ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Convolutional Transformer via Graph Embeddings for Few-shot Toxicity and Side Effect Prediction

Luis H. M. Torres[1], Bernardete Ribeiro[1] and Joel P. Arrais[1] *

1 - Univ Coimbra, Centre for Informatics and Systems of the University of Coimbra Department of Informatics Engineering, Coimbra, 3030-290, Portugal

Abstract. The prediction of chemical toxicity and adverse side effects is a crucial task in drug discovery. Graph neural networks (GNNs) have accelerated the discovery of compounds with improved molecular profiles for effective drug development. Recently, Transformer networks have also managed to capture the long-range dependence in molecules to preserve the global aspects of molecular embeddings for molecular property prediction. In this paper, we propose a few-shot GNN-Transformer, FS-GNNCvTR to face the challenge of low-data toxicity and side effect prediction. Specifically, we introduce a convolutional Transformer to model the local spatial context of molecular graph embeddings while preserving the global information of deep representations. Furthermore, a two-module meta-learning framework is proposed to iteratively update model parameters across few-shot tasks with limited available data. Experiments on small-sized biological datasets for toxicity and side effect prediction, Tox21 and SIDER, demonstrate a superior performance of FS-GNNCvTR compared to standard graph-based methods. The code and data underlying this article are available in the repository, `https://github.com/larngroup/FS-GNNCvTR`.

1 Introduction

Toxicity and side effect prediction are essential tasks in drug discovery. Drugs that were previously approved can often be removed from the market due to the occurrence of toxic side effects and off-target interactions. To minimize costs and mitigate risks, it is crucial to select compounds with desirable molecular properties that can reduce chemical toxicity and decrease the risk of adverse drug reactions. Computational methods identify potential issues before clinical trials, saving time, resources and providing more efficient and safe treatments for patients. In particular, deep learning (DL) methods learn generalizable and transferable representations to model non-linear systems of toxicity and side effect prediction, which can improve the accuracy and efficiency of drug discovery efforts [1, 2]. However, due to the limited amount of labeled information available in a large chemical space, DL models struggle to generalize to new toxicity and side effect properties. Thus, finding ways to effectively learn from a small number of labeled molecules remains a key challenge in drug discovery [3]. In this paper, we propose a few-shot GNN-based convolutional Transformer,

*This work is funded by the FCT - Foundation for Science and Technology, I.P./MCTES through national funds (PIDDAC), within the scope of CISUC R&D Unit.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

FS-GNNCvTR to predict toxicity and drug side effects with a small amount of labeled molecules. FS-GNNCvTR leverages the hierarchical structure of convolutional architectures to learn both local and global connections in graph embeddings at varying levels of complexity. Small data collections, Tox21 and SIDER [4], with high-level molecular property measurements are considered for toxicity and side effect prediction. To address the problem of low-data, we propose a two-module meta-learning framework to iteratively update model parameters across few-shot tasks with limited available data. Few-shot experiments demonstrate the superior performance of FS-GNNCvTR over standard graph-based methods.

2 Methods

2.1 Graph Neural Network Module

Molecules can be described by molecular graphs $G = (V, E)$, where V is the set of nodes v (atoms) and E is the set of edges e (chemical bonds). An edge is defined by $e = (v, u)$, where v and u are nodes connected in a neighborhood $N(v)$. In this work, a graph isomorphism network (GIN) with $L_{GIN} = 5$ layers is used to iteratively aggregate node embeddings h_v^l at message-passing layers l and compute graph embeddings h_G. The GIN performs AGGREGATE and COMBINE steps as a sum of node and edge features. An UPDATE step applies a multi-layer perceptron MLP followed by non-linear activation $\sigma = ReLU$.

$$h_v^l = ReLU(MLP^l(\sum_{u \in N(v) \cup v} h_u^{l-1} + \sum_{e=(v,u): u \in N(v) \cup v} h_e^{l-1})). \quad (1)$$

A READOUT operation pools node embeddings to obtain a graph embedding h_G by averaging node embeddings h_v (mean-pooling) at the final layer, $h_G = mean(\{h_v^L : v \in V\})$. Input node and edge features (h_v^0, h_e^0) are described by atom and bond attributes including atom number (AN), atom chirality (AC) with $h_v^0 = \{v_{AN}, v_{AC}\}$, and bond type (BT), bond direction (BD) with $h_e^0 = \{e_{BT}, e_{BD}\}$. Pre-trained models of Hu et al. (2019) [5] are used to pre-train the model. The proposed model architecture is depicted in Figure 1. [1]

2.2 Convolutional Transformer Module

In this section, we investigate how to combine Transformers [6] and GNNs to capture the local spatial context and global information in molecular graph embeddings h_G. Inspired by the work of Wu et al. (2021) [7], we adapt a convolutional vision Transformer divided in $L_i = 3$ different steps as an efficient hierarchical structure for toxicity and side effect prediction. In this study, a

[1] In this figure, the nodes being operated on are shown in blue, while neighboring nodes are displayed in black. Blue and white squares represent node and graph embeddings h_v and h_G. For AGGREGATE, COMBINE, and UPDATE steps, graph operations are performed simultaneously on all nodes $v \in V$. We consider graph operations for $L_{GIN} = 5$ GIN layers, and a READOUT mean-pooling operation is performed at the final layer. The convolutional Transformer computes deep representations h_T using graph embeddings h_G of size 300.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

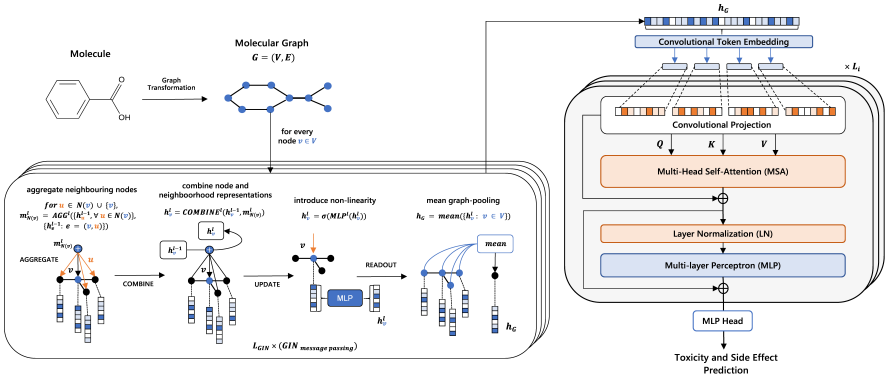

Fig. 1: Graphical depiction of the proposed GNN-Transformer, FS-GNNCvTR

convolutional Transformer converts graph embeddings h_G into a sequence of patches $p(h_G) = [x_p^1, x_p^2, ..., x_p^N]$ where x_p^k is the k-th patch vector with N the number of patches. For each step i, a convolutional embedding layer performs a convolution operation on a sequence of patches p to compute token embeddings h_T as the input tokens for a convolutional projection. Formally, we learn a function f_{conv} that converts patch tokens or 1D token sequences from the previous Transformer step $i-1$, $h_T^{i-1} \in \mathbb{R}^{P_{i-1} \times 1 \times C_{i-1}}$ with size P_{i-1} and channel size C_{i-1} into a new 1D token sequence $f_{conv}(h_T^{i-1})$. The function f_{conv} is a convolution operation with a kernel size s, output size o, stride $s-o$ and padding size k. The new token sequence h_T^i has size $P_i = [\frac{P_{i-1}+2k-s}{s-o} + 1]$ and followed by layer normalization (LN), reducing the number of tokens and increasing feature complexity across Transformer layers. Transformer blocks propagate token embeddings h_T across multi-head self-attention (MSA) layers which take queries, keys, and values (q, k, v) stacked into matrices (Q, K, V). Here, we consider multiple projection heads H in MSA and the attention scores are given by

$$MSA(Q, K, V) = CONCAT(head_1, ..., head_H)W \tag{2}$$

$$head_j = Attention(Q, K, V) = softmax(\frac{QW_j^Q(KW_j^K)^T}{\sqrt{d}})VW_j^V \tag{3}$$

with (W_j^Q, W_j^K, W_j^V) the projection matrices of (Q, K, V) and d the Transformer dimension for each head j. The convolutional projection is applied using a depth-wise separable convolution formulated by $h_T^{i*} = Conv(h_T^i, s)$ where h_T^{i*} are the input tokens for Q, K and V matrices at step i, s is the kernel size, and h_T^i are the input tokens for the convolutional projection followed by batch normalization and a point-wise convolution operation. Finally, a cls token at the last Transformer step summarizes the information captured by h_T is applied to a MLP followed by sigmoid activation to predict a molecular property label (condensed in a value $\in \{0, 1\}$).

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

2.3 Two-Module Meta-Learning Framework

In this study, we propose two-module meta-learning framework consisting of: a graph neural network (GNN) module and a convolutional Transformer (CvTR) module (see Figure 2). Both modules are trained to update model parameters across few-shot tasks (meta-training) using a task-specific support set for training and a disjoint query set for evaluation. The updated parameters are used to generalize to new representations in the test data (meta-testing). In this case, we focus on predicting the chemical toxicity and adverse side effects on Tox21 and SIDER datasets, so that $\{f_\theta(G), g_{\theta^*}(h_G)\} : M \Rightarrow \{0,1\} \in Y$, where M is the space of all molecular graph structures G, h_G are the output graph embeddings from a GNN f_θ, g_{θ^*} is the Convolutional Transformer (CvTR), and Y is the toxicity or side effect property labels. A GIN f_θ with model parameters θ and a Convolutional Transformer g_{θ^*} with parameters θ^* are trained across few-shot tasks $t \in \{1..., T\}$. For each task, meta-models f_θ and g_{θ^*} are trained on a task-specific support set S_t of molecular graphs $G_{S_{t_i}}$ and evaluated on a query set Q_t of molecular graphs $G_{Q_{t_i}}$. In meta-training, a support set of size k is randomly sampled to serve as an input to the GNN-Transformer and compute the support losses $\mathcal{L}_t^{GNN}$, $\mathcal{L}_t^{CvTR}$ for each task $t \in \{1..., n_{train}\}$. Support losses are then used to iteratively update model parameters $\theta \rightarrow \theta'$, $\theta^* \rightarrow \theta^{*'}$. Both meta-models compute the query losses $\mathcal{L}_t^{GNN'}$,$\mathcal{L}_t^{CvTR'}$ using the remaining n samples for that task. In meta-training, to update model parameters, we apply a few gradient steps, $\theta_t = \theta - \alpha \nabla_\theta \mathcal{L}_t^{GNN}(\theta)$ and $\theta_t^* = \theta^* - \alpha^* \nabla_{\theta^*} \mathcal{L}_t^{CvTR}(\theta^*)$ where α and α^* are the size of the steps for the gradient descent updates. In meta-testing, a support set of k examples is randomly sampled for a new test task to update model parameters $\theta \rightarrow \theta'$, $\theta^* \rightarrow \theta^{*'}$ for each task $t \in \{n_{train} + 1..., T\}$. Then, we evaluate both meta-models on a new query set using the remaining n samples, to predict the toxicity and side effect properties with just a few labeled molecules. The loss for both models, $\mathcal{L}^{GNN}$ and $\mathcal{L}^{CvTR}$ is a binary cross-entropy loss. To

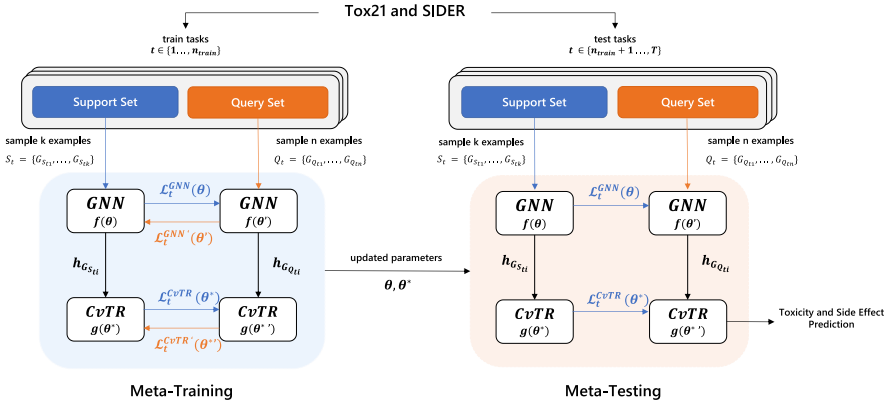

Fig. 2: Few-shot meta-learning framework for toxicity and side effect prediction

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

address class imbalance, we use a weighted binary cross-entropy loss to assign a higher penalty to failed predictions on the minority class by defining a weight c,

$$\mathcal{L} = -\frac{1}{k}\sum_{i=1}^{k} c\ y_i\ log(y_i') + (1 - y_i)\ log(1 - y_i') \tag{4}$$

where y' are the predictions and y are the molecular property labels with k representing the number of samples. A value of $c = 25$ was effective for Tox21, while a value of $c = 1$ was more effective for SIDER due to task variability.

2.4 Details of Model Training and Implementation

Models were implemented using PyTorch version 1.10.1 with CUDA version 11.3 and Python 3.7. RDKit libraries are used to compute the molecular graph features including atom-type, atom chirality, bond type and bond direction in a given compound. These attributes formed the initial set of molecular features fed into GNN layers. Models were trained across ($n_{train} \times epochs$) training episodes with n_{train} number of training tasks and $epochs$ the total number of epochs. In general, models stopped improving significantly after 1000 epochs. In this study, our primary focus was not on hyper-parameter optimization, particularly for GNN baselines. Consequently, we did not put extensive effort in fine-tuning model hyper-parameters, leaving it for future research. Specifically, we used a learning rate of $1e^{-5}$ and an update step of 5 for training and 10 for testing.

3 Results and Conclusion

In this work, we evaluated the binary classification across few-shots tasks for toxicity and side effect prediction on Tox21 and SIDER [4]. Tox21 includes the screening results for 12 different toxic effects in 7831 molecules. Here, we consider a train-test split with 9 tasks for meta-training and 3 for meta-testing. SIDER is a compound database that includes information on the potential side effects of 1427 molecules on 27 different organ classes. Here, information is divided in 21 tasks for meta-training and 6 for meta-testing. ROC-AUC scores are evaluated on the query set of each test task. For each few-shot task, we sample a random support set that includes a set of positive samples $k+$ and a set of negative samples $k-$ for training. The remaining datapoints of that task form the query set for evaluation. Here, we conduct 5-shot and 10-shot experiments with random support sets of size $(k+, k-)$, with $k = 5$ and $k = 10$, respectively. Each experiment is repeated 30 times, using different random support sets each time. In Table 1, we present the mean and standard deviation of ROC-AUC scores for 5-shot and 10-shot experiments with 30 random support sets. ROC-AUC results reported show a superior performance of the proposed model over other graph-based methods (GCN, GraphSAGE and GIN) and more robust results with a smaller variance across few-shot tasks. Ultimately, we posit that the FS-GNNCvTR model combines the strengths of both GNNs and convolutional

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Transformers to capture both local and global information of molecular embeddings and improve the process of drug discovery through the accurate prediction of toxicity and side effect properties with a small amount of labeled molecules.

Dataset	Task	GIN	GCN	GraphSAGE	FS-GNNCvTR (GIN+CvTR)	$\triangle(AUC)$
			5-shot $(5+, 5-)$			
Tox21	**SR-HSE**	0.612 ± 0.009	0.661 ± 0.019	0.651 ± 0.039	$\mathbf{0.778 \pm 0.002}$	$+0.117$
	SR-MMP	0.578 ± 0.009	0.657 ± 0.015	0.667 ± 0.032	$\mathbf{0.796 \pm 0.001}$	$+0.129$
	SR-p53	0.590 ± 0.011	0.621 ± 0.015	0.644 ± 0.025	$\mathbf{0.740 \pm 0.002}$	$+0.096$
	Average	0.593	0.646	0.654	**0.771**	$+0.117$
SIDER	**R.U.D.**	0.697 ± 0.012	0.609 ± 0.011	0.630 ± 0.009	$\mathbf{0.715 \pm 0.002}$	$+0.018$
	P.P.P.C.	$\mathbf{0.769 \pm 0.006}$	0.719 ± 0.014	0.723 ± 0.012	0.738 ± 0.003	-0.031
	E.L.D.	0.703 ± 0.007	0.631 ± 0.014	0.648 ± 0.008	$\mathbf{0.723 \pm 0.002}$	$+0.020$
	C.D.	0.683 ± 0.010	0.607 ± 0.013	0.628 ± 0.012	$\mathbf{0.729 \pm 0.002}$	$+0.046$
	N.S.D.	0.650 ± 0.008	0.585 ± 0.021	0.596 ± 0.011	$\mathbf{0.672 \pm 0.005}$	$+0.022$
	I.P.P.C.	0.731 ± 0.011	0.658 ± 0.015	0.677 ± 0.012	$\mathbf{0.738 \pm 0.002}$	$+0.007$
	Average	0.706	0.635	0.650	**0.719**	$+0.013$
			10-shot $(10+, 10-)$			
Tox21	**SR-HSE**	0.655 ± 0.013	0.652 ± 0.021	0.668 ± 0.023	$\mathbf{0.784 \pm 0.002}$	$+0.116$
	SR-MMP	0.626 ± 0.014	0.651 ± 0.014	0.691 ± 0.016	$\mathbf{0.801 \pm 0.001}$	$+0.110$
	SR-p53	0.629 ± 0.010	0.633 ± 0.013	0.651 ± 0.014	$\mathbf{0.741 \pm 0.003}$	$+0.090$
	Average	0.637	0.652	0.670	**0.775**	$+0.105$
SIDER	**R.U.D.**	0.691 ± 0.004	0.600 ± 0.006	0.637 ± 0.006	$\mathbf{0.712 \pm 0.002}$	$+0.021$
	P.P.P.C.	$\mathbf{0.780 \pm 0.008}$	0.716 ± 0.015	0.735 ± 0.009	0.725 ± 0.004	-0.055
	E.L.D.	0.709 ± 0.004	0.617 ± 0.009	0.645 ± 0.007	$\mathbf{0.723 \pm 0.003}$	$+0.014$
	C.D.	0.676 ± 0.008	0.589 ± 0.012	0.633 ± 0.008	$\mathbf{0.732 \pm 0.002}$	$+0.056$
	N.S.D.	0.660 ± 0.009	0.577 ± 0.016	0.593 ± 0.016	$\mathbf{0.674 \pm 0.006}$	$+0.014$
	I.P.P.C.	$\mathbf{0.738 \pm 0.009}$	0.639 ± 0.016	0.687 ± 0.010	0.734 ± 0.002	-0.004
	Average	0.709	0.623	0.655	**0.717**	$+0.008$

Table 1: Average ROC-AUC scores obtained across 30 experiments with support sets of size $(5+, 5-)$ (5-shot) and $(10+, 10-)$ (10-shot) on Tox21 and SIDER

References

[1] Han Altae-Tran, Bharath Ramsundar, Aneesh S. Pappu, and Vijay Pande. Low data drug discovery with one-shot learning. *ACS Central Science*, 3, 2017.

[2] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. volume 3, pages 2053–2070. International Machine Learning Society (IMLS), 2017.

[3] Zhichun Guo, Chuxu Zhang, Wenhao Yu, John Herr, Olaf Wiest, Meng Jiang, and Nitesh V. Chawla. Few-shot graph learning for molecular property prediction. 2021.

[4] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2), 2018.

[5] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. 5 2019.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 10 2020.

[7] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. pages 22–31. Institute of Electrical and Electronics Engineers Inc., 2021.