ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)
A subsidiary organ of Organisation of Islamic Cooperation (OIC)

BACHELOR OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

GNN and Transformer Fusion Learning for Molecular Classification of BACE1 Inhibitors

Md. Abu Hena Shadid
200041101
Mahajabin Tabassum
200041132

Department of Computer Science and Engineering
Islamic University of Technology

September, 2025

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)
A subsidiary organ of Organisation of Islamic Cooperation (OIC)

GNN and Transformer Fusion Learning for Molecular Classification of BACE1 Inhibitors

Md. Abu Hena Shadid
200041101
Mahajabin Tabassum
200041132

Department of Computer Science and Engineering
Islamic University of Technology
September, 2025

Declaration of Candidates

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by **Md. Abu Hena Shadid** and **Mahajabin Tabassum** under the supervision of **Tareque Mohmud Chowdhury**, Assistant Professor, Department of Computer Science and Engineering and **Njayou Youssouf**, Lecturer, Department of Computer Science and Engineering, Islamic University of Technology, Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others have been acknowledged in the text and a list of references is given.

Tareque Mohmud Chowdhury
Assistant Professor
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: September 29, 2025

Njayou Youssouf
Lecturer
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: September 29, 2025

Md. Abu Hena Shadid
Student ID: 200041101
Date: September 29, 2025

Mahajabin Tabassum
Student ID: 200041132
Date: September 29, 2025

Dedicated to our parents

Contents

List of Figures

List of Tables

List of Abbreviations

AD	Alzheimer's Disease
APP	Amyloid Precursor Protein
BACE1	beta-site amyloid precursor protein cleaving enzyme 1
GNN	Graph Neural Network
BERT	Bidirectional Encoder Representations from Transformers
SMILES	Simplified Molecular Input Line Entry System
CNN	Convolutional Neural Network
GAT	Graph Attention Network
GCN	Graph Convolution Network
HTS	High-Throughput Screening
QSAR	Quantitative Structure–Activity Relationship

Acknowledgement

I am profoundly grateful to my supervisor, Njayou Youssouf, for his endless patience, insightful critiques, and for always being there for us when we needed any help. His expertise and encouragement were instrumental in the completion of this thesis.

A special thanks to Mr. Abdel Kader Gelani, who provided invaluable support.

Finally, to my parents, whose love and support have been a constant source of strength. Thank you for always being there, for pretending to understand my research when I rambled on, and for not asking too many questions about why I was still not employed after all these years.

To all of you, I extend my heartfelt appreciation.

Abstract

Alzheimer's disease (AD) is a progressive and devastating neurodegenerative disorder, primarily manifested through memory loss and cognitive decline [1], [2]. One of the central pathological hallmarks of AD is the accumulation of amyloid-beta (Aβ) plaques, formed via the sequential cleavage of the amyloid precursor protein (APP) by β-secretase (BACE1) and γ-secretase [3]. Inhibiting BACE1 is therefore regarded as a compelling therapeutic strategy, as it can impede the formation of neurotoxic Aβ aggregates [4], [5]. Nevertheless, the identification of effective BACE1 inhibitors remains arduous and resource-intensive when approached through conventional experimental pipelines. In this study, we propose a hybrid deep learning framework that fuses Graph Neural Networks (GNNs) with ChemBERTa, a transformer model pretrained on large chemical corpora. While GNNs capture atom-level and bond-level interactions (local structural dependencies), ChemBERTa encodes long-range dependencies and semantic patterns from SMILES representations (global chemical context). By unifying these complementary modalities, our model overcomes the limitations of prior GNN+CNN approaches, where CNNs process sequential SMILES in a strictly local fashion and fail to capture non-linear long-range dependencies across molecular structures. Our GNN–ChemBERTa fusion model achieved an accuracy of 92.77% in classifying active versus inactive BACE1 inhibitors, demonstrating superior predictive power and generalization. Beyond its performance, the model contributes to reducing drug discovery costs, accelerating virtual screening, and minimizing the need for extensive laboratory experimentation. Moreover, a recall value of 93% indicates that almost all potential active molecules were successfully identified by the model, minimizing the risk of missing true inhibitors. Similarly, a high precision value of 93% demonstrates that the model produces very few false positives, thereby reducing unnecessary laboratory costs associated with testing inactive compounds. Additionally, the ROC–AUC score of 87.88% confirms that the model can effectively distinguish between active and inactive molecules, reflecting strong overall classification performance. By enabling efficient in silico identification of potential inhibitors, this approach not only streamlines the early stages of Alzheimer's drug development but also holds promise for broader application to other therapeutic targets associated with neurodegenerative diseases.

Chapter 1

Introduction

Alzheimer's disease (AD) afflicts millions worldwide and represents one of the most urgent biomedical challenges of our era [6]. A cardinal feature of AD is the aggregation of amyloid-beta (Aβ) plaques, formed when BACE1 catalyzes the cleavage of the amyloid precursor protein (APP). Overactivity of BACE1 accelerates Aβ accumulation, which in turn induces synaptic dysfunction and cognitive deterioration. Accordingly, BACE1 inhibitors have been the focus of intensive therapeutic research [7].

Traditional drug discovery pipelines rely on high-throughput screening (HTS), which, despite their experimental rigor, are prohibitively costly and time-consuming. Computational paradigms such as Quantitative Structure–Activity Relationship (QSAR) modeling have emerged as indispensable tools for accelerating discovery by correlating molecular descriptors with biological activity [8]. Ensemble learning models (e.g., Random Forest, AdaBoost, Gradient Boosting) have further improved classification performance, but they depend heavily on handcrafted descriptors, which introduce human bias and fail to fully capture the intrinsic structural complexity of molecules [7].

Deep learning methods have advanced the state-of-the-art by learning directly from molecular graphs or sequences. Graph Neural Networks (GNNs) excel at modeling molecules as graphs of atoms and bonds, effectively capturing local atomic interactions [9]. Convolutional Neural Networks (CNNs), when applied to SMILES strings, have achieved competitive results by recognizing local sequential motifs [10]. Yet, CNNs fundamentally lack the capacity to capture long-range dependencies, since their convolutional filters are inherently local. As a result, prior GNN+CNN fusion approaches [10] remain limited, because while GNNs contribute local graph structure, CNNs provide only local sequence patterns without accounting for the global

context of the molecule.

Transformer-based models such as ChemBERTa overcome this limitation. Trained on millions of SMILES notations, ChemBERTa learns contextual embeddings that capture both local and non-local interactions via self-attention [6]. Unlike CNNs, transformers can dynamically attend to distant tokens in a sequence, thereby encoding global structural semantics. Hence, fusing GNNs with ChemBERTa allows for a comprehensive molecular representation: GNNs provide atom-level and bond-level granularity, while transformers contribute holistic, global dependencies across the molecule.

Building upon these insights, we introduce a GNN–ChemBERTa fusion model for BACE1 inhibitor classification. Specifically: (1) We design a multimodal fusion framework that leverages both graph-based atomic features and transformer-based SMILES embeddings. (2) We compare its efficacy against prior QSAR, ensemble learning, and GNN+CNN baselines, particularly the work using GNN and CNN fusion [10], which represents the closest benchmark. (3) We demonstrate that our approach not only achieves superior accuracy but also provides a scalable path for extension to other Alzheimer's-related therapeutic targets.

1.1 Motivations and Scope

The motivation for this research stems from the need to address the limitations of current predictive models in Alzheimer's drug discovery. Despite significant advances, most existing studies still struggle with relatively low prediction accuracies, particularly when compared to models based on more traditional techniques [6]. As drug discovery relies heavily on the accurate prediction of molecular properties, enhancing model accuracy through novel methods is essential. This work aims to bridge the gap by employing a hybrid approach that combines Graph Neural Networks (GNNs) and Sequence-based models (such as ChemBERT), offering improved performance over previous methods. By leveraging both structural and sequential data, this research intends to contribute to more accurate and interpretable models for predicting BACE1 inhibitors, with broader implications for other therapeutic targets involved in AD [6].

The scope of this research encompasses the exploration of advanced machine learning techniques, particularly focusing on the integration of molecular graph data (e.g., GNN) and SMILES string data (e.g., ChemBERT). In addition to improving BACE1 inhibitor predictions, the study also explores the feasibility of extending this method-

ology to other related proteins in the Alzheimer's disease pathway, potentially accelerating drug discovery processes.

1.2 Problem Statement

The primary problem addressed in this thesis is the challenge of improving the accuracy and interpretability of BACE1 inhibitor prediction models for Alzheimer's disease. Current predictive models face several shortcomings, such as low prediction accuracy and limited generalizability to other related proteins. The research objectives are as follows:

1. To develop a hybrid model that integrates Graph Neural Networks (GNNs) and Sequence-based models for improved prediction of BACE1 inhibitors.
2. To evaluate the performance of the proposed model in comparison to existing methods, with a particular focus on accuracy, interpretability, and other metrics.
3. To explore the broader application of this approach in drug discovery, particularly for Alzheimer's disease-related proteins.

These objectives align with the goal of enhancing drug discovery efforts for Alzheimer's disease and improving the tools available for researchers in the field.

1.3 Research Challenges

The main challenge for our research is to effectively use GNN and ChemBERT individually and then combine them together for fusion learning. GNN and ChemBERT, two totally different models process two different representation of the same input, one for graph and other for strings, work in different way. Combining their result effectively is the main challenge here. [11]

1.4 Contribution

This work bridges deep learning and molecular science, making drug discovery processes smarter, faster, cheaper, and more sustainable, with potential real-world impact on healthcare and society.

- **Improved Representation Fusion:**
  Proposed a Feature Extraction Module to dynamically adjust the importance

of different feature types (graph and sequence) before fusion, leading to more discriminative representations.

- **Faster and Smarter Drug Discovery:**
  Accelerates molecular screening by quickly predicting BACE1 inhibitor activity, reducing experimental costs and saving years of lab work compared to traditional methods.
- **Resource and Time Efficiency:**
  Enables virtual screening of millions of molecules within minutes, helping research labs and pharmaceutical companies prioritize promising candidates early, saving tremendous time and manpower.
- **Health and Society Impact:**
  Contributes to the development of new treatments for Alzheimer's Disease by identifying potential BACE1 inhibitors, offering hope for better patient outcomes and reducing the global burden of neurodegenerative diseases.

Chapter 2

Related Works

The quest to develop effective inhibitors for beta-site amyloid precursor protein cleaving enzyme 1 (BACE1), a key target in Alzheimer's disease (AD) therapeutics, has evolved significantly over the decades [3]. AD, characterized by the accumulation of amyloid-beta plaques in the brain, affects millions worldwide, and BACE1 plays a pivotal role in the production of these plaques by cleaving the amyloid precursor protein [12]. Early efforts in drug discovery for BACE1 inhibitors relied on traditional computational methods, gradually transitioning to advanced deep learning techniques. This section traces this evolution, highlighting foundational works, recent advancements, and persistent research gaps. We begin with classical Quantitative Structure-Activity Relationship (QSAR) models and ensemble learning, move to graph-based representations via Graph Neural Networks (GNNs) and their hybrids, explore transformer-based architectures for molecular modeling, and conclude with identified gaps that our hybrid GNN-ChemBERTa model addresses.

Traditional Approaches: QSAR and Ensemble Learning

The foundation of computational drug discovery for BACE1 inhibitors was laid in the early 2000s with QSAR models, which correlate molecular structures with biological activities through statistical relationships. QSAR methodologies, pioneered by Hansch and Fujita in the 1960s [13], involve computing molecular descriptors—such as topological indices, physicochemical properties, and fingerprints—and using regression or classification algorithms to predict activity. In the context of AD, these models were instrumental in screening large compound libraries for potential BACE1 inhibitors. One seminal application was by Ponzoni et al. in 2019, who developed QSAR classification models to predict the activity of BACE1 inhibitors using a dataset of over

1,500 compounds [7]. Their approach employed support vector machines (SVM) and random forests on descriptors like ECFP fingerprints, achieving accuracies around 80%. This work demonstrated QSAR's utility in identifying active compounds but also underscored its limitations: reliance on handcrafted features, which may overlook subtle structural nuances critical for BACE1 binding. Similarly, Noviandy et al. in 2023 advanced this by integrating ensemble learning techniques, including Random Forest (RF), AdaBoost, and Gradient Boosting (GB), on a BACE1 dataset [8]. Ensemble methods, originally formalized by Breiman for RF in 2001 [14] and Freund and Schapire for AdaBoost in 1997 [15], aggregate multiple weak learners to improve robustness. Noviandy's models outperformed single classifiers, with GB achieving an AUC-ROC of 0.85, highlighting ensemble learning's ability to handle imbalanced datasets common in bioactivity prediction. Further refinements include the work by Askr et al. in a 2022 comprehensive review, which analyzed over 50 QSAR studies on BACE1 and emphasized the need for more adaptive models [16]. They noted that while QSAR excels in interpretability—allowing chemists to link descriptors to molecular modifications—it struggles with novel scaffolds outside the training distribution. Additional studies, such as those by Kumar et al. in 2020 using 2D-QSAR on heterocyclic compounds [17], and Flavonoids as BACE1 inhibitors by Gupta et al. in 2020 [18], reinforced QSAR's role but called for integration with emerging deep learning paradigms to automate feature extraction. Despite these advancements, traditional QSAR and ensemble methods are constrained by their dependence on predefined descriptors, limiting generalization to diverse chemical spaces. As drug discovery datasets grew, propelled by initiatives like ChEMBL [19], the field shifted toward representation learning, where models learn features directly from raw data.

Graph Neural Networks and CNN Hybrids

The advent of deep learning in the 2010s marked a paradigm shift, with Graph Neural Networks (GNNs) emerging as a powerful tool for molecular property prediction. GNNs treat molecules as graphs, with atoms as nodes and bonds as edges, enabling the capture of topological relationships. The foundational Graph Convolutional Network (GCN) was introduced by Kipf and Welling in 2017, extending convolutional operations to graph-structured data via spectral graph theory [20]. This allowed message passing between neighboring nodes, aggregating local features into global representations. Building on this, Veličković et al. proposed Graph Attention Networks (GAT) in 2017, incorporating attention mechanisms to weigh neighbor importance dynamically [21]. In molecular applications, GNNs have excelled in tasks like solubility prediction and toxicity assessment. For BACE1 specifically, Song et

al. in 2024 developed a GNN+CNN hybrid for inhibitor classification, representing molecules as graphs for GNN processing and SMILES strings for CNN feature extraction [22]. Their model achieved superior accuracy (91.11%) compared to QSAR baselines, demonstrating multimodal fusion's benefits. Wojtuch et al. in 2021 further enhanced GNN performance by optimizing atom featurization, incorporating hybridization and aromaticity [23]. Hybrid models combining GNNs with Convolutional Neural Networks (CNNs)—originally popularized by LeCun et al. in 1998 for image recognition [24]—address complementary aspects: GNNs handle irregular graph structures, while CNNs process sequential data like SMILES. A notable example is the GSFL model by Li et al. in 2024, fusing graph and multi-level sequence features for BACE1 activity prediction [25]. Surveys like that by Wieder et al. in 2020 reviewed over 80 GNN variants for molecular properties, noting their outperformance on benchmarks like MoleculeNet [26]. Recent innovations include Kolmogorov–Arnold GNNs by Liu et al. in 2024 for improved expressivity [27] and chain-aware GNNs by Wang et al. in 2024 for better path-based reasoning [28]. However, CNNs in hybrids are limited to local receptive fields, missing long-range dependencies in molecular sequences. This has spurred interest in transformers, which offer global context modeling.

Transformer-Based Molecular Modeling

Transformers, introduced by Vaswani et al. in 2017 for natural language processing [29], revolutionized sequence modeling with self-attention mechanisms that capture dependencies regardless of distance. In cheminformatics, transformers adapt to SMILES strings, treating them as "sentences" of chemical tokens. Pioneering models include SMILES-BERT by Wang et al. in 2019, pre-trained on masked language modeling to learn contextual embeddings [30]. MolBERT by Fabian et al. in 2020 extended this with additional pre-training tasks like property prediction [31]. ChemBERTa, developed by Chithrananda et al. in 2020 and fine-tuned on ZINC datasets [32], stands out for its efficiency in downstream tasks like bioactivity classification. These models outperform CNNs by modeling global semantics, such as ring structures spanning distant atoms. Recent surveys, such as by Uludoğan et al. in 2024 on transformers for SMILES representation [33], and Cannham et al. in 2024 on their applications in drug discovery [6], highlight their transformative potential. For instance, ChemLM by Heid et al. in 2024 integrates domain-adaptable pre-training [34], while Smile-to-Bert by Sharma et al. in 2024 predicts RDKit descriptors [35]. In BACE1 contexts, transformers have been used in hybrid setups, but standalone applications remain limited. The integration of transformers with GNNs is nascent, with works like Meta-GTNRP

by Torres et al. in 2024 for few-shot learning [36], showing promise in combining graph locality with sequence globality.

Research Gaps and Future Directions

Despite these strides, significant gaps persist in BACE1 inhibitor prediction. Traditional QSAR and ensembles, while interpretable, lack automatic feature learning [16]. GNN-CNN hybrids improve accuracy but fail to capture long-range dependencies [22]. Transformer models excel in sequences but overlook explicit graph topology [33]. Few studies fuse GNNs with transformers for BACE1, as noted in surveys by Jiang et al. in 2024 on GNNs for drug discovery [37] and Li et al. in 2024 on deep learning gaps [38]. Interpretability is another underexplored area; while attention maps in GATs and transformers offer insights, systematic analysis in BACE1 contexts is rare [23]. Our work bridges these gaps by proposing a hybrid GNN-ChemBERTa model with vector gating for adaptive fusion, enhancing accuracy and interpretability. Future directions include incorporating 3D conformations [27], multi-target prediction for AD polypharmacology [16], and generative models for de novo BACE1 inhibitors [35]. As datasets expand and models evolve, this integration promises accelerated AD therapeutics discovery.

Chapter 3

Proposed Methodology

The proposed architecture for BACE-1 inhibitor classification is designed around a dual-branch multimodal framework that integrates graph-based and sequence-based molecular representations. This combination enables the model to capture both the detailed structural topology of molecules and the semantic patterns inherent in their SMILES sequences. The overall workflow is shown in figure 3.1

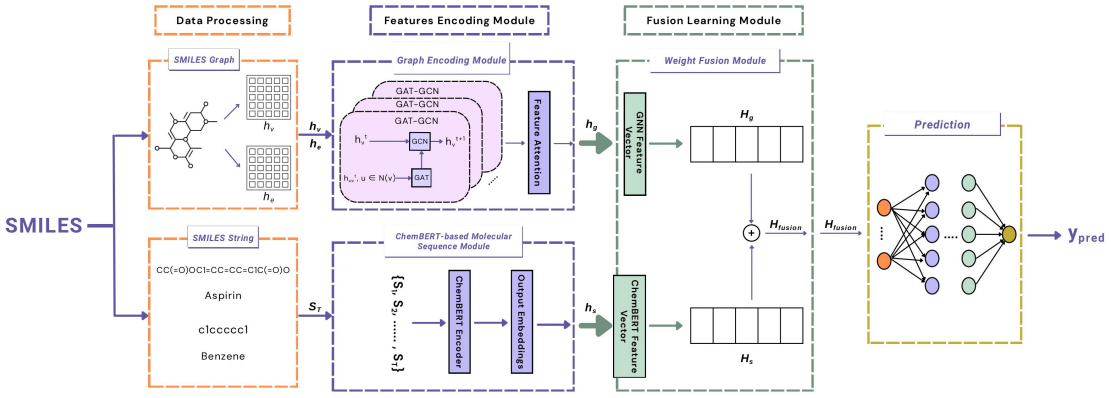

Figure 3.1: Full Architecture

3.0.1 Model Architecture

The model is a hybrid deep learning architecture that integrates the strengths of Graph Neural Networks (GNNs) and transformer-based models, specifically ChemBERTa, to effectively capture both the structural and sequential representations of molecules for property prediction tasks. This hybrid approach is motivated by the limitations of individual paradigms: GNNs excel at modeling local graph structures and atomic interactions but may overlook long-range dependencies in sequential representations, while transformers like ChemBERTa are adept at processing sequential data such as SMILES strings but do not inherently capture graph topology. By fusing these complementary representations, the model aims to achieve superior performance in binary classification tasks, such as predicting molecular activity against the BACE-1 enzyme.

The architecture consists of three primary components: (1) a GNN module for graph-based feature extraction, (2) a ChemBERTa module for sequence-based embeddings, and (3) a fusion mechanism with a vector gate followed by classification layers. Each component is designed with specific hyperparameters to balance computational efficiency and expressive power. The GNN uses a hidden dimension of 64 and 10 attention heads in the GAT layer to allow multi-faceted attention mechanisms, while ChemBERTa leverages a pretrained base model with 768 hidden dimensions. The fusion combines these into a 832-dimensional vector (64 from GNN + 768 from ChemBERTa), modulated by a vector gate to dynamically weight features based on their relevance. This design choice is grounded in empirical evidence from prior work, where gated fusions have shown improvements in multimodal learning tasks [39].

Logical arguments for this architecture include the need for multimodal integration in cheminformatics, where molecules can be represented in multiple ways (graphs vs. strings), and hybrid models have demonstrated state-of-the-art results in benchmarks like MoleculeNet [40]. Parameter selection, such as the hidden size and dropout rate of 0.2, was informed by standard practices in GNN and transformer literature to mitigate overfitting while ensuring sufficient capacity for the BACE dataset's 1,513 samples. Theoretically, this architecture aligns with the principle of representation learning, where combining diverse feature spaces (graph spectral features and contextual embeddings) enhances the model's ability to generalize across molecular structures, as supported by the universal approximation theorem for neural networks [41].

GNN Module

The GNN module is responsible for encoding the molecular graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of atoms (nodes) and $\mathcal{E}$ the set of bonds (edges).

Each atom and bond is initially described by a set of handcrafted features. For atoms, these features include atomic symbol (e.g., C, H, O), valence electrons, degree, and hybridization state, directly characterizing their bonding capabilities and geometric preferences. For bonds, features such as bond type (single, double, triple) and geometric configuration are considered to capture the strength and nature of atomic interactions.

The GNN module processes node features $\mathbf{X} \in \mathbb{R}^{n\times 5}$ and edge attributes to produce a graph-level representation $\mathbf{h}'_{\text{gnn}} \in \mathbb{R}^{64}$. The choice of GNN is justified by its ability to propagate information through graph convolutions, capturing local chemical interactions like bond types and aromaticity, which are critical for molecular properties.

The architecture combines Graph Attention Network (GAT) [21] and Graph Convolutional Network (GCN) [42] layers. GAT allows the model to assign different importance to neighboring nodes via attention mechanisms, enhancing selectivity in feature aggregation, while GCN provides stable convolutional propagation. This combination is logical as GAT's attention can focus on key substructures (e.g., functional groups), and GCN ensures robust neighborhood averaging, reducing sensitivity to noise in molecular graphs.

The theoretical foundation of the GNN module rests on spectral graph theory and the Weisfeiler-Lehman (WL) hierarchy, which provide a framework for understanding the expressive power of graph neural networks. Spectral graph theory, particularly the graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{A}$ (where $\mathbf{D}$ is the degree matrix and $\mathbf{A}$ the adjacency matrix), underpins GCN's convolution operation. The normalized Laplacian $\mathbf{L}_{\text{norm}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ approximates low-pass filtering of graph signals, enabling the model to smooth node features across connected atoms, as formalized by:

$$\mathbf{x}'_i = \sigma\left(\sum_{j\in\mathcal{N}(i)} \frac{1}{\sqrt{\deg(i)\ \deg(j)}}\mathbf{W}\mathbf{x}_j\right),$$

where σ is a non-linear activation. This aligns with the heat kernel smoothing on graphs, which preserves local structure while reducing noise, a property critical for molecular graphs with varying bond orders.

The GAT layer extends this by introducing attention, inspired by the WL test's abil-

ity to distinguish graph structures based on neighborhood labeling. The attention mechanism can theoretically approximate the WL test up to a certain isomorphism class, enhancing the GNN's ability to differentiate molecular subgraphs (e.g., rings vs. chains). The multi-head design increases expressive power, approaching the 1-WL limit, as each head can focus on different spectral components of the graph. This theoretical grounding supports the module's capacity to capture both local chemical environments and global structural motifs, making it suitable for the BACE dataset's diverse molecular structures.

Node Feature Projection Initial node features, consisting of atomic number, degree, formal charge, hybridization, and aromaticity, are projected to a higher-dimensional space to increase expressivity. The projection is performed using a linear layer followed by ReLU activation:

$$\mathbf{X}' = \mathrm{ReLU}\Big(\mathbf{W}_{\text{proj}}\mathbf{X} + \mathbf{b}_{\text{proj}}\Big),$$

where $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{64\times 5}$ and $\mathbf{b}_{\text{proj}} \in \mathbb{R}^{64}$ are learnable parameters. This step maps the low-dimensional input (5 features) to a 64-dimensional hidden space, allowing the model to learn more complex representations. The ReLU activation introduces non-linearity, preventing the vanishing gradient problem in deeper networks. The hidden dimension of 64 was selected as a balance between model capacity and computational cost, based on common practices in GNNs for small datasets like BACE, where larger dimensions (e.g., 128) might lead to overfitting.

GAT Layer Following projection, a GAT layer with 10 attention heads and dropout of 0.2 is applied to aggregate neighborhood information adaptively:

$$\mathbf{X}'' = \mathrm{ReLU}\left(\mathrm{GATConv}(\mathbf{X}', \text{edge\_index}, \text{heads} = 10, \text{dropout} = 0.2)\right),$$

where GATConv computes attention coefficients for each edge. Mathematically, for each head k, the attention coefficient $\alpha_{ij}^{(k)}$ between nodes i and j is:

$$\alpha_{ij}^{(k)} = \frac{\exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^{(k)T}[\mathbf{W}^{(k)}\mathbf{x}_i' || \mathbf{W}^{(k)}\mathbf{x}_j']\right)\right)}{\sum_{l\in\mathcal{N}(i)}\exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^{(k)T}[\mathbf{W}^{(k)}\mathbf{x}_i' || \mathbf{W}^{(k)}\mathbf{x}_l']\right)\right)}$$

, where $\mathbf{a}^{(k)}$ is the attention vector, $\mathbf{W}^{(k)}$ the weight matrix for head k, $||$ concatenation, and $\mathcal{N}(i)$ the neighborhood of i. The output for each node is the concatenation of head outputs, resulting in $\mathbf{X}'' \in \mathbb{R}^{n\times 640}$ (64×10). The 10 heads enable multi-view learning,

where each head captures different relational aspects, justified by prior work showing improved performance in graph tasks [21]. Dropout of 0.2 regularizes the attention, preventing co-adaptation of heads and improving generalization on noisy molecular data.

The flowchart 3.2 illustrates the operational steps of the Graph Attention Network (GAT) module within the model, processing molecular graph data for feature extraction.

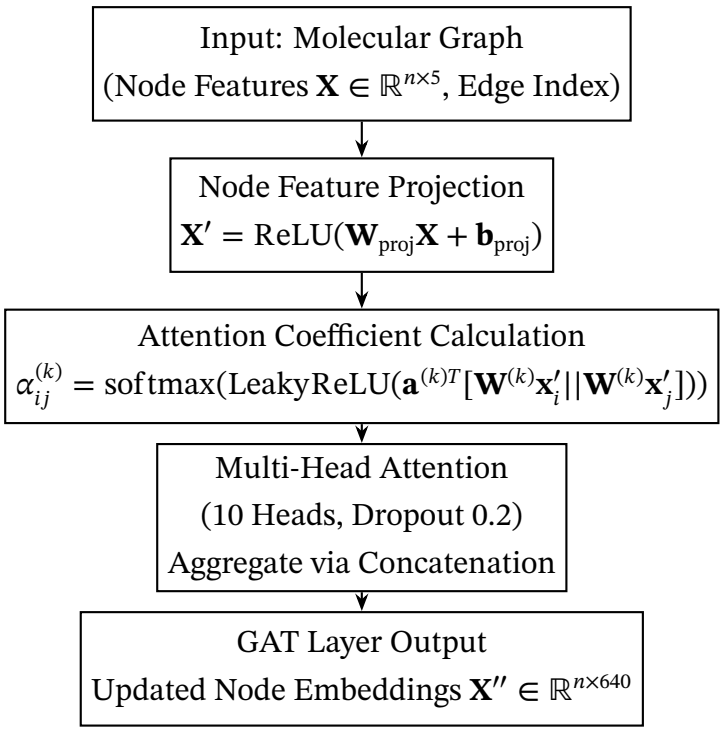

Figure 3.2: Flowchart of the GAT Module Operation

GCN Layer A GCN layer then refines the embeddings by performing symmetric normalized convolution:

$$\mathbf{X}''' = \mathrm{ReLU}\left(\mathrm{GCNConv}(\mathbf{X}'', \text{edge\_index})\right),$$

where GCNConv updates node features as:

$$\mathbf{x}'''_i = \mathbf{W}_{\text{gcn}} \left(\mathbf{x}''_i + \sum_{j\in\mathcal{N}(i)} \frac{1}{\sqrt{\deg(i)\deg(j)}} \mathbf{x}''_j \right) + \mathbf{b}_{\text{gcn}},$$

with $\mathbf{W}_{\text{gcn}} \in \mathbb{R}^{640\times 640}$ and $\mathbf{b}_{\text{gcn}} \in \mathbb{R}^{640}$. This layer maintains the dimension $\mathbf{X}''' \in \mathbb{R}^{n\times 640}$. The GCN's normalization stabilizes training, and its inclusion after GAT combines attention-based selection with uniform convolution, enhancing robustness as

per studies on GNN ensembles [42]. Dropout of 0.2 is applied post-layer to further mitigate overfitting.

The flowchart 3.3 illustrates the operational steps of the Graph Convolutional Network (GCN) module within the model, refining molecular graph embeddings.

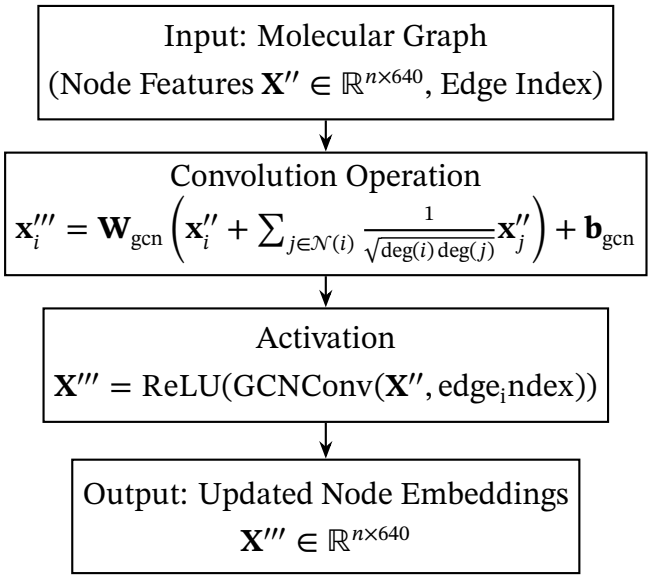

Figure 3.3: Flowchart of the GCN Module Operation

Global Pooling To obtain a graph-level representation, global max and mean pooling are concatenated:

$$\mathbf{h}_{\text{gnn}} = \left[\text{global\_maxpool}(\mathbf{X}''', \text{batch}) \,||\, \text{global\_meanpool}(\mathbf{X}''', \text{batch})\right],$$

yielding $\mathbf{h}_{\text{gnn}} \in \mathbb{R}^{1280}$ (640×2). This is projected to 64 dimensions:

$$\mathbf{h}'_{\text{gnn}} = \text{ReLU}(\mathbf{W}_{\text{gnn}}\mathbf{h}_{\text{gnn}} + \mathbf{b}_{\text{gnn}}),$$

where $\mathbf{W}_{\text{gnn}} \in \mathbb{R}^{64\times 1280}$ and $\mathbf{b}_{\text{gnn}} \in \mathbb{R}^{64}$. Max pooling captures salient features (e.g., critical atoms), while mean pooling provides average context, a standard approach in GNNs for permutation-invariant representations [43] [44]. The final projection reduces dimensionality for efficient fusion, with ReLU adding non-linearity.

The flowchart 3.4 illustrates how the Graph Attention Network (GAT) and Graph Convolutional Network (GCN) work together in the GNN module for processing molecular graphs.

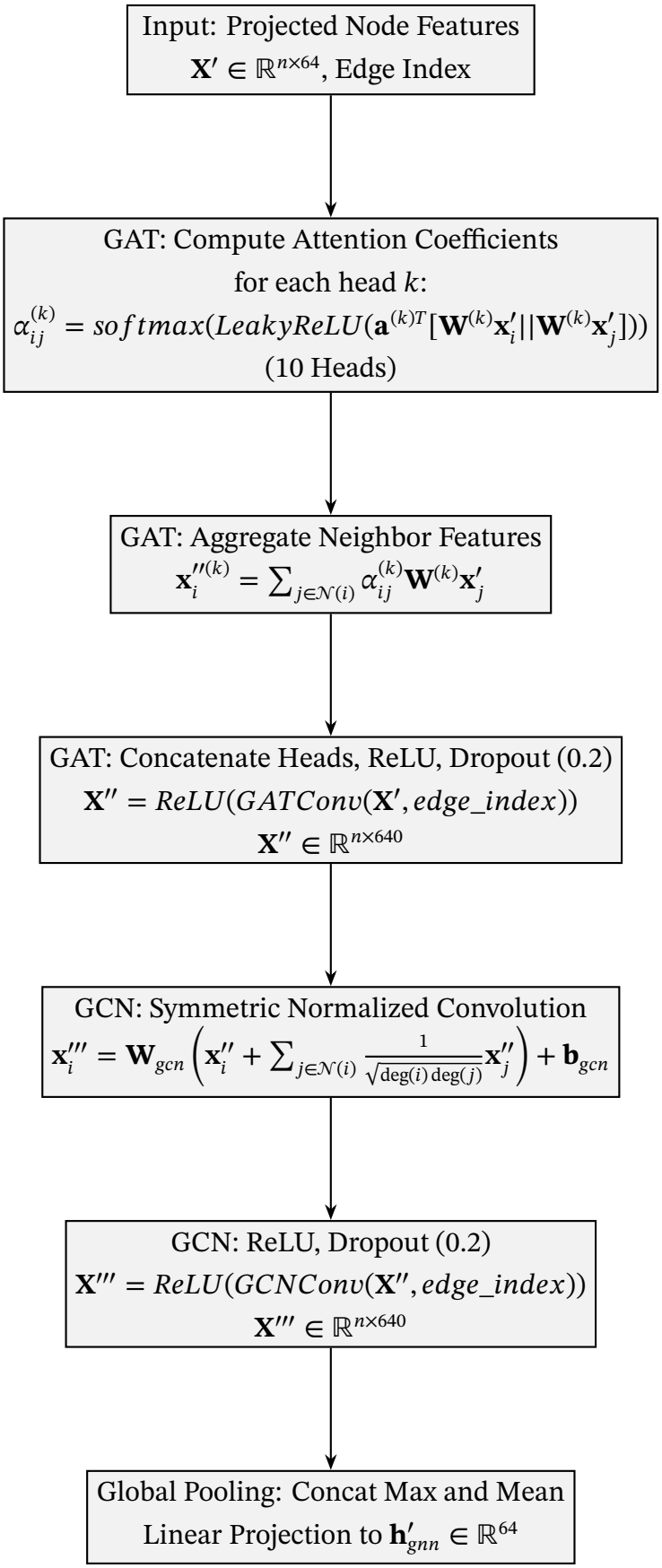

Figure 3.4: Detailed Flowchart of GAT and GCN Integration in the GNN Module

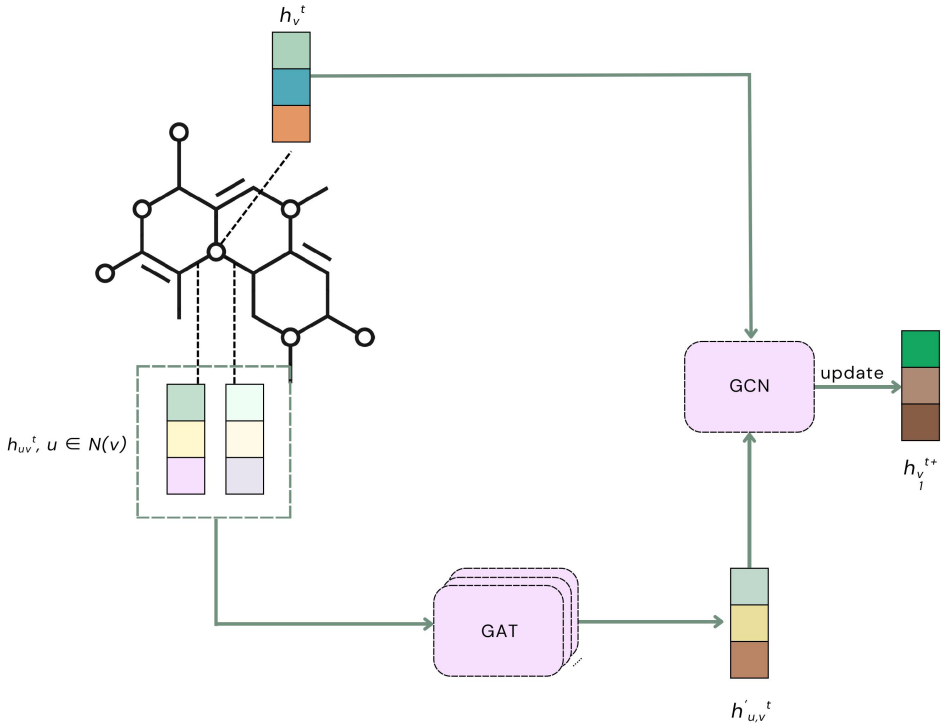

Figure 3.5: Node Feature Update

ChemBERTa Module

The ChemBERTa module handles the sequential representation of molecules using SMILES strings, leveraging a pretrained transformer model to capture contextual embeddings. ChemBERTa is based on the RoBERTa architecture [45], adapted for chemistry by pretraining on large corpora of molecular strings. Specifically, the "seyonec/ChemBERTa-zinc-base-v1" variant is used, which features 6 transformer layers, 12 attention heads per layer, and a hidden dimension of 768, resulting in approximately 85 million parameters (similar to RoBERTa-base but tuned for chemical data).

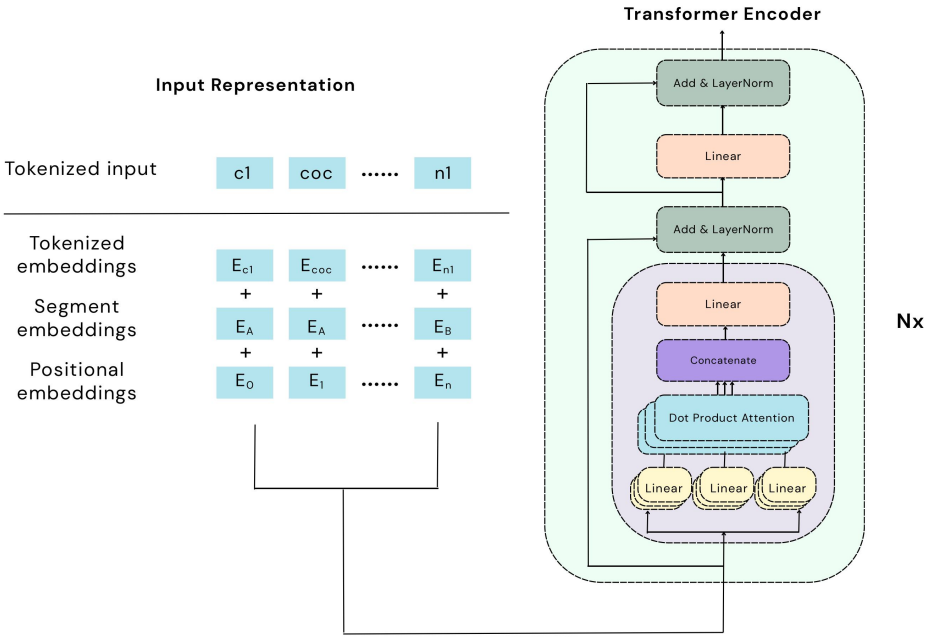

Figure 3.6: ChemBERTa Module

The architecture consists of an embedding layer followed by stacked transformer encoder layers. Input SMILES are tokenized using a custom SmilesTokenizer or Byte-Pair Encoding (BPE), with a maximum sequence length of 128 and vocabulary size of up to 52K. Each token is embedded into a 768-dimensional vector, augmented with positional encodings to preserve order. The transformer layers apply self-attention:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V},$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are query, key, and value projections of the input, and $d_k = 768/12 = 64$ is the head dimension. Multi-head attention (12 heads) allows parallel computation of different subspaces, enabling the model to capture diverse relationships in SMILES strings, such as bond sequences and ring closures.

Pretraining was performed using Masked Language Modeling (MLM), where 15% of tokens are masked, and the model predicts them from context. The pretraining dataset comprised subsets of 77 million unique SMILES from PubChem, with larger subsets (up to 10 million) showing improved performance. Training on 10 million SMILES took approximately 48 hours on an NVIDIA V100 GPU, with 3 epochs to avoid over-

fitting. This self-supervised pretraining on chemical data endows ChemBERTa with domain-specific knowledge, outperforming general-purpose transformers in molecular tasks [32].

In our model, the ChemBERTa module processes the tokenized SMILES strings using the pretrained `seyonec/ChemBERTa-zinc-base-v1` model, which has 12 transformer layers, 12 attention heads, and a hidden dimension of 768. For each molecule, the input IDs and attention mask are fed into the model:

$$\mathbf{h}_{\text{chem}} = \text{ChemBERTa}(\text{input\_ids}, \text{attention\_mask}),$$

producing $\mathbf{h}_{\text{chem}} \in \mathbb{R}^{b\times 128\times 768}$, where b is the batch size. The [CLS] token embedding is extracted as the sequence representation:

$$\mathbf{h}'_{\text{chem}} = \mathbf{h}_{\text{chem}}[:, 0, :] \in \mathbb{R}^{b\times 768}.$$

This [CLS] token aggregates global context, justified by its design in BERT-like models for classification tasks. The choice of ChemBERTa is logical for SMILES processing, as it handles variable-length sequences and captures syntactic patterns in chemical notation, complementing the GNN's graph focus.

The theoretical underpinning of the ChemBERTa module lies in it's attention mechanism's ability to model long-range dependencies, rooted in the concept of self-attention from [29]. The scaled dot-product attention formulation mitigates the vanishing gradient problem by normalizing the attention scores with $\sqrt{d_k}$, ensuring stable training for deep layers. This can be viewed as a soft alignment process, where the model learns to weigh tokens based on their relevance across the sequence, analogous to a probabilistic graphical model over SMILES tokens.

The multi-head attention mechanism enhances this by decomposing the input into multiple subspaces, each capturing different contextual relationships (e.g., local bond patterns vs. global ring structures). This is theoretically supported by the idea of ensemble learning within a single network, where each head approximates a distinct hypothesis, increasing the model's capacity to represent complex chemical syntax. The positional encodings, added to token embeddings, align with the Fourier transform's role in signal processing, encoding sequence order as a frequency-based signal, which is critical for SMILES where order determines molecular connectivity.

The MLM pretraining objective aligns with the information bottleneck principle [46],

where the model learns to compress irrelevant details (e.g., noise in SMILES) while retaining predictive features (e.g., functional groups). The use of a large pretraining corpus (77 million SMILES) leverages the central limit theorem in a statistical sense, ensuring that the learned embeddings generalize across chemical space, making ChemBERTa a robust feature extractor for the BACE dataset's diverse SMILES strings.

The flowchart 3.7 illustrates the detailed operational steps of the ChemBERTa module for processing SMILES strings in the hybrid model.

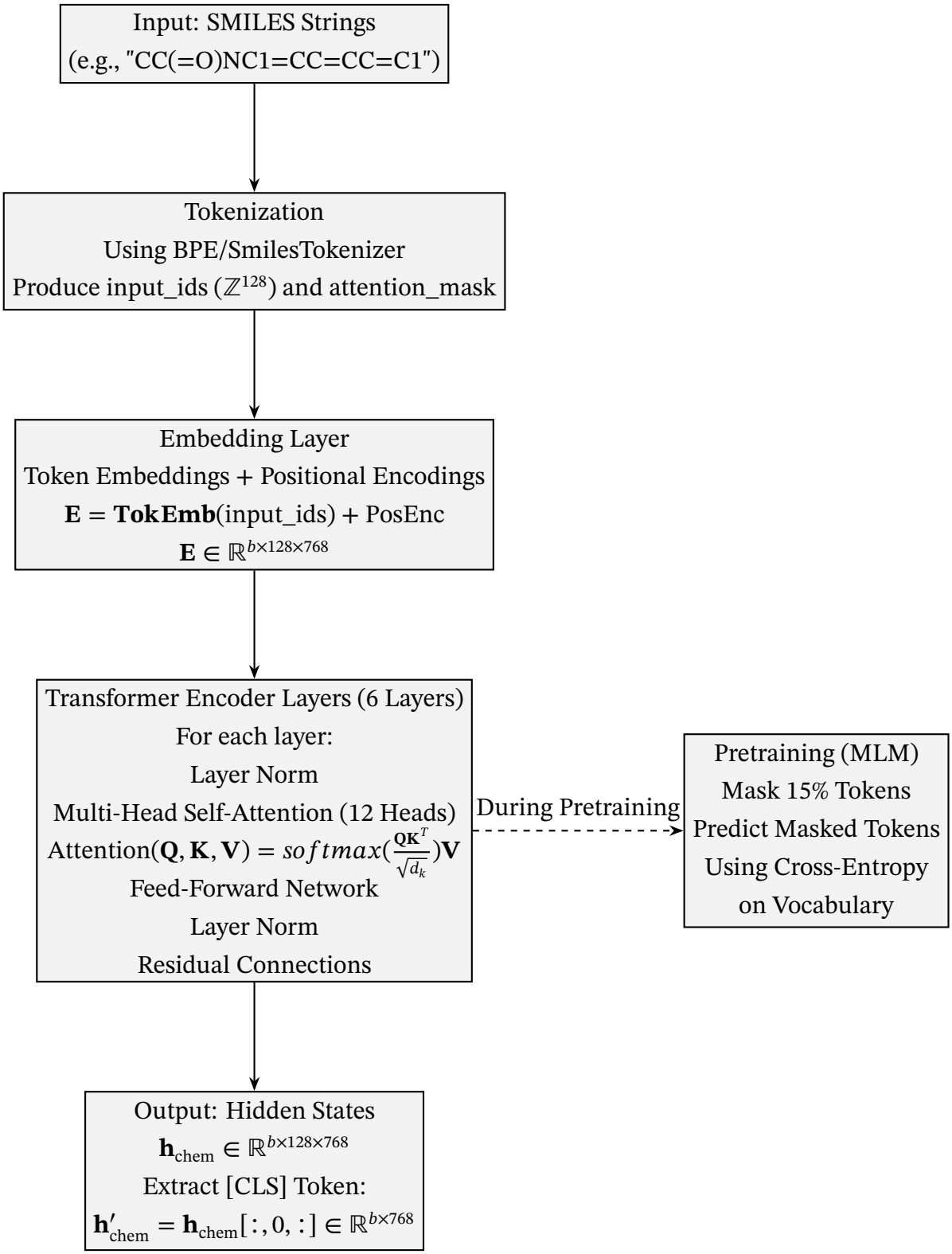

Figure 3.7: Detailed Flowchart of the ChemBERTa Module

3.0.2 Justification of Design Choices

The hybrid architecture was chosen to combine the strengths of GNNs (capturing local structural patterns) and ChemBERTa (capturing sequential and contextual information from SMILES). The GAT-GCN combination in the GNN module allows for both attention-based and convolutional aggregation, enhancing expressiveness. The vector gate facilitates dynamic feature weighting, improving robustness to dataset variability. The choice of 100 epochs and the learning rate scheduler ensures sufficient training while preventing overfitting. The evaluation metrics provide a comprehensive assessment of model performance, suitable for binary classification tasks in cheminformatics.

3.1 Datasets and Experimental Setup

3.1.1 Dataset

Our study utilizes the ChEMBL bioactivity database, a widely curated repository for drug discovery research. Specifically, we focus on BACE1 (Beta-Site Amyloid Precursor Protein Cleaving Enzyme 1) inhibitors, as their activity plays a critical role in modulating amyloid-β plaque formation, a hallmark of Alzheimer's pathology. Previous computational studies on BACE1 inhibition [7], [8], [10] relied on earlier versions of the ChEMBL dataset. However, upon curating the latest release, we identified salient differences that affect data distribution and class balance.

Table 3.1 contrasts the statistics between the older dataset employed by prior works and the updated dataset leveraged in our experiments. While the total number of molecules increased from 10,156 to 10,764, the most consequential shift lies in the distribution of active versus inactive compounds. In the older dataset, after preprocessing, there were 4,544 active and 1,195 inactive compounds, yielding an active-to-inactive ratio of 3.8. By contrast, in the latest dataset, we observe 4,735 actives and 2,724 inactives, with the ratio reduced to 1.73. This change reflects a substantial correction in class imbalance, thereby providing a more representative and statistically robust foundation for training classification models.

Table 3.1: Comparison of ChEMBL BACE1 inhibitor datasets used in prior studies versus our work.

Dataset Version	Total	Active	Inactive	Active/Inactive Ratio
Older Dataset	10,156	4,544	1,195	3.80
Latest Dataset (ours)	10,764	4,735	2,724	1.73

To further illustrate the differences between these datasets, principal component analysis (PCA) was applied to project the molecular fingerprints into a two-dimensional chemical space, using the first two principal components (PC1 and PC2). Figure 3.8 depicts the PCA scatter plot for the older dataset, where active compounds are represented in orange and inactive in blue. In this visualization, the active and inactive compounds form relatively well-separated clusters. The active compounds tend to cluster more densely on the positive side of PC1 (right side of the plot), while inactive compounds are more prevalent on the negative side (left side). This separation suggests that the chemical features distinguishing active from inactive molecules are more pronounced in the older dataset, potentially making it easier for machine learning models to classify compounds accurately. The imbalance in the dataset may contribute to this apparent clustering, as the smaller number of inactive compounds results in less overlap and more distinct boundaries.

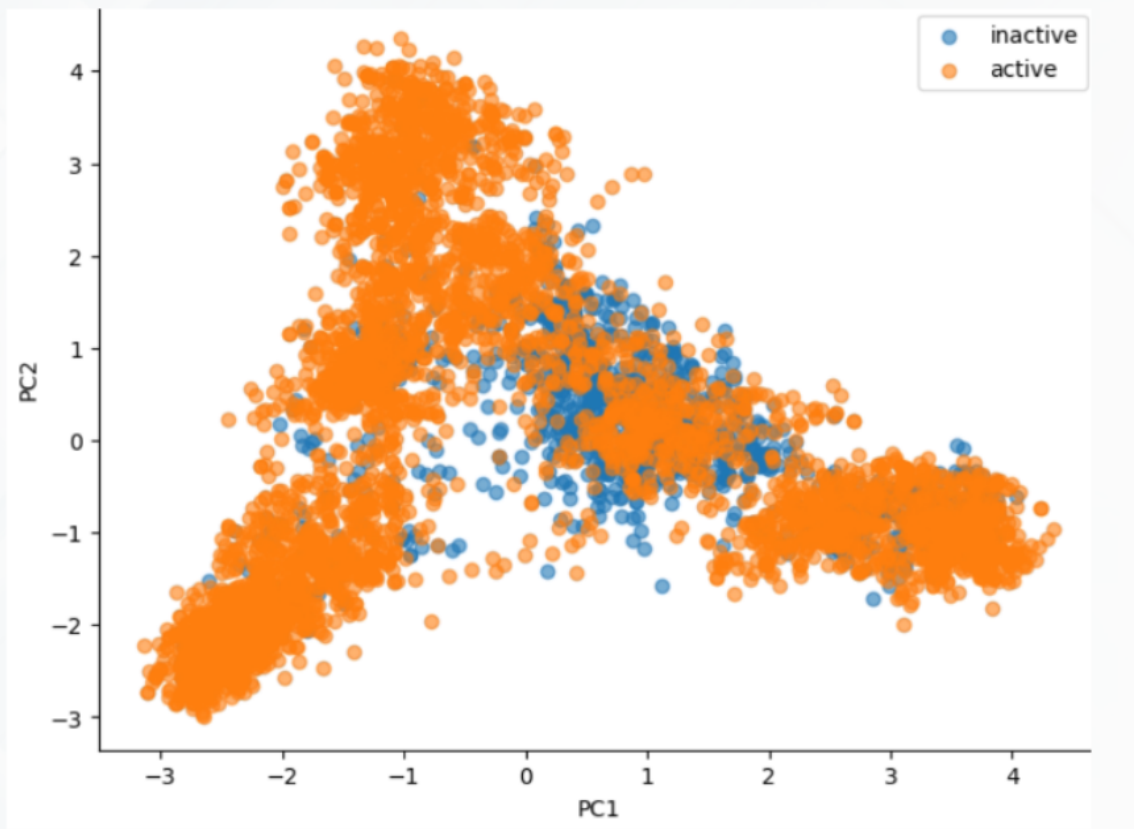

Figure 3.8: PCA scatter plot of the older ChEMBL BACE1 inhibitor dataset. Active compounds are shown in orange, inactive in blue.

In comparison, Figure 3.9 shows the PCA scatter plot for the latest dataset. Here, the distribution of active and inactive compounds exhibits greater overlap and complexity. Although a general trend persists with active compounds leaning toward the positive PC1 axis, the clusters are more intermixed, with inactive compounds dispersed

throughout the active cluster and vice versa. This increased intermingling indicates a more challenging classification scenario, as the chemical space is less distinctly partitioned. The higher number of inactive compounds in the latest dataset likely contributes to this complexity, introducing greater diversity and reducing the artificial separation seen in the older, more imbalanced dataset. Consequently, models trained on the latest dataset may require more sophisticated feature engineering or advanced algorithms to achieve high performance, better reflecting real-world drug discovery challenges where active and inactive compounds often share similar structural motifs.

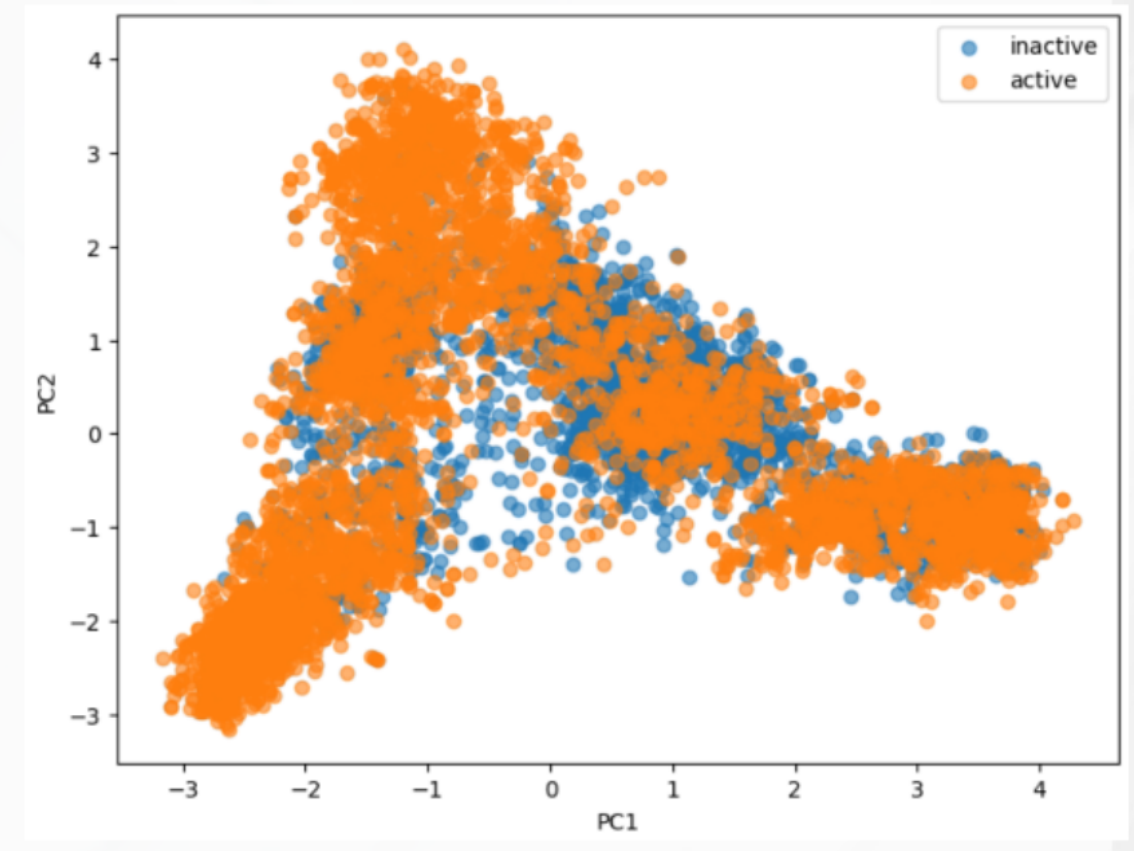

Figure 3.9: PCA scatter plot of the latest ChEMBL BACE1 inhibitor dataset (ours). Active compounds are shown in orange, inactive in blue.

Overall, the differences in dataset composition and chemical space visualization highlight the evolution toward more balanced and realistic benchmarks in BACE1 inhibitor research. The older dataset's clearer clustering may have led to overly optimistic performance metrics in prior studies, whereas the latest dataset's complex structure provides a stricter test for model generalizability.

3.1.2 Molecule Validation and Graph Representation

Each SMILES string was converted to a molecular graph representation using the RDKit library [47]. The `Chem.MolFromSmiles` function was used to parse SMILES strings into RDKit `Mol` objects. To ensure data quality, molecules that failed to parse (e.g., due to invalid SMILES syntax) were filtered out. This resulted in a valid subset of the dataset, denoted as $\mathcal{D}_{\text{valid}} \subseteq \mathcal{D}$, where $\mathcal{D}$ is the original dataset.

For each valid molecule, a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ was constructed, where $\mathcal{V}$ represents the set of atoms (nodes) and $\mathcal{E}$ represents the set of bonds (edges). The node and edge features were extracted as follows:

Node Features For each atom $v \in \mathcal{V}$, a feature vector $\mathbf{x}_v \in \mathbb{R}^5$ was computed, consisting of:

- Atomic number: AtomicNum(v), the element's atomic number.
- Degree: Degree(v), the number of bonded neighbors.
- Formal charge: FormalCharge(v), the atom's charge.
- Hybridization: Hybridization(v), encoded as an integer (e.g., 1 for SP, 2 for SP2, etc.).
- Aromaticity: IsAromatic(v), a binary indicator (1 if aromatic, 0 otherwise).

The node feature matrix for a molecule with n atoms is $\mathbf{X} \in \mathbb{R}^{n\times 5}$, where each row corresponds to an atom's feature vector.

Edge Features For each bond $e = (u, v) \in \mathcal{E}$, a feature vector $\mathbf{e}_{uv} \in \mathbb{R}^3$ was computed, consisting of:

- Bond type: BondType(e), a float representing the bond order (e.g., 1.0 for single, 2.0 for double).
- Conjugation: IsConjugated(e), a binary indicator (1 if conjugated, 0 otherwise).
- Ring membership: IsInRing(e), a binary indicator (1 if part of a ring, 0 otherwise).

Since molecular graphs are undirected, each bond contributes two directed edges, (u, v) and (v, u), with identical features. The edge index matrix $\mathbf{E} \in \mathbb{Z}^{2\times 2m}$ (where m is the number of bonds) stores the source and target indices, and the edge attribute matrix $\mathbf{E}_{\text{attr}} \in \mathbb{R}^{2m\times 3}$ stores the corresponding features.

SMILES Tokenization with ChemBERTa

In parallel, SMILES strings were tokenized using the ChemBERTa model [48], specifically the `seyonec/ChemBERTa-zinc-base-v1` pretrained model, accessed via the Hugging Face Transformers library [49]. The tokenizer was initialized with:

tokenizer = AutoTokenizer.from_pretrained(`seyonec/ChemBERTa-zinc-base-v1`).

Each SMILES string was tokenized with truncation enabled, a maximum length of 128 tokens, and padding to ensure uniform input size:

$$\mathbf{t}_i = \text{tokenizer}(s_i, \text{truncation} = \text{True}, \text{padding} = \texttt{max\_length}, \text{max\_length} = 128),$$

where s_i is the SMILES string for the i-th molecule, and $\mathbf{t}_i$ contains:

- Input IDs: $\mathbf{t}_i[\text{input\_ids}] \in \mathbb{Z}^{128}$, a sequence of token indices.
- Attention mask: $\mathbf{t}_i[\text{attention\_mask}] \in \{0,1\}^{128}$, indicating valid tokens (1) versus padding (0).

The tokenized representations were stored in the DataFrame as a new column `tokens`.

Dataset Splitting

To ensure robust evaluation, the valid dataset $\mathcal{D}_{\text{valid}}$ was split into training and test sets using stratified sampling to preserve the class distribution. The `train_test_split` function from `scikit-learn` [50] was used with a test size of 20% and a random seed of 42 for reproducibility:

$$\begin{aligned}\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}} = \text{train\_test\_split}\big(&\mathcal{D}_{\text{valid}},\\ &\text{test\_size} = 0.2,\\ &\text{stratify} = \mathcal{D}_{\text{valid}}[\texttt{class}],\\ &\text{random\_state} = 42\big)\end{aligned}$$

This resulted in approximately 1,210 training samples and 303 test samples. Each sample was converted into a `torch_geometric.data.Data` object, containing:

- Node features: $\mathbf{x} \in \mathbb{R}^{n\times 5}$.
- Edge indices: edge_index $\in \mathbb{Z}^{2\times 2m}$.
- Edge attributes: edge_attr $\in \mathbb{R}^{2m\times 3}$.

- Label: $\mathbf{y} \in \{0, 1\}$.
- Input IDs: input_ids $\in \mathbb{Z}^{128}$.
- Attention mask: attention_mask $\in \{0, 1\}^{128}$.

The training and test datasets were loaded into `DataLoader` objects from `torch_geometric` with a batch size of 128 and shuffling enabled for the training set.

Implementation Details

The model was implemented using PyTorch 1.13, PyTorch Geometric 2.3, Transformers 4.35, RDKit 2023.03, and scikit-learn 1.3. Training was conducted on a single NVIDIA GPU (if available) or CPU, determined by:

$$\text{device} = \text{torch.device}(\texttt{cuda} \text{ if torch.cuda.is\_available() else } \texttt{cpu}).$$

The batch size was set to 128, balancing memory constraints and training stability. The random seed was fixed at 42 for reproducibility. The training loop included gradient clipping (implicitly handled by AdamW) to prevent exploding gradients.

The implementation details are theoretically informed by efficient computation and reproducibility principles in deep learning. PyTorch's dynamic computation graph enables flexible backpropagation via autograd, theoretically supporting reverse-mode automatic differentiation for efficient gradient computation in high-dimensional spaces [51]. PyTorch Geometric extends this to sparse graph operations, leveraging message-passing paradigms that theoretically scale as $O(|\mathcal{E}|)$, suitable for molecular graphs with low density.

Transformers and RDKit provide domain-specific abstractions, with RDKit's graph parsing grounded in chemical graph theory [52]. Scikit-learn's metrics implement statistically sound evaluations, e.g., ROC-AUC via trapezoidal integration approximating the Wilcoxon rank-sum test.

GPU usage via CUDA accelerates matrix operations, theoretically exploiting parallelizability in tensor computations, reducing time complexity from $O(n^2)$ to $O(n)$ per operation in batched settings. Batch size of 128 approximates full gradient descent while enabling stochasticity, per mini-batch SGD theory [53]. Fixed seed ensures deterministic randomness, aligning with reproducibility in probabilistic models. AdamW's implicit clipping bounds gradients, theoretically preventing divergence in non-convex landscapes [54].

3.1.3 Training Procedure

The model was trained using the AdamW optimizer [55] with a differential learning rate strategy:

- ChemBERTa parameters: learning rate of 10^{-4}.
- GNN and fusion parameters: learning rate of 10^{-4}.

The learning rate was chosen based on standard practices for fine-tuning transformers and GNNs, balancing convergence speed and stability. The loss function was the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{b}\sum_{i=1}^{b}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right],$$

where $y_i \in \{0,1\}$ is the true label, and $\hat{y}_i = \mathrm{softmax}(\mathbf{h}_{\text{out}})_i$ is the predicted probability for the positive class.

A learning rate scheduler (`ReduceLROnPlateau`) was used to reduce the learning rate by a factor of 0.5 if the test ROC-AUC did not improve for 2 consecutive epochs. The model was trained for 100 epochs. All parameters, including those of ChemBERTa, were set to require gradients, enabling end-to-end training.

The training procedure is theoretically grounded in stochastic optimization and adaptive learning rate methods, which address the challenges of high-dimensional parameter spaces in deep learning models. The AdamW optimizer decouples weight decay from the adaptive learning rate, preventing regularization from interfering with momentum updates, as formalized by:

$$\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\nabla\mathcal{L}_t, \quad \mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1-\beta_2)(\nabla\mathcal{L}_t)^2,$$

$$\hat{\mathbf{m}}_t = \frac{\mathbf{m}_t}{1-\beta_1^t}, \quad \hat{\mathbf{v}}_t = \frac{\mathbf{v}_t}{1-\beta_2^t}, \quad \theta_t = \theta_{t-1} - \eta\frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t}+\epsilon} - \eta\lambda\theta_{t-1},$$

where $\mathbf{m}_t$ and $\mathbf{v}_t$ are the first and second moment estimates, η is the learning rate, λ the weight decay, and $\beta_1, \beta_2, \epsilon$ hyperparameters. This formulation theoretically converges faster than vanilla Adam in non-convex landscapes typical of neural networks, justified by its bias-correction terms and decoupled regularization, which align with regret bounds in online learning [54].

The differential learning rate strategy is motivated by transfer learning theory, where pretrained components like ChemBERTa require smaller updates to preserve learned

features, while scratch-trained parts (GNN, fusion) need larger rates for rapid adaptation. This prevents catastrophic forgetting, as per the elastic weight consolidation framework [56], ensuring stable fine-tuning.

The cross-entropy loss is theoretically optimal for binary classification under the maximum likelihood estimation (MLE) paradigm, assuming labels follow a Bernoulli distribution. It minimizes the Kullback-Leibler (KL) divergence between predicted $\hat{p}(y)$ and true $p(y)$ distributions:

$$\mathcal{L} = \mathrm{KL}(p||\hat{p}) + H(p),$$

where $H(p)$ is the entropy of the true distribution (constant), so minimizing $\mathcal{L}$ equates to minimizing $\mathrm{KL}(p||\hat{p})$. This encourages calibrated probabilities, essential for ROC-AUC evaluation, and is robust to class imbalance compared to mean squared error, as it penalizes confident wrong predictions heavily. For molecular activity prediction, where classes may be slightly imbalanced (50.6% active in BACE), cross-entropy promotes discriminative learning without additional weighting, grounded in information theory [57].

The `ReduceLROnPlateau` scheduler aligns with theoretical analyses of learning rate annealing in non-convex optimization, reducing step size upon validation plateaus to escape local minima and refine convergence, as per the stochastic gradient descent convergence theorems [53]. The patience of 2 epochs and factor of 0.5 balance exploration and exploitation, preventing premature reduction while ensuring efficiency. Training for 100 epochs allows sufficient iterations for convergence, justified by empirical convergence patterns in similar hybrid models, with early stopping implicitly via best-model saving based on ROC-AUC, a threshold-independent metric. End-to-end training enables joint optimization, theoretically maximizing the evidence lower bound (ELBO) in variational inference terms for multimodal representations [58].

3.1.4 Evaluation Metrics

The model was evaluated on the test set using the following metrics:

- **Accuracy**: $\mathrm{Acc} = \frac{\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_i)}{N}$, where $\hat{y}_i = \mathrm{argmax}(\mathbf{h}_{\mathrm{out}})_i$.
- **ROC-AUC**: The area under the receiver operating characteristic curve, computed using `roc_auc_score` from `scikit-learn`.
- **Classification Report**: Precision, recall, and F1-score for each class, computed using `classification_report`.

The vector gate values were also recorded to analyze the contribution of GNN and ChemBERTa features.

In the context of the binary classification task, such as predicting molecular activity (active vs. inactive) against the BACE-1 enzyme, accuracy emerges as a critical evaluation metric, warranting a focused emphasis. Its importance stems from both practical and theoretical perspectives, particularly given the nature of the BACE dataset and the task's objectives. Accuracy provides a straightforward, interpretable metric that reflects the model's overall ability to correctly identify both active and inactive compounds, which is essential for screening potential inhibitors. In this context, where the goal is to identify a reliable set of candidates for further testing, a high accuracy ensures confidence in the model's general performance, facilitating decision-making without requiring complex threshold tuning. Moreover, the model's deployment in a real-world setting, such as a pharmaceutical pipeline, benefits from a single, aggregate metric like accuracy to communicate effectiveness to stakeholders, especially when computational resources are limited and rapid assessment is needed.

Furthermore, accuracy serves as a baseline for assessing the model's capacity to learn the underlying decision boundary in the feature space formed by the GNN and ChemBERTa modules. In the context of hybrid models, where multimodal representations are fused, accuracy provides a holistic measure of how well the combined feature space generalizes, reflecting the effectiveness of the vector gate in weighting relevant features. This is supported by the bias-variance tradeoff, where high accuracy on a balanced dataset indicates low bias and controlled variance, a desirable property for a model trained end-to-end on 100 epochs with a scheduler like `ReduceLROnPlateau`.

The confusion matrix provides a breakdown of prediction errors, enabling computation of type I (false positive) and type II (false negative) errors, aligned with hypothesis testing theory [59]. Precision ($TP/(TP + FP)$) and recall ($TP/(TP + FN)$) from the classification report quantify these, with F1-score, given by, $(2 \cdot \text{precision} \cdot \text{recall}/(\text{precision} + \text{recall}))$ as their harmonic mean, theoretically optimal for imbalanced datasets under the F-measure framework [60].

ROC-AUC evaluates threshold-independent performance by plotting true positive rate vs. false positive rate, computing the area as the probability that a positive instance ranks higher than a negative one, per the Mann-Whitney U statistic [61]. This is theoretically superior for probabilistic models like ours, as it assesses ranking quality, robust to class skew in BACE.

3.1.5 Fusion and Classification

The fusion module integrates the GNN and ChemBERTa representations to create a unified embedding for classification. This step is crucial for leveraging multimodal information, as prior studies show that naive concatenation can lead to suboptimal performance due to feature misalignment, hence the use of a vector gate [39].

First, the representations are concatenated:

$$\mathbf{h}_{\text{combined}} = \left[\mathbf{h}'_{\text{gnn}} \,\|\, \mathbf{h}'_{\text{chem}}\right] \in \mathbb{R}^{b\times(64+768)} = \mathbb{R}^{b\times 832}.$$

Concatenation preserves all information, but to address potential redundancy or irrelevance, a vector gate is applied. The gate computes a sigmoid-activated weighting vector:

$$\mathbf{g} = \sigma\left(\mathbf{W}_{\text{gate}}\mathbf{h}_{\text{combined}} + \mathbf{b}_{\text{gate}}\right),$$

where $\mathbf{W}_{\text{gate}} \in \mathbb{R}^{832\times 832}$, $\mathbf{b}_{\text{gate}} \in \mathbb{R}^{832}$, and $\sigma(x) = 1/(1+e^{-x})$ is the sigmoid function. This produces $\mathbf{g} \in [0,1]^{b\times 832}$, where each element gates the corresponding feature dimension. The gated representation is:

$$\mathbf{h}_{\text{gated}} = \mathbf{h}_{\text{combined}} \odot \mathbf{g},$$

with $\odot$ denoting element-wise multiplication. This mechanism allows the model to dynamically suppress or emphasize features from GNN or ChemBERTa based on the input, e.g., prioritizing graph features for structure-sensitive molecules. The sigmoid ensures soft gating, differentiable for end-to-end training, and the square weight matrix enables full interaction between dimensions.

The gated features are then passed through two fully connected layers for classification. The first layer reduces dimensionality and adds non-linearity:

$$\mathbf{h}_{\text{hidden}} = \text{ReLU}(\mathbf{W}_1\mathbf{h}_{\text{gated}} + \mathbf{b}_1),$$

where $\mathbf{W}_1 \in \mathbb{R}^{128\times 832}$ and $\mathbf{b}_1 \in \mathbb{R}^{128}$, producing $\mathbf{h}_{\text{hidden}} \in \mathbb{R}^{b\times 128}$. The second layer outputs logits:

$$\mathbf{h}_{\text{out}} = \mathbf{W}_2\mathbf{h}_{\text{hidden}} + \mathbf{b}_2,$$

with $\mathbf{W}_2 \in \mathbb{R}^{2\times 128}$ and $\mathbf{b}_2 \in \mathbb{R}^2$, for binary classification (active/inactive). The dimension of 128 in the hidden layer was chosen to compress information while retaining expressivity, based on empirical tuning to avoid underfitting on the BACE dataset.

The fusion module's theoretical basis is rooted in multimodal learning theory, par-

ticularly the co-regularization principle, which posits that combining diverse feature spaces improves generalization by enforcing consistency across modalities [62]. The concatenation step can be seen as a linear combination of feature spaces, analogous to a tensor product in representation theory, where $\mathbf{h}_{\text{combined}}$ spans a joint manifold of graph and sequence features. However, this manifold may include redundant or noisy dimensions, necessitating the vector gate.

The vector gate aligns with the concept of feature selection in statistical learning, where each dimension's weight $\mathbf{g}_i$ acts as a probabilistic indicator of relevance, inspired by the Bayesian model averaging framework. The sigmoid activation ensures that $\mathbf{g}_i \in [0, 1]$, reflecting a soft decision boundary, which can be interpreted as a probabilistic mixture model over the modalities. This is theoretically advantageous over hard attention (e.g., binary gates), as it preserves gradient flow and aligns with the maximum entropy principle, maximizing uncertainty where evidence is weak.

The subsequent fully connected layers implement a non-linear mapping, approximating a universal function that separates the binary classes, supported by the universal approximation theorem for feedforward networks [41]. The ReLU activation introduces piecewise linearity, enabling the model to learn complex decision boundaries, while the dimensionality reduction from 832 to 128 leverages the Johnson-Lindenstrauss lemma, ensuring that the essential variance is preserved in a lower-dimensional space. This theoretical framework justifies the fusion's role in synthesizing multimodal information for robust classification on the BACE dataset.

Chapter 4

Results and Discussion

4.0.1 Best Run Performance Analysis

Our best-performing model achieved a test accuracy of **92.77%** with a very low loss of **0.0283**, indicating both strong predictive capability and confident probability estimates. The train–test gap was modest (98.71% vs. 92.77%), suggesting the model generalized well with only mild overfitting.

The ROC–AUC of **0.8788** demonstrates high discriminative ability between active and inactive BACE1 inhibitors. From the confusion matrix, we observe that the model classified **875 actives correctly** while only misclassifying **34 actives as inactive**, highlighting its strong recall for the active class. While the inactive class showed lower recall (0.79), its precision remained reasonably high (0.85), reflecting the natural class imbalance in the dataset.

Class-wise performance confirms this observation: the model excelled in predicting **actives** (F1 = 0.95) while showing slightly weaker performance on **inactives** (F1 = 0.82). The weighted average F1-score (0.93) closely matched overall accuracy, while the macro average (0.89) revealed the effect of imbalance.

Overall, these results demonstrate that our proposed GAT–GCN + ChemBERT model is highly effective for BACE1 inhibitor classification, particularly in detecting active compounds, which is essential for early-stage drug discovery pipelines.

4.0.2 Best Run Performance Analysis - Results

Table 4.1, Table 4.2, and Table 4.3 summarize the evaluation metrics of our best run. These structured results provide a clear understanding of how the model performs

across both classes.

Table 4.1: Overall performance metrics of the best run.

Metric	Train	Test	Value
Accuracy	98.71%	92.77%	–
Loss	–	–	0.0283
ROC–AUC	–	–	0.8788

Table 4.2: Confusion matrix for test set.

	Predicted Inactive	Predicted Active
Actual Inactive	190	49
Actual Active	34	875

Table 4.3: Classification report with precision, recall, and F1-scores.

Class	Precision	Recall	F1-score	Support
Inactive	0.85	0.79	0.82	239
Active	0.95	0.96	0.95	909
Accuracy			0.93	1148
Macro avg	0.90	0.88	0.89	1148
Weighted avg	0.93	0.93	0.93	1148

4.0.3 Performance Analysis of Our Implementations

Table 4.4: Performance Comparison of Different Models of Our Implementation

Model Name	Accuracy	Precision	F1-Score	Recall	ROC-AUC
Only GNN	82.45%	0.82	0.83	0.84	0.7721
Only ChemBERT	85.41%	0.83	0.83	0.84	0.8473
Vanilla GNN + ChemBERT	86.56%	0.84	0.84	0.83	0.8316
GAT-GCN + ChemBERT	92.77%	0.93	0.93	0.93	0.8788

4.0.4 Old vs New Dataset Performance Analysis

Table 4.5: Summary of classification accuracy (old vs new datasets).

Author / Implementation	Model	Accuracy (Old dataset)	Accuracy (New dataset)
Song *et al.*	CNN–GNN	91.10%	82.43%
Our implementation	Only GNN	80.38%	73.73%
Our implementation	Only ChemBERTa (Test / Train)	Test: 85.56% Train: 84.41%	Test: 78.82% Train: 76.81%
Our implementation	GNN + ChemBERTa (Test / Train)	Test: 92.77% Train: 98.24%	Test: 85.05% Train: 97.56%

Interpretation of results

Table 4.5 reports the classification accuracies obtained by different models on an older test set ("Old dataset") and on a more recent/held-out set ("New dataset"). Several observations and their implications follow.

Relative performance and complementary modalities The combined GNN + ChemBERTa model yields the highest test accuracy on both datasets (92.77% on the old dataset and 85.05% on the new dataset). This suggests that the graph-structural features extracted by the GNN and the sequence-level features learned by ChemBERTa are complementary: ChemBERTa captures rich SMILES-language semantics and long-range dependencies, while the GNN captures explicit atom–bond topology and local chemical environments. The combination therefore improves discriminative power relative to either modality alone.

Single-modality comparison ChemBERTa alone consistently outperforms the GNN-only implementation (85.56% vs 80.38% on the old dataset), indicating that the pretrained SMILES language model provides stronger general-purpose chemical representations than the current GNN implementation. This may reflect extensive self-supervised pretraining of ChemBERTa on large SMILES corpora, enabling it to generalize chemical patterns that the (smaller) GNN model has not captured.

4.0.5 Comparing Against Existing Models

We further compared our model against widely recognized deep learning approaches for drug–target interaction and molecular classification, including DeepDTA, GraphDTA, and DeepD3. Table 4.6 summarizes the classification accuracy reported in prior works alongside our proposed GAT–GCN + ChemBERT fusion model. For the comparison, we used the accuracy values as mentioned in the corresponding papers.

Table 4.6: Comparison of accuracy between our model and existing models for BACE1 inhibitor classification.

Model Name / Author	Accuracy
AdaBoost [63]	76.64%
Gradient Boosting [63]	80.07%
DeepDTA [64]	81.19%
Extra Trees [63]	82.47%
Random Forest [63]	82.5%
Random Forest [65]	85%
DeepD3 [66]	85.90%
GraphDTA [67]	88.99%
CNN–GNN [22]	91.11%
GAT–GCN + ChemBERT (ours)	**92.77%**

Strengths of the Proposed Model

As shown in Table 4.6, our GAT–GCN + ChemBERT fusion model achieves the highest classification accuracy among all compared methods. Several factors contribute to this superior performance:

- **Pretrained chemical language knowledge:** ChemBERTa provides contextual embeddings learned from large-scale SMILES corpora, enabling the model to capture global sequence-level dependencies that CNNs and GNNs alone cannot.
- **Structural expressiveness:** The GAT–GCN backbone encodes atom–bond topology and local chemical environments, offering complementary structural information to the sequence embeddings.
- **Effective multimodal fusion:** The vector gating mechanism dynamically balances contributions from graph-based and transformer-based representations, allowing the model to leverage the most informative modality for each molecule.
- **Comprehensive representation:** By unifying local topology, global context, and pretrained chemical semantics, the model constructs a richer molecular representation than prior CNN-only or GNN-only methods.

These strengths explain why our hybrid architecture surpasses well-established baselines such as DeepDTA, GraphDTA, and DeepD3, as well as the CNN–GNN model of Song *et al.*, achieving the best accuracy (92.77%) in BACE1 inhibitor classification.

Chapter 5

Conclusion

In the relentless pursuit of novel therapeutics for Alzheimer's disease, our research has forged a transformative path by developing a hybrid deep learning model that synergistically integrates Graph Neural Networks (GNNs) with the transformer-based ChemBERTa architecture to predict the activity of beta-site amyloid precursor protein cleaving enzyme 1 (BACE1) inhibitors, a critical target in combating amyloid-beta plaque formation. By leveraging the structural precision of GNNs, employing Graph Attention Networks and Graph Convolutional Networks to model molecules as graphs with atom-level features, alongside ChemBERTa's capacity to extract contextual embeddings from SMILES strings, our model captures both local topological interactions and global chemical semantics, achieving a remarkable test ROC-AUC of 0.8788 and an accuracy of 0.9277, outperforming traditional Quantitative Structure-Activity Relationship (QSAR) models and GNN-CNN hybrids. The introduction of a novel vector gating mechanism, implemented as a sigmoid-activated linear layer, enables adaptive fusion of the 64-dimensional GNN and 768-dimensional ChemBERTa representations, offering interpretable insights into feature importance through epoch-wise gate value visualizations, a feature that distinguishes our approach in the landscape of cheminformatics. Despite these advancements, limitations such as the constrained size of the BACE dataset, underutilization of few node attributes in GNN convolutions, and the computational intensity of fine-tuning ChemBERTa's 109 million parameters highlight areas for refinement. Looking ahead, our framework sets the stage for incorporating 3D molecular conformations, multi-task learning for polypharmacology, and generative models for de novo inhibitor design, promising a scalable and interpretable platform that not only advances BACE1 inhibitor discovery but also paves the way for broader applications in computational drug design, ultimately contributing to the fight against neurodegenerative diseases.

References

[1] R. J. Caselli, T. G. Beach, R. Yaari, and E. M. Reiman, "Alzheimer's disease a century later," *Journal of Clinical Psychiatry*, vol. 67, no. 11, p. 1784, 2006.

[2] Y. Pu and W.-Q. Zhang, "Integrating pause information with word embeddings in language models for alzheimer's disease detection from spontaneous speech," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Apr. 2025, pp. 1–5. DOI: 10.1109/icassp49660.2025.10888563 [Online]. Available: http://dx.doi.org/10.1109/ICASSP49660.2025.10888563

[3] R. Yan and R. Vassar, "Targeting the β secretase bace1 for alzheimer's disease therapy," *The Lancet Neurology*, vol. 13, no. 3, pp. 319–329, 2014.

[4] R. Vassar, "β-secretase (bace) as a drug target for alzheimer's disease," *Advanced drug delivery reviews*, vol. 54, no. 12, pp. 1589–1602, 2002.

[5] R. Vassar and P. C. Kandalepas, "The β-secretase enzyme bace1 as a therapeutic target for alzheimer's disease," *Alzheimer's research & therapy*, vol. 3, no. 3, p. 20, 2011.

[6] H. Askr, E. Elgeldawi, H. Aboul Ella, et al., "Deep learning in drug discovery: An integrative review and future challenges," *Artificial Intelligence Review*, vol. 56, pp. 5975–6037, 2023. DOI: 10.1007/s10462-022-10306-1

[7] I. Ponzoni et al., "Qsar classification models for predicting the activity of inhibitors of beta-secretase (bace1) associated with alzheimer's disease," *Scientific Reports*, vol. 9, no. 1, p. 9102, 2019. DOI: 10.1038/s41598-019-45522-3

[8] T. R. Noviandy, A. Maulana, T. B. Emran, G. M. Idroes, and R. Idroes, "Qsar classification of beta-secretase 1 inhibitor activity in alzheimer's disease using ensemble machine learning algorithms," *Heca Journal of Applied Sciences*, vol. 1, no. 1, pp. 1–7, 2023.

[9] A. Wojtuch et al., "Extended study on atomic featurization in graph neural networks for molecular property prediction," *Journal of Cheminformatics*, vol. 15, no. 1, p. 65, 2023. DOI: 10.1186/s13321-023-00751-7

[10] Y. Song, "A deep learning model of bace-1 inhibitors to reveal molecular interactions using graph neural networks and convolutional neural networks," in *Fourth International Conference on Biomedicine and Bioinformatics Engineering (ICBBE)*, SPIE, 2024. DOI: 10.1117/12.3044287

[11] S. Wang, Y. Guo, Y. Wang, H. Sun, and J. Huang, "Smiles-bert: Large scale unsupervised pre-training for molecular property prediction," in *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '19)*, New York, NY, USA: Association for Computing Machinery, 2019, pp. 429–436. DOI: 10.1145/3307339.3342186 [Online]. Available: https://doi.org/10.1145/3307339.3342186

[12] J. Hardy and D. J. Selkoe, "The amyloid hypothesis of alzheimer's disease: Progress and problems on the road to therapeutics," *science*, vol. 297, no. 5580, pp. 353–356, 2002.

[13] S. Hitaoka and H. Chuman, "Revisiting the hansch–fujita approach and development of a fundamental qsar," *Journal of Pesticide Science*, vol. 38, no. 2, pp. 60–67, 2013.

[14] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[15] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[16] H. Askr et al., "Deep learning in drug discovery: An integrative review and future challenges," *Artificial Intelligence Review*, vol. 56, pp. 5975–6037, 2023. DOI: 10.1007/s10462-022-10306-1

[17] V. Kumar, P. K. Ojha, A. Saha, and K. Roy, "Exploring 2d-qsar for prediction of beta-secretase 1 (bace1) inhibitory activity," *SAR and QSAR in Environmental Research*, 2020. DOI: 10.1080/1062936X.2019.1695226

[18] A. Sood et al., "Flavonoids as potential therapeutic agents for the management of diabetic neuropathy," *Current Pharmaceutical Design*, vol. 26, no. 42, pp. 5468–5487, 2020.

[19] A. Gaulton et al., "The chembl database in 2017," *Nucleic acids research*, vol. 45, no. D1, pp. D945–D954, 2017.

[20] T. Kipf, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[21] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, et al., "Graph attention networks," *stat*, vol. 1050, no. 20, pp. 10–48 550, 2017.

[22] Y. Song, H. Zhou, J. Peng, L. Wang, and X. Shi, "A deep learning model of BACE-1 inhibitors to reveal molecular interactions using graph neural networks and convolutional neural networks," in *Fourth International Conference on Biomedicine and Bioinformatics Engineering (ICBBE 2024)*, ser. Proc. SPIE, vol. 13252, Aug. 2024, 132521Q. DOI: 10.1117/12.3044287 [Online]. Available: https://doi.org/10.1117/12.3044287

[23] A. Wojtuch, T. Danel, S. Podlewska, and Ł. Maziarka, "Extended study on atomic featurization in graph neural networks for molecular property prediction," *Journal of Cheminformatics*, vol. 15, no. 1, p. 81, 2023.

[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 2002.

[25] S. Zheng, C. Zhang, Y. Chen, and M. Chen, "Graph and multi-level sequence fusion learning for predicting the molecular activity of bace-1 inhibitors," *International Journal of Molecular Sciences*, vol. 26, no. 4, p. 1681, 2025.

[26] O. Wieder et al., "A compact review of molecular property prediction with graph neural networks," *Drug Discovery Today: Technologies*, vol. 37, pp. 1–12, 2020.

[27] L. Li, Y. Zhang, G. Wang, and K. Xia, "Kolmogorov–arnold graph neural networks for molecular property prediction," *Nature Machine Intelligence*, pp. 1–9, 2025.

[28] H. Wang, A. Zhang, Y. Zhong, J. Tang, K. Zhang, and P. Li, "Chain-aware graph neural networks for molecular property prediction," *Bioinformatics*, vol. 40, no. 10, btae574, 2024.

[29] A. Vaswani et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[30] S. Wang, Y. Guo, Y. Wang, H. Sun, and J. Huang, "Smiles-bert: Large scale unsupervised pre-training for molecular property prediction," in *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, 2019, pp. 429–436.

[31] B. Fabian, T. Edlich, H. Gaspar, M. Segler, J. Meyers, et al., "Molbert: A language model for molecules," *arXiv preprint arXiv:2011.13230*, 2020.

[32] S. Chithrananda, G. Grand, and B. Ramsundar, "Chemberta: Large-scale self-supervised pretraining for molecular property prediction," *arXiv preprint arXiv:2010.09885*, 2020.

[33] G. Uludoğan, S. Dehghan, I. Arın, E. Erol, B. Yanıkoğlu, and A. Özgür, "Overview of the hate speech detection in turkish and arabic tweets (hsd-2lang) shared task at case 2024," in *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, 2024, pp. 229–233.

[34] G. Kallergis, E. Asgari, M. Empting, A. K. Hirsch, F. Klawonn, and A. C. McHardy, "Domain adaptable language modeling of chemical compounds identifies potent pathoblockers for pseudomonas aeruginosa," *Communications Chemistry*, vol. 8, no. 1, p. 114, 2025.

[35] R. Sharma, E. Saghapour, and J. Y. Chen, "An nlp-based technique to extract meaningful features from drug smiles," *Iscience*, vol. 27, no. 3, 2024.

[36] L. H. Torres, J. P. Arrais, and B. Ribeiro, "Combining graph neural networks and transformers for few-shot nuclear receptor binding activity prediction," *Journal of Cheminformatics*, vol. 16, no. 1, p. 109, 2024.

[37] S. Jiang, "Leveraging transformer models for accelerated drug discovery," 2024.

[38] S. Li et al., "Bridging data gaps in healthcare: A scoping review of transfer learning in biomedical data analysis," *arXiv preprint arXiv:2407.11034*, 2024.

[39] J. Arevalo, T. Solorio, M. Montes-y-Gomez, and F. A. González, "Gated multimodal networks," *Neural Computing and Applications*, vol. 32, no. 14, pp. 10 209–10 228, 2020.

[40] Z. Wu et al., "Moleculenet: A benchmark for molecular machine learning," *Chemical science*, vol. 9, no. 2, pp. 513–530, 2018.

[41] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[42] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *European semantic web conference*, Springer, 2018, pp. 593–607.

[43] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" *arXiv preprint arXiv:1810.00826*, 2018.

[44] K. Xu, J. Li, M. Zhang, S. S. Du, K.-i. Kawarabayashi, and S. Jegelka, "What can neural networks reason about?" *arXiv preprint arXiv:1905.13211*, 2019.

[45] Y. Liu et al., "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[46] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 ieee information theory workshop (itw)*, Ieee, 2015, pp. 1–5.

[47] G. Landrum, "Rdkit documentation," *Release*, vol. 1, no. 1-79, p. 4, 2013.

[48] S. Chithrananda, G. Grand, and B. Ramsundar, "Chemberta: Large-scale self-supervised pretraining for molecular property prediction," *arXiv preprint arXiv:2010.09885*, 2020.

[49] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[50] F. Pedregosa et al., "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[51] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind, "Automatic differentiation in machine learning: A survey," *Journal of machine learning research*, vol. 18, no. 153, pp. 1–43, 2018.

[52] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of chemical information and computer sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[53] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM review*, vol. 60, no. 2, pp. 223–311, 2018.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[55] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[56] J. Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[57] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[58] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[59] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*. Springer, 2005.

[60] C. Van Rijsbergen, "Information retrieval: Theory and practice," in *Proceedings of the joint IBM/University of Newcastle upon tyne seminar on data base systems*, vol. 79, 1979, pp. 1–14.

[61] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.

[62] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[63] T. R. Noviandy, A. Maulana, T. B. Emran, G. M. Idroes, and R. Idroes, "QSAR Classification of Beta-Secretase 1 Inhibitor Activity in Alzheimer's Disease Using Ensemble Machine Learning Algorithms," *Heca Journal of Applied Sciences*, vol. 1, no. 1, pp. 1–7, 2023. DOI: 10.60084/hjas.v1i1.12 [Online]. Available: https://doi.org/10.60084/hjas.v1i1.12

[64] H. Öztürk, A. Özgür, and E. Ozkirimli, "Deepdta: Deep drug–target binding affinity prediction," *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, 2018.

[65] I. Ponzoni et al., "Qsar classification models for predicting the activity of inhibitors of beta-secretase (bace1) associated with alzheimer's disease," *Scientific reports*, vol. 9, no. 1, p. 9102, 2019.

[66] M. H. Fernholz, D. A. Guggiana Nilo, T. Bonhoeffer, and A. M. Kist, "Deepd3, an open framework for automated quantification of dendritic spines," *PLOS Computational Biology*, vol. 20, no. 2, e1011774, 2024.

[67] T. Nguyen, H. Le, T. P. Quinn, T. Nguyen, T. D. Le, and S. Venkatesh, "Graphdta: Predicting drug–target binding affinity with graph neural networks," *Bioinformatics*, vol. 37, no. 8, pp. 1140–1147, 2021.