Graph Convolutional Model to Understand Quantitative Systems Pharmacology

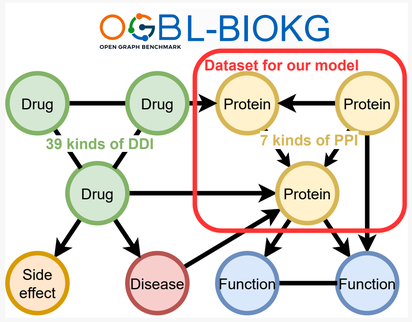
Sungwoo Goo[1] Sangkeun Jung* [1,3] Jung-woo CHAE* [1,2] Hwi-yeol Yun*[1,2]

[1]Department of Bio-AI convergence, Chungnam National University, Daejeon
[2]College of Pharmacy, Chungnam National University, Daejeon
[3]Department of Computer Science and Engineering, Chungnam National University, Daejeon

Objectives

Understanding quantitative systems pharmacology is essential for drug development. Protein-Protein Interaction (PPI) is the basic unit of life activity, and to analyze these PPIs holistically, Protein Graph Network is analyzed. Protein information can be represented by amino acid sequences, and it is reasonable to use a model pre-trained on amino acid sequences. Since none of the state-of-the-art models in the OBGL-BIOKG dataset competition[1] used a model pre-trained on amino acid sequences, we built a Graph Convolutional Model using a Large Language Model (LLM) for amino acid sequences to predict seven types of PPIs.

Methods



Node	Count
Disease	10687
Protein	17499
Drug	10533
Side effect	9969
Function	45085

PPI Edge	Count
Activation	73044
Binding	292254
Catalysis	303434
Expression	1952
Inhibition	25732
Post-Translational modification	15120
Reaction	352546

Figure 1: **Exploratory data analysis:**
open graph benchmark dataset scheme of biology knowledge graph and Counts for every node and every PPI edge of OGBL-BIOKG dataset

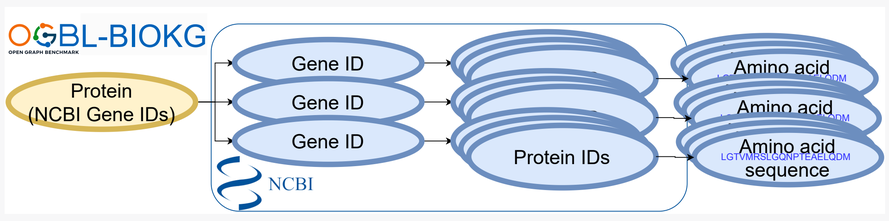


Figure 2: **Data Mining:**
For building nodes in the protein graph, amino acid sequences were collected. Protein nodes were represented only by Gene IDs in the OGBL-BIOKG dataset. Using the gene IDs, we imported the amino acid sequences from NCBI

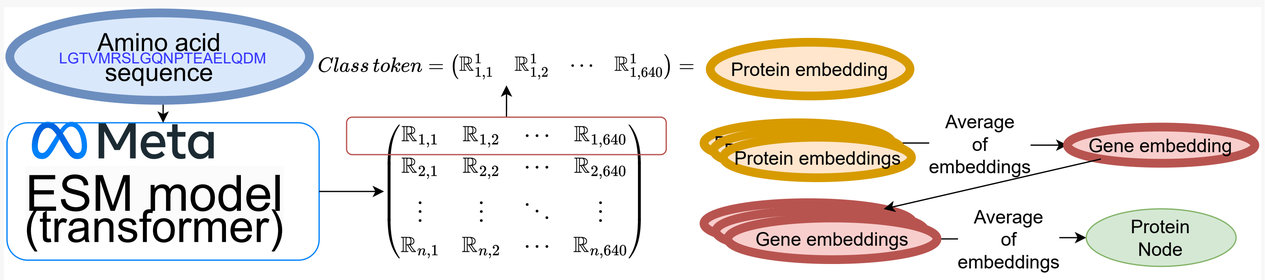


Figure 3: **Data Process:**
Amino acid sequences were processed with Meta's ESM model (Large Language Model)[2] to embed the protein nodes into a 640-dimensional vector.

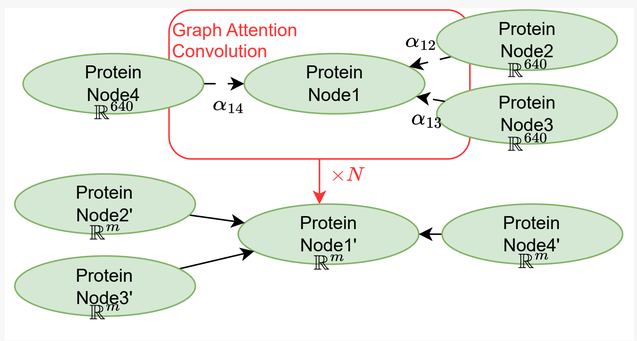
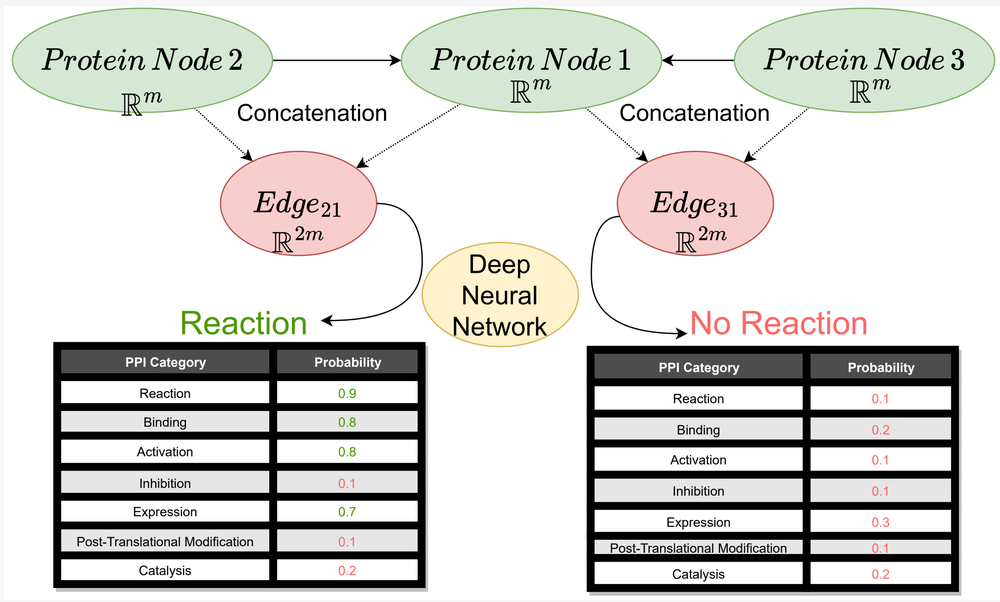


Figure 4: **Graph Convolutional Network:**
GAT(Graph ATtention Network)[3] were used to calculate PPI. α is the attention score for a protein node directed to another protein node. In the example above, the sum of α_{12}, α_{13}, and α_{14} is 1. The αs are weights for how much protein 1 was affected by protein 2, protein 3, and protein 4, respectively.



PPI Category	Probability
Reaction	0.9
Binding	0.8
Activation	0.8
Inhibition	0.1
Expression	0.7
Post-Translational Modification	0.1
Catalysis	0.2

PPI Category	Probability
Reaction	0.1
Binding	0.2
Activation	0.1
Inhibition	0.1
Expression	0.3
Post-Translational Modification	0.1
Catalysis	0.2

Figure 5: **PPI Prediction:**
The protein node vectors were computed by graph convolution using the GAT layer. Then, the PPI edge vectors to be predicted were generated by concatenating the two protein node vectors, and the PPI edge vector is fed into the deep neural network to predict 7 probabilities.

Results

Figure 6: **Train records with pre-trained protein nodes**

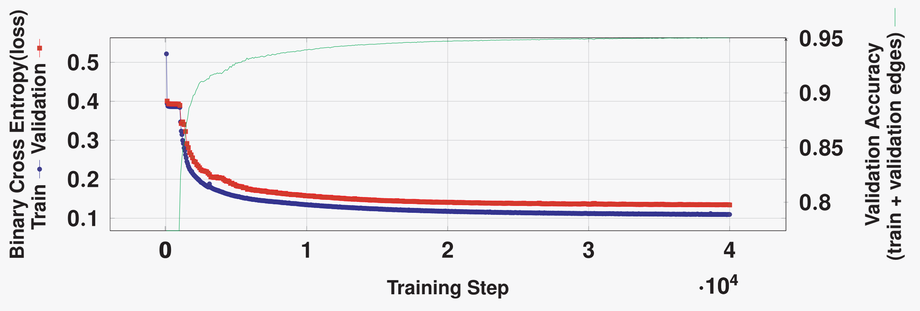


Figure 7: **Train records without pre-trained protein nodes**

	With pretrain		Without pretrain	
PPI	Accuracy	AUROC	Accuracy	AUROC
Total	0.81	0.88	0.81	0.89
Activation	0.92	0.95	0.93	0.95
Binding	0.65	0.76	0.65	0.75
Catalysis	0.61	0.69	0.61	0.71
Expression	1.00	0.88	1.00	0.88
Inhibition	0.97	0.96	0.97	0.96
Post-translational modification	0.98	0.97	0.98	0.96
Reaction	0.57	0.66	0.57	0.67

Table 1: **validation edges' prediction accuracy and Area Under Receiver Operating characteristic Curves(AUROC). (without train edges)**

	Number of parameters	
Layer	with pre-train	without pre-train
GCN	5.8*10^6	5.8*10^6
Classifier	287	287
Protein Embedding	0	11.2*10^6

Table 2: **Numbers of model parameters. Using the pre-train model, we achieved similar performance with fewer parameters.**

Conclusions

We built a deep learning model to understand protein-protein interactions using the Graph Convolutional Model. With appropriate prior knowledge, we were able to achieve similar performance with fewer parameters, and we built models that used more biological knowledge than models that did not use amino acid sequences. and Since our model uses the pre-train model when adding new protein nodes, we can maintain relatively high performance without retraining the GCN. The Large Language Model (LLM), which is currently in the spotlight, has the disadvantage of being non-explanatory, so the output of the LLM needs to be verified by experts in each domains. Graph models can be understood by experts in different domains and are therefore more suitable as a decision-making tool in the pharmaceutical industry, where several experts work together.

References

[1] **Weihua Hu et al.** "Open graph benchmark: Datasets for machine learning on graphs". In: *Advances in neural information processing systems* 33 (2020), pp. 22118–22133.

[2] **Zeming Lin et al.** "Evolutionary-scale prediction of atomic-level protein structure with a language model". In: *Science* 379.6637 (2023), pp. 1123–1130.

[3] **Shaked Brody, Uri Alon, and Eran Yahav.** "How attentive are graph attention networks?" In: *arXiv preprint arXiv:2105.14491* (2021).

Acknowledgement

This work was partly supported by Institute of Information Communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-00155857, Artificial Intelligence Convergence Innovation Human Resources Development (Chungnam National University)), Research Foundation of Korea (NRF) grants from the Korean government (MSIT) (Nos. NRF-2022R1A2C1010929 and 2022R1F1A1071047) and Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. 2022R1A5A7085156, Senior Health Convergence Research Center based on Life Cycle (Chungnam National University))

swgoo@o.cnu.ac.kr, {hugman, jwchae, hyyun}@cnu.ac.kr