A deep learning model of BACE-1 inhibitors to reveal molecular interactions using graph neural networks and convolutional neural networks

Yuzhe Song[a], Han Zhou[a], Jiaqi Peng[a], Lu Wang[b], Xiumin Shi*[a]
[a]School of Information and Electronics, Beijing Institute of Technology, Beijing, 100081 China
[b]Dept. of Critical Care Medicine, Renmin Hospital of Wuhan University, Wuhan, 430060 China
* Corresponding author: sxm@bit.edu.cn

ABSTRACT

Significant emphasis has been placed on advancing therapeutic interventions and medicines to treat Alzheimer's disease, the leading cause of dementia. BACE1 inhibitors have shown considerable promise in reducing amyloid-β levels in the brain and potentially halting the progression of Alzheimer disease. However, human clinical trials are fraught with risk and exorbitant cost. In addressing these challenges, this investigation has developed a deep learning model for the prediction of interactions between BACE1 inhibitors and ligand. The model represents compounds as molecular graphs and SMILES strings, which are then processed through Graph Neural Network and Convolutional Neural Network channels, respectively. This approach allows comprehensive prediction of IC50 values and classification of compound activity with the BACE1 inhibitor. The model can be used as a convenient tool for the development of BACE1 inhibitors and also for virtual screening of molecules to identify potential inhibitors.

Keywords: β-Secretase 1 (BACE1) inhibitor, alzheimer disease, drug discover, deep learning

1. INTRODUCTION

Dementia is a significant health issue that affects not only the individuals who suffer from it but also their families, caregivers, and society at large [1]. Current estimates indicate that approximately 47 million individuals are affected by dementia. The prevalence of this condition is anticipated to escalate to 75 million by the year 2030, with a projected tripling of the affected population by 2050. Alzheimer's disease (AD), the predominant form of dementia, is characterized as a chronic and progressive neurodegenerative disorder. It is currently classified as incurable, with its pathological hallmarks such as the progressive degeneration of brain tissue. As a mild cognitive impairment disease, it predominantly affects the geriatric population, resulting in a progressive yet subtle decline in cognitive faculties, particularly impacting short-term memory and executive functions.

In the spectrum of hypotheses proposed to elucidate the pathogenesis of AD, the amyloid hypothesis is considered pivotal, which is strongly associated with the accumulation of toxic amyloid-β (Aβ) peptides. Aβ arises from the proteolytic cleavage of the amyloid precursor protein (APP), a process mediated by β and γ-secretases [2]. β-Secretase, identified as β-site amyloid precursor protein cleaving enzyme 1 (BACE1), catalyzes the initial and rate-limiting step of APP cleavage, thereby initiating the cascade that leads to Aβ production.

BACE1 is an aspartic acid protease composed of 501 amino acids, with a gene located on chromosome 1. In particular, the inhibition of BACE1 has been identified as a potential intervention to reduce the production of Aβ. By limiting the generation of these peptides, the progression of AD could be mitigated. BACE1 inhibitors represent a class of pharmaceutical agents designed to target and inhibit the activity of BACE1. Furthermore, BACE1 proteins and their mRNA transcripts have been identified in a variety of human cell types, albeit typically in modest abundance. Beyond its established association with the pathogenesis of AD, BACE1 has also been implicated in the development and progression of other conditions, such as type 2 diabetes, epilepsy, and schizophrenia. Emerging research suggests that BACE1 inhibitors may offer a promising therapeutic avenue for the treatment of type 2 diabetes [3].

Although many pharmaceutical companies are actively investigating BACE1 inhibitors as potential drugs for the treatment of Alzheimer's disease (AD), the challenges encountered in clinical trials cannot be ignored. BACE1 is primarily active in the central nervous system, which poses a challenge as it needs to overcome the blood-brain barrier. Even if the drug is able to penetrate the BBB, the inhibitor may trigger a complex response that can cause damage to the brain. Clinical trials involving BACE1 inhibitors are fraught with challenges, including safety concerns, cost, and time required. Computational

Fourth International Conference on Biomedicine and Bioinformatics Engineering (ICBBE 2024), edited by Pier Paolo Piccaluga, Ahmed El-Hashash, Xiangqian Guo, Proc. of SPIE Vol. 13252, 132521Q · © 2024 SPIE · 0277-786X · doi: 10.1117/12.3044287

modelling has the potential to aid in the screening of drug compounds, and although it cannot completely replace clinical trials, efficient predictive modelling techniques can reduce the reliance on human subjects in studies and increase the efficiency of the drug development process.

Several models have been proposed to classify BACE1 inhibitors. Research conducted by Subramaniam et al. in 2016 published a QSAR model of BACE1 inhibitors using the Ligand-Based approach [4]. A study by Ravi Singh et al. utilized various machine learning models, including Naive Bayes (NB), k-Nearest Neighbors (kNN), Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGB) to perform a classification assignment [5].Drug-target affinity (DTA) predicting can also help to evaluate the toxicity and adverse effects of drugs to enhance drug safty. In 2018, Ozturk et al. introduced the DeepDTA, which employs one-dimensional representations of drug and protein sequences to predict the strength values of protein-ligand binding affinity [6]. Then in 2020, Nguyen et al. introduced the GraphDTA model, which represents drugs as graphs and uses graph neural networks to predict drug–target affinity [7], they have consistently demonstrated superior predictive performance in estimating the DTA values. And various predictive models in DTA have experienced significant growth, including XG-DTA [8] and DGS-DTA [9].

Although the GraphDTA model is an improvement over DeepDTA and provides better results, each of these two models has unique advantages. The one-dimensional (1D) representation used in DeepDTA places more emphasis on the atomic composition of compounds, providing a macroscopic description of compounds. The graph representation utilized in GraphDTA, on the other hand, emphasizes the specific connections between atoms, refining the microscopic structure of compounds. And both macro and micro features can influence the effectiveness of drug prediction. To improve the accuracy of our predictions, we introduce a hybrid computational model that integrates both one-dimensional sequence features and graph-based molecular representations of the drug. Then, to identify active BACE1 inhibitors, we apply the developed model to datasets obtained from ChEMBL. Our objective is s to develop a model that will serve as a robust tool for virtual screening of BACE1 inhibitors, thereby accelerating the discovery process and reducing the need for extensive experimental testing.

2. MATERIAL AND METHOD

2.1 Dataset

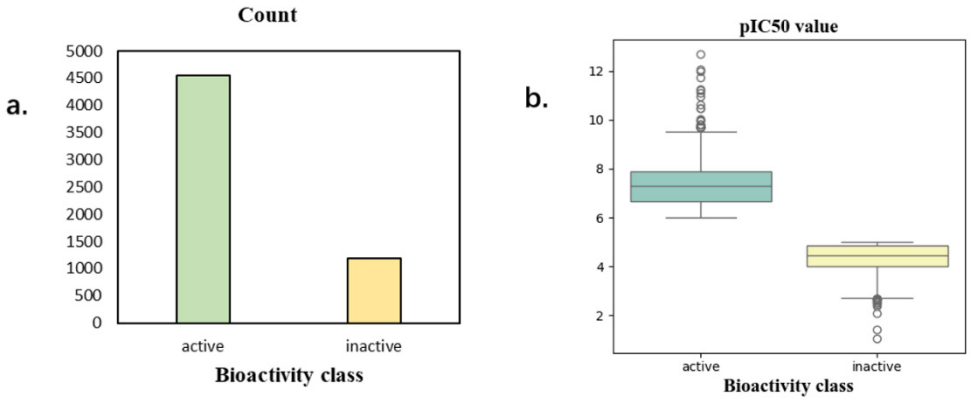

Figure 1. Data analysis of dataset (a) Sample number of each class (b) Statistical analysis of pIC50 values.

ChEMBL[10] is an open-access database that contains curated biological characteristics of over two billion compounds. We obtained bioactivity information solely for Human BACE1 (CHEMBL4822) encompassing 10156 compounds, from the database. And then we excluded the compound without a standard value (IC50 values can be derived from the standard values) or canonical smiles, as well as the compound categorized as intermediate. After all we collected data on 5739 compounds containing SMILES, IC50 values, and activity class sourced from ChEMBL to form the dataset used in this investigation. The process resulted in 4544 active and 1195 inactive compounds, as shown in Figure1a.

In order to ensure a more uniform distribution of the data, the IC50 value was transformed into the pIC50 value, calculated as the negative logarithm base 10 of the IC50 value, essentially -log10 (IC50). The data analysis presented in Figure 1b illustrates the relationship between the pIC50 value and bioactivity. In the dataset, compounds with pIC50 values greater

than 6 were classified as active, while those with pIC50 values less than 6 were categorized as inactive. Based on the relationship, we convert the classification problem into a regression task by using the SMILES of compounds as the input and the pIC50 value as the output.

2.2 Input representation

2.2.1 1D representation

We can employ a numerical encoding scheme where each character in the SMILES notation of BACE1 inhibitors is assigned a unique integer value. So, we can encode the SMILES with 64 labels. Evidently, the length of these SMILES representations varies among the compounds. To enhance the efficacy of the approach, we imposed a constraint on the maximum character length to 200, as indicated in Figure 2. The maximum length covers almost all compounds, allowing us to distinguish each compound and minimize unnecessary calculations. This constraint involves truncating sequences that exceed this length and padding shorter sequences with zeros to achieve uniformity in length.

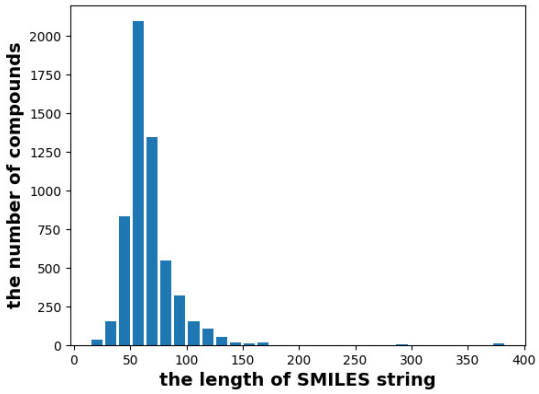

Figure 2. Distribution of the lengths of the SMILES strings.

2.2.2 Graph representation

Each drug compound is a graph of the interactions between atoms. The SMILES representation can be transformed into its respective molecular graph utilizing the RDKit software. Within this graph, each atom node is meticulously characterized by a set of five key attributes: the atomic symbol, the degree, the count of adjacent hydrogen atoms, the implicit valence value, and a binary indicator of whether the atom is part of an aromatic ring structure. By integrating these elemental data points, we construct a multi-dimensional feature matrix that represents the compound in a comprehensive and structured manner.

2.3 Proposed model

In order to process the one-dimensional representation of compounds, our method employed a sequence of two consecutive 1D-convolutional layers, followed by a global max pooling layer, and succeeded with two fully connected (FC) layers. Besides, we employ the Rectified Linear Unit (ReLU) as the activation function in every convolutional layer, and integrate a dropout layer with a 0.4 rate following the Fully Connected layers, to mitigate the risk of overfitting.

We chose the GAT-GCN model to process the graph representation of compounds within the graph channel. The model integrates the Graph Convolutional Network (GCN) and the Graph Attention Network (GAT) within a deep learning framework. The network commences with a GAT layer, which applies a linear transformation to each input node using a weight matrix Wh. The nodes denoted as j represent the immediate neighbors of each given input node i. Consequently, the attention coefficients between the node i and nodes j can be computed as

$$e_{ij} = a\big([Wh_i][Wh_j]\big), j \in \mathcal{N}_i \tag{1}$$

Then, we utilize the soft-max function to make the values normalized and aggregate the outputs to get the attention features of nodes as

$$h_i' = \sigma\big(\sum_{j\in\mathcal{N}_i} a_{ij}^k W^k h_j\big) \tag{2}$$

The σ() is characterized as a non-linear activation function., and the a_{ij} are the normalized outcomes. Attention mechanisms enable the model to dynamically allocate varying weights to distinct connections based on the specific task and the underlying graph structure.

The network then proceeds to a subsequent GCN layer where it convolves the feature matrix of the connections. GCN represents an innovative extension of CNN, which is capable of extracting graph features and facilitating model training. The procession can be descripted by matrix as

$$Z = \widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}X \tag{3}$$

$$D_{ii} = \sum_j A_{ij} \tag{4}$$

$\widetilde{A}$ is the adjacent matrix of A, and the $\widetilde{D}^{-\frac{1}{2}}$ can be viewed as the scale factor of $\widetilde{A}$. Following the implementation of two GNN layers, we employed global max pooling and global mean pooling operations to concatenate the network accordingly. The outputs from these two distinct pooling layers were combined through concatenation to produce the final result of the graph channel.

The proposed model that combines graph neural network and traditional convolutional neural network architectures is illustrated in Figure 3. And the hyperparameters employed in our experimental setup are outlined in Table 1.

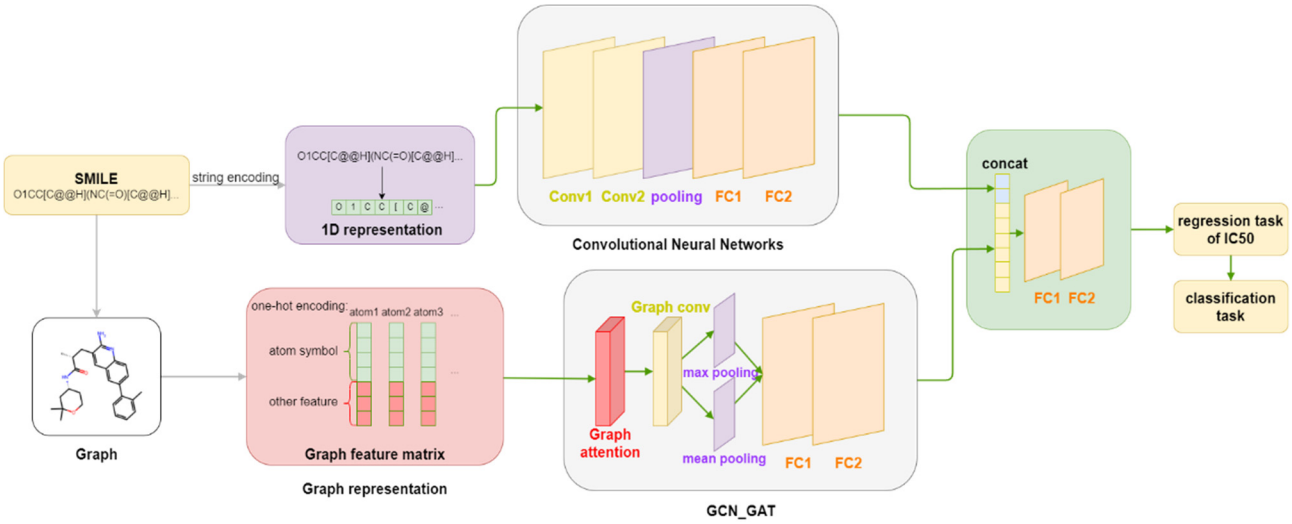

Figure 3. The network architecture of the proposed model.

Table 1. Hyper-parameters for the neural network.

Hyper-parameters	Setting value
Learning rate	0.0001
Batch size	512
optimizer	Adam
Dropout rate	0.4

3. RESULTS AND DISCUSSION

The loss value of the regression task pertaining to the IC50 value of BACE1 inhibitors can be visualized in Figure 4.a, illustrating the model's regression performance. The accuracy (ACC) and area under the curve (AUC) metrics can be computed based on the classification task, with the results presented in Figure 4.b and Figure 4.c, respectively. Additionally, we computed the Root Mean Square Error (RMSE) and Concordance Index (CI) values for the regression analysis. The best results of our model are presented in Table 2.

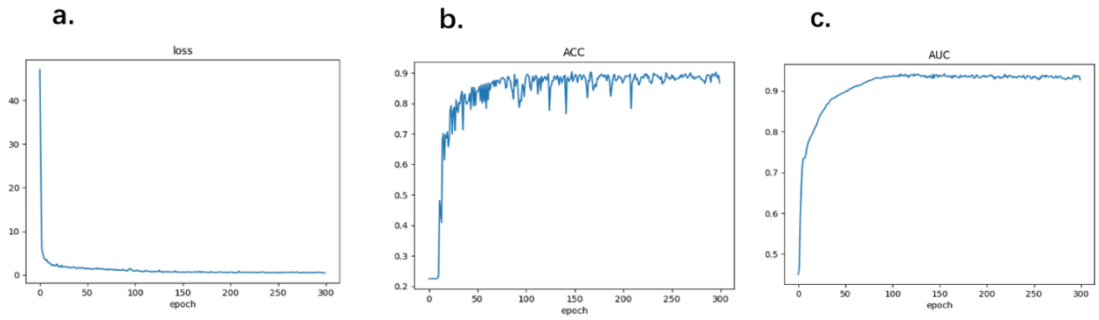

Figure 4. The learning curve of the model, which consists of three components: a) loss, b) accuracy (ACC), and c) area under the curve (AUC).

Table 2. The best outcomes of evaluation parameters.

Evaluation parameters	Outcome values
Best RMSE	0.71155
Best CI	0.80226
Best ACC	0.91115
Best AUC	0.93821

Table 3 compares the performance of our model with the existing baseline model on the same dataset. Our model is named CNN-GNN and the results obtained using regression analysis will be used for subsequent classification tasks. We compared the root mean square error values with other regression models and evaluated the accuracy values compared to other classification models. Baseline models include FP-GNN[11], DeepDTA, GraphDTA and Deep-d3[12] in comparison experiment. Compared to the baseline model, the CNN-GNN shows superior performance on every evaluation metric. The results demonstrate that the approach of combining one-dimensional sequence information with two-dimensional structural information can capture the characteristics of drug molecules in a more comprehensive way, thus improving the accuracy and reliability of prediction.

Table 3. Comparison results of the proposed model and baselines on the ChEMBL dataset.

Method	Task	ACC	RMSE
FP-GNN	Regression	-	0.76042
DeepDTA	Regression and classification	0.81192	0.81398
GraphDTA	Regression and classification	0.88998	0.72093
Deep-d3	Classification (CNN and NLP)	0.86520	-
	Classification (CNN)	0.86176	-
	Classification (NLP)	0.85901	-
	Classification	0.85901	-
CNN-GNN	Regression and classification	**0.91115**	**0.71155**

4. CONCLUSION

We have developed a predictive model based on deep learning to Identify potential inhibitors of Human β-secretase 1 for their potential therapeutic use in Alzheimer's disease. Our approach incorporates two distinct channels to independently train the one-dimensional (1D) and graph representations of compound SMILES. The optimal outcomes achieved by our model demonstrate a regression Root Mean Squared Error of 0.7115 and an Index of Concordance value of 0.80226. Furthermore, we have attained superior results in terms of predictive accuracy at 0.91115 and predictive area under the curve at 0.93821. The results indicate that our approach shows potential as a valuable tool for forecasting BACE-1 inhibitors.

ACKNOWLEDGMENTS

This research was partially supported by BIT Training Program for Innovation and Entrepreneurship, National Undergraduate Training Program for Innovation and Entrepreneurship: 202310007087.

DATA AND CODE AVAILABLITY

The BACE1 inhibitors dataset are available at https://www.ebi.ac.uk/chembl/. The codes of CNN-GNN are publicly available on https://github.com/ShiLab-GitHub/CNN-GNN.

REFERENCES

[1] Ulep, M.G., Saraon, S.K., McLea, S. "Alzheimer disease." The Journal for Nurse Practitioners 14(3), 129-135 (2018).
[2] Ponzoni, I., Sebastián-Pérez, V., et al. "QSAR Classification models for predicting the activity of inhibitors of beta-secretase (BACE1) associated with Alzheimer's disease. " Sci Rep 9, 9102 (2019).
[3] Dekeryte, R., Franklin, Z., Hull, C., et al. "The BACE1 inhibitor LY2886721 improves diabetic phenotypes of BACE1 knock-in mice." Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease 1867(7), 166149, (2021).
[4] Subramanian, G., Ramsundar, B., Pande, V., et al. "Computational modeling of β-secretase 1 (BACE-1) inhibitors using ligand based approaches." Journal of chemical information and modeling 56(10), 1936-1949 (2016).
[5] Singh, R., Ganeshpurkar, A., Ghosh, P., et al. "Classification of beta-site amyloid precursor protein cleaving enzyme 1 inhibitors by using machine learning methods." Chemical Biology & Drug Design 98(6), 1079-1097 (2021).
[6] Öztürk, H., Özgür, A., Ozkirimli, E.. "DeepDTA: deep drug–target binding affinity prediction." Bioinformatics 34(17), i821-i829 (2018).
[7] Nguyen, T., Le, H., Quinn, T.P., et al. "GraphDTA: predicting drug–target binding affinity with graph neural networks." Bioinformatics 37(8), 1140-1147 (2021).
[8] Zhou, H., Shi, X., Wang, Y., et al. "XG-DTA: Drug-Target Affinity Prediction Based on Drug Molecular Graph and Protein Sequence combined with XLNet," in *2023 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, Beijing, China: IEEE, 189–196(2023).
[9] Guo, Y., Shi, X., Zhou, H. "A Deep Learning Drug-Target Binding Affinity Prediction Based on Compound Microstructure and Its Application in COVID-19 Drug Screening." Journal of Beijing Institute of Technology 32(4), 396-405 (2023).
[10] Gaulton, A., Bellis, L.J., Bento, A.P., et al. "ChEMBL: a large-scale bioactivity database for drug discovery." Nucleic acids research, D1100-D1107 (2012).
[11] Cai, H., Zhang, H., Zhao, D., et al. "FP-GNN: a versatile deep learning architecture for enhanced molecular property prediction." Briefings in bioinformatics 23(6), bbac408 (2022).
[12] Qiang, T., Fulei, N., Qi, Z., Wei, C., "A merged molecular representation deep learning method for blood–brain barrier permeability prediction." Briefings in Bioinformatics 23(5), bbac357 (2022).