Enhancing Drug Discovery by Fusing Graph and Sequence Encoder Representations

*Abhishek K P *Madhurya R Nair *Sonika Rajesh Pillai *Vyshakh G Nair **Dr. Anuraj Mohan
*UG Scholar, **Associate Professor, Department of Computer Science and Engineering,
NSS College of Engineering, Palakkad

Problem Statement

- Drug discovery through wet-lab experimentation is a time-consuming and expensive process.
- The selection of potential drug candidates requires extensive biological testing in the lab, leading to high attrition rate.
- Deep learning offers a potential solution to these challenges by leveraging computational methods to expedite the process and reduce resource wastage.

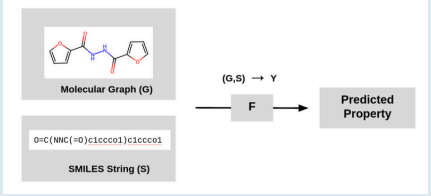
Problem Definition

- In this approach, molecules are represented either as **graphs** (where atoms are nodes and chemical bonds are edges) or as **sequences** (such as SMILES strings representing the chemical structure).
- The goal is to predict certain properties or characteristics of these molecules, such as their effectiveness as drugs or their toxicity.

Formally, the problem is defined as follows:

- **Given a molecular representation X, which can be either a molecular graph, G or a molecular sequence S, the task is to predict a corresponding property, Y.**
- **The function *f* maps the input molecular representation to the output property:**

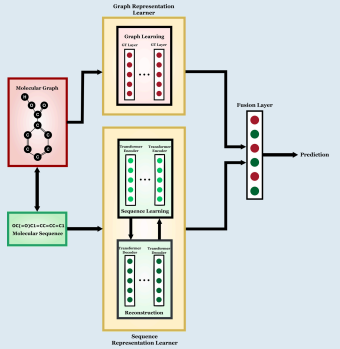
$$f: X \rightarrow Y$$



Objectives

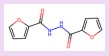
- To employ various preprocessing techniques to ensure the conversion of SMILES (Simplified Molecular Input Line Entry System) representation that is a suitable input for graph and sequence model.
- To explore multiple graph and sequence models that yields the best learning.
- To develop fusion-model - TraGT, TraGT-R (with reconstruction) that is capable of learning both graph and sequence representations from the SMILES of molecules from the dataset.
- To conduct comprehensive training and evaluation of the developed fusion model and employ appropriate evaluation metrics to assess the performance of the model.

Methodology



Input Representations

Input representations involve molecular graph input, G and SMILES sequence input S. The graph data and sequence data is fed into the **Graph Representation Learner** and **Sequence Representation Learner.**



O=C(NNC(=O)c1ccccl)c1ccccl

Graph Transformer

- It transforms graphs into numerical embeddings extracted from nodes & edges.
- They focus on a node's neighbors (like attention in sequences) considering edge information.
- This creates a final embedding for each node, capturing its features and connections within the graph.

$$q_{c,j}^{(l)} = W_{c,q}^{(l)} h_i^{(l)} + b_{c,q}^{(l)} \qquad k_{c,j}^{(l)} = W_{c,k}^{(l)} h_j^{(l)} + b_{c,k}^{(l)}$$

$$v_{c,j}^{(l)} = W_{c,v}^{(l)} h_j^{(l)} + b_{c,v}^{(l)} \qquad e_{c,ij} = W_{c,e} e_{ij} + b_{c,e}$$

$$a_{c,ij}^{(l)} = \frac{\langle q_{c,i}^{(l)}, k_{c,j}^{(l)} + e_{c,ij} \rangle}{\sum_{u \in N(i)} \langle q_{c,i}^{(l)}, k_{c,u}^{(l)} + e_{c,iu} \rangle}$$

$$\hat{h}_i^{l+1} = \|_{c=1}^{C} \left[\sum_{j \in N(i)} a_{c,ij}^{(l)} (V_{c,j}^{(l)} + e_{c,ij}) \right]$$

Transformer Encoder

- It transforms the SMILES strings to numerical representations.
- Each encoder use self-attention to understand how each element relates to others in the sequence, and feed-forward networks for further processing. This captures both individual meanings and contextual relationships.

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

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$
$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Transformer Decoder

- Transformer Decoder is also added to the Sequence Representation Learner to recreate the original sequence data.
- This helps ensure that the model isn't losing too much information during processing and serves as a form of regularization. The reconstruction process contributes to the loss function.

Fusion Layer

The graph and sequence embeddings are then passed through a fusion layer that uses attention to focus on relevant parts of the sequence data based on the context provided by the graph data. It then combines this focused sequence information with the graph data to create a more comprehensive representation for the final prediction.

$$x = [g, s]$$
$$attention(x) = W_{att} \cdot x + b_{att}$$
$$\alpha = \sigma(attention(x))$$
$$s' = \alpha \cdot s$$
$$x' = [g, s']$$
$$fusion(x) = W_{fusion} \cdot x' + b_{fusion}$$
$$y = fusion(x)$$

Experimental Setup

Dataset Description

DATASET	DESCRIPTION	TASK
LogP	Solubility of molecules	Binary labels for LogP >= 1.88 and LogP < 1.88
BBBP	Blood-brain barrier permeability	Binary labels for penetration/non-penetration
BACE	Quantitative (IC50) and qualitative (binary label) binding results for a set of inhibitors of human β-secretase 1(BACE-1)	Binary labels for inhibitor
FDA	Approved compounds by Food and Drug Administration	Approval analysis

Tools Specifications

PARAM UTKARSH Supercomputing Infrastructure at C-DAC

Hardware	Specifications
CPU	2x Intel Xeon G-6248
Clock Speed	2.5 GHz
Memory (per node)	192 GB DDR4
SSD Storage	480 GB (per node)
GPU (per node)	2x NVIDIA V100
GPU Memory (per node)	32 GB HBM2
Software	**Version**
PyTorch	2.2.1
RDKit	2023.9.6
PyTorch Geometric	2.5.3

Evaluation Metrics

The proposed model was evaluated to validate the chosen architecture and hyper parameters over the four datasets based on **Accuracy, ROC-AUC** (Receiver Operating Characteristics - Area Under the Curve) **, F1-score, Precision and Recall.**

Experimental Results

TraGT without Reconstruction

Datasets	Metrics				
	AUC	Accuracy	F1 Score	Recall	Precision
LogP	97.98	91.30	90.45	87.78	100
BBBP	97.22	92.74	95.30	100	95.74
FDA	97.54	95.05	95.27	96.36	96.45
BACE	74.30	74.35	62.61	60.66	76.19

TraGT with Reconstruction

Datasets	Metrics				
	AUC	Accuracy	F1 Score	Recall	Precision
LogP	98.28	92.30	91.75	87.17	99.5
BBBP	96.39	91.71	94.74	100	91.72
FDA	96.87	93.81	94.05	95.86	96.15
BACE	75.55	72.37	64.61	65.57	88.89

- It is quite evident that all the four datasets have obtained good ROC-AUC values.
- A **high ROC-AUC** value indicates that the model has a higher ability to distinguish between the classes

Comparative Analysis

The table shown below compares the performance of the proposed model with the Graph Transformer model, Transformer Encoder model, and GraSeq model.

Models	Datasets			
	LogP	BBBP	FDA	BACE
GT	97.81	96.96	97.31	70.73
TE	94.86	95.90	95.29	74.74
GraSeq-F	98.57	94.26	99.10	83.82
GraSeq-RF	98.83	91.36	98.38	81.21
TraGT	97.98	**97.22**	97.54	74.30
TraGT-R	98.28	**96.39**	96.87	75.55

The above plot compares the proposed model with the GraSeq model over the BBBP dataset. The proposed model outperformed the GraSeq model.

Conclusion and Future Work

- The proposed TraGT fusion model outperforms individual models and existing GraSeq fusion model.
- Future work may involve integration of 3D structure learning into graph embeddings.
- Introducing geometric graph contrastive learning could enhance molecular representation.
- While the fusion model excels, its interpretability is a challenge. Incorporating explainability could offer insight into its decision-making process.

References

[1] Zhou, Jie, et al. "Graph neural networks: A review of methods and applications." AI open 1 (2020): 57-81.

[2] Dwivedi, Vijay Prakash, and Xavier Bresson. "A generalization of transformer networks to graphs." arXiv preprint arXiv:2012.09699 (2020).

[3] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017)

[4] Zhang, Chuxu, et al. "GraSeq: Graph and Sequence Fusion Learning for Molecular Property Prediction." CIKM'20: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. Association for Computing Machinery, 2020.

Paper Presentation

- Presented a paper titled Leveraging Deep Learning and Molecular Representation for Sustainable Drug Discovery in APJ AKTU Techfest and KETCON 2024, at Ahalia School of Engineering and Technology, Palakkad on 16th February, 2024.
- Registered a paper titled Leveraging Deep Learning and Molecular Representation for Drug Discovery in International Conference on "Integrated Circuits, Communication, and Computing Systems (ICIC3S-2024)", organised by School of Electronics, Indian Institute of Information Technology Una, Himachal Pradesh.