. While overall mortality rates are still falling, this progress is jeopardized by a number of worrying patterns that underscore the critical need for new therapeutic approaches.

Tumor-promoting inflammation is a biological process that uses the body's inflammatory pathways to promote tumor development, angiogenesis, and metastasis. The COX-II enzyme, overexpressed in malignant tissues but low in most normal tissues, plays a critical role in this process [5]. COX-II's fluctuating expression makes it an intriguing and verified target for the development of tailored anticancer therapies. However, the traditional drug development pipeline for identifying novel and effective inhibitors is a famously slow, costly, and high-risk enterprise, taking over a decade and billions of dollars to bring a single new medicine to market [6]. This inefficiency is a substantial barrier to translating promising biological targets like COX-II into therapeutic benefits.

The significant human and economic burden of cancer, both globally and in Indonesia, combined with the profound inefficiencies of the traditional drug discovery model, creates an urgent and unmet need for innovative strategies to accelerate the identification of new therapies [4]. COX-II expression patterns have been investigated in a variety of human cancer tissues, and they have been found to be often enhanced in many types of cancer, particularly colon cancer. Furthermore, the functional role of COX-II has been explored in encouraging tumor growth through processes such as angiogenesis and the production of new blood vessels, employing animal models and pharmacological inhibitors to validate COX-II as a therapeutic target for cancer [7]. The well-established role of COX-II in cancer progression presents a clear and appealing opportunity for targeted treatment.

To accelerate the drug development process, computational tools ranging from molecular docking to advanced deep learning models were employed. However, all strategies have limitations. Molecular docking and conventional QSAR typically employ simplified scoring systems or hand-crafted descriptors [8]. Modern deep learning models, such as Artificial Neural Networks (ANN) [9, 10] and Long Short-Term Memory (LSTM) [11, 12], still confront obstacles. ANNs reduce molecules to flat feature descriptors [13], deleting structural information, while LSTMs treat molecules as text sequences [14], ignoring their natural graph topology. This methodological gap, driven by both reliance on manual qualities and a failure to take advantage of molecules' inherent graph structure, highlights the critical need for a better strategy.

As a result, this study proposes developing a GCN—a cutting-edge artificial intelligence model—to accurately predict the bioactivity of potential COX-II inhibitors, with the goal of improving the first, most critical phase of drug discovery and accelerating the path to novel anti-cancer agents.

1.2 Theoretical Framework

This study follows the QSAR paradigm, which says that a compound's biological activity is directly proportional to its molecular structure [15]. Conventional QSAR techniques translate a molecule's structure into a fixed set of numerical values called molecular descriptors [16]. QSAR aims to speed up drug discovery by creating a statistical model that links these properties to known bioactivity.

However, the reliance on built descriptors poses a significant theoretical constraint. Reducing a complicated molecule, commonly represented as a 1D Simplified Molecular-Input Line-Entry System (SMILES) string [17], to a predetermined feature vector may obscure or lose crucial structural information. The model's prediction power is restricted by the quality and relevance of manually picked descriptors [18]. This underlying issue necessitates a more sophisticated technique capable of learning feature representations directly from the chemical network. This immediately leads to the consideration of GCNs

as a more appropriate theoretical framework for this task, as they are designed to function on graph-structured data without the need for defined feature engineering.

1.3 Conceptual Framework

Figure 1.1 depicts the multi-phase development and evaluation of a computational model to predict the bioactivity of COX-II inhibitors. The paradigm is essentially separated into three stages: input and foundation, development and evaluation, and output and contribution.

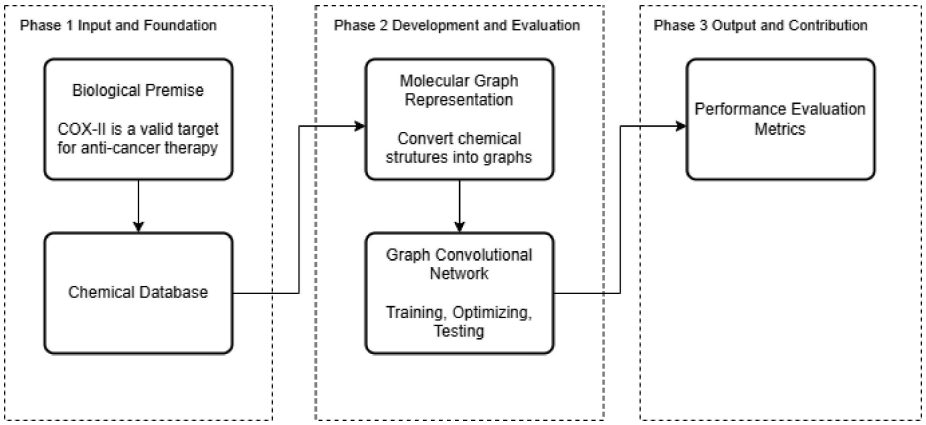

Figure 1.1: Research Conceptual Framework

1. **Phase 1: Input and Foundation**
   This initial step lays the scientific and data-driven foundation for the investigation. It starts with the Biological Premise, which states that COX-II is a scientifically validated and important target for anti-cancer therapy. This assumption guides the data collection method and supports the research emphasis. The next step after this assumption is to get information from the Chemical Database, ChEMBL [19]. Key raw materials for the model, a carefully selected collection of chemical compounds and their corresponding molecular structures, and experimentally verified bioactivity against the COX-2 target are all included in this database.

2. **Phase 2: Development and Evaluation**
   This stage explains the creation and assessment of the computational model, which is the technical core of the investigation. The first step in the process is Molecular Graph Representation, which preprocesses data from the chemical database. This crucial stage transforms linear chemical notations into a graph-based structure, where chemical bonds serve as edges and atoms as nodes. The GCN, a specialized deep learning architecture designed to learn directly from graph data, is then fed this graph-structured data. Iterative development of the GCN involves testing, optimization, and training. To ascertain the relationship between molecular structure

and bioactivity, the model is first trained on a subset of data. After that, the parameters are changed to improve the accuracy of the predictions. In order to ascertain the model's capacity for generalization, its performance is lastly thoroughly assessed using a dataset that has been seen before.

3. **Phase 3: Output and Contribution**
   By formally stating the research's conclusions and contributions, this last stage brings the study to a close. The final outcome of the GCN testing stage is a binary categorization for every molecule. In particular, the model gives a value that corresponds to its estimated ability to inhibit COX-II. A comprehensive set of Performance Evaluation Metrics is created based on these predictions in order to statistically assess the model's efficacy. This includes measurements such as F1-Score, MCC, AUC-ROC, recall, accuracy, and precision. The ability of the GCN model to distinguish between active and inactive inhibitors may be objectively and reliably assessed thanks to these data. All things considered, this study creates a quantitative benchmark for the application of GCNs in drug development and validates a particular computational methodology.

1.4 Statement of the Problem

The main issues in this study are:

1. Does a GCN-based strategy that learns directly from the molecular graph structure provide a more accurate depiction of the features necessary for COX-II inhibition than methods that employ manually created molecular descriptors?

2. Can a GCN-based predictive model correctly and consistently classify compounds as COX-II inhibitors or non-inhibitors? What is the predictive performance of the trained GCN model?

By investigating these issues, the researchers seek to improve the accuracy of bioactivity predictions, leading in more successful pharmaceutical development.

1.5 Objective and Hypotheses

The goal of this study is to develop a predictive model using GCNs to classify chemical substances as COX-II inhibitors or non-inhibitors by utilizing graph representations of COX-II molecules.

The hypotheses of this study include:

1. **A graph-based model outperforms a traditional manual feature-engineering technique in terms of prediction accuracy.**

Traditional Quantitative Structure-Activity Relationship (QSAR) models rely on manually created molecular descriptors, which might conceal or ignore key structural information needed for bioactivity prediction. citewang2015quantitative. In contrast, graph models, which identify atoms as nodes and bonds as edges, are believed to fully capture the essential information needed for predicting biological activity against a target like COX-II [20]. Using this structural approach, a Graph Convolutional Network (GCN) may learn feature representations directly from molecular graphs, bypassing the theoretical limits of descriptor-based approaches [21].

2. **A GCN methodology outperforms non-graph-based deep learning methods in terms of prediction accuracy.**
   Deep learning architectures, such as Artificial Neural Networks (ANNs) and Long Short-Term Memory (LSTM) networks, face inherent challenges in molecular modeling [9–12]. ANNs remove topological information by reducing molecules to flattened feature vectors, whereas LSTMs analyze them as linear text sequences (SMILES), discarding their inherent graph topology [13, 14]. Graph Convolutional Networks (GCNs) are a theoretical framework meant to operate on graph-structured data, directly addressing the methodological failure to use a molecule's inherent structure [21].

1.6 Assumption

This study is based on the following key assumptions.

- **Data Quality and Integrity**, the bioactivity data from the public database, ChEMBL [19], is presumed to be accurate and credible. The experimental procedures utilized to create the data were valid and yielded consistent results, and any related molecular structure information accurately represented the tested substances.
- **Adequacy of molecular representation**, it is assumed that describing molecules as graphs, where atoms are nodes and bonds are edges, captures the necessary structural information to predict their biological activity against COX-II [22].

1.7 Scope and Limitation

This study focuses on the creation and deployment of a deep learning system for identifying potential anti-cancer drugs that target the COX-II enzyme. At the heart of this research is the creation, training, and validation of a GCN, a customized deep learning model capable of learning from molecular structures. To train and test this model, the study will use publicly available biochemical data obtained from an existing chemical database. The key

goals are to create a reliable and generalizable predictive model, as well as discover novel compounds with a high anticipated inhibitory potential against COX-2.

Several distinct boundaries have been created to assure the study's focus and achievement of its aims. Crucially, this research is totally computational in nature. It will not include the chemical synthesis and experimental validation of the molecule identified. Furthermore, the investigation is primarily focused on the COX-II target. Finally, the model's accuracy and applicability are inextricably linked to the quality and breadth of data available in the public domain during the study period.

1.8 Significance of the Study

This study has great value since it adds new knowledge at the junction of artificial intelligence and cancer therapy development. The discoveries are intended to benefit numerous specific groups by providing a more conclusive and efficient method for identifying novel medicinal drugs. The fundamental contribution of this research is the creation and validation of a specific GCN model for predicting the bioactivity of COX-II inhibitors. This makes a novel contribution by bringing cutting-edge deep learning architecture to a vital and specialized anti-cancer target. This work has significant practical implications for medicinal chemistry and discovery researchers. By effectively filtering and selecting chemicals, the methodology allows scientists to concentrate their limited resources on producing and evaluating only the most promising possibilities.

For future researchers and academics, this thesis will serve as a platform for further research, whether by testing confirming anticipated molecules, refining the GCN design, or adapting the methodology to various biological challenges. This work advances the long-term goal of creating new cancer medicines by improving the efficiency of the discovery phase.