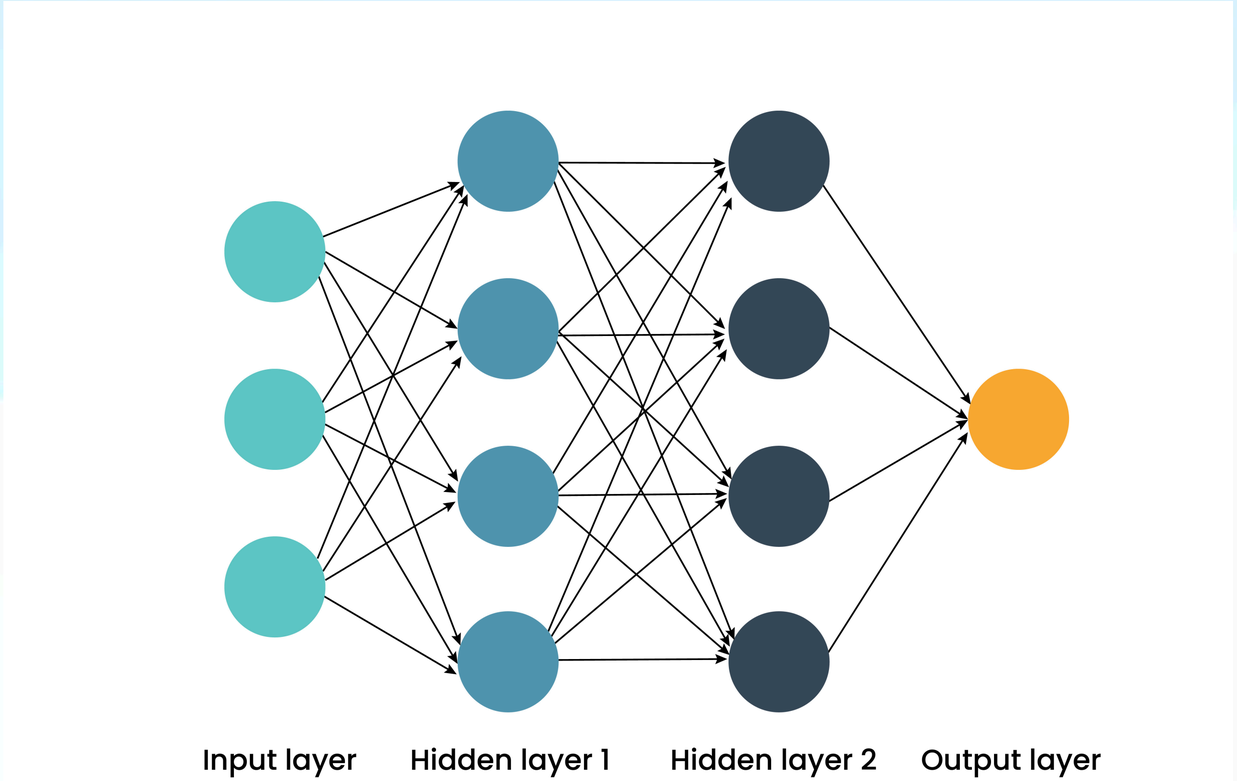
Towards Graph Foundation Models: A Survey and Beyond

Abdellah Rahmani

Deep learning model

Powerful tool with different limitation



What are the limitations of a deep learning model have ?

Deep learning models

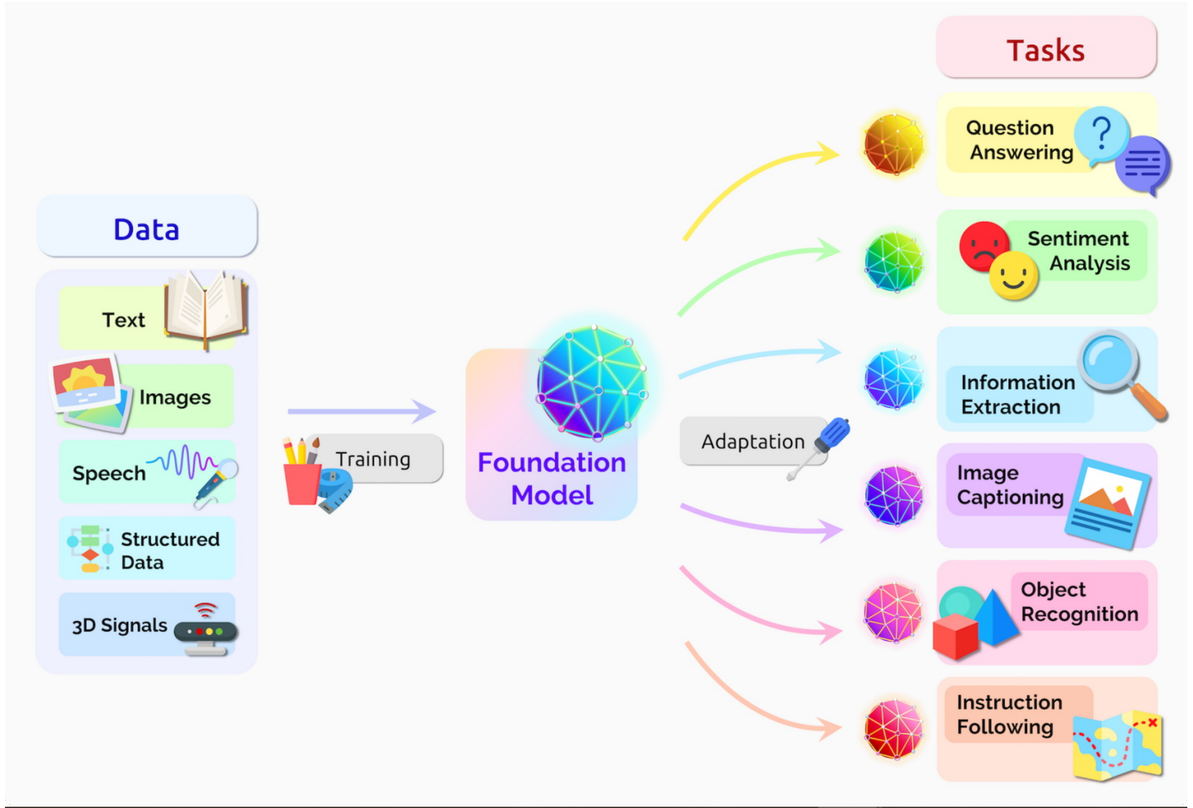
Limitations

- Huge amount of labeled data
- Can not use unlabelled data
- Can be trained and build for one task
- Face challenges to adapt to new tasks/datasets
- Out of distribution issue

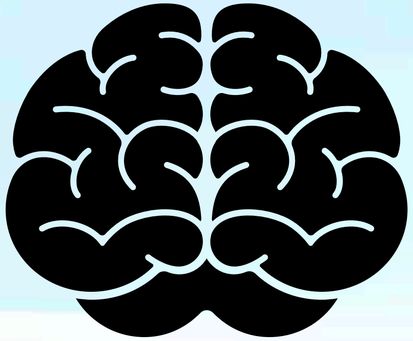
Foundation model

Definition

A foundation model is any model that is trained on broad data and can be adapted to a wide range of downstream tasks

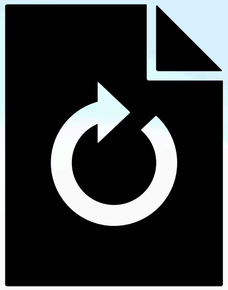


Foundation models



Emergence:

Manifesting novel capabilities



Homogenisation

Deployment across diverse application

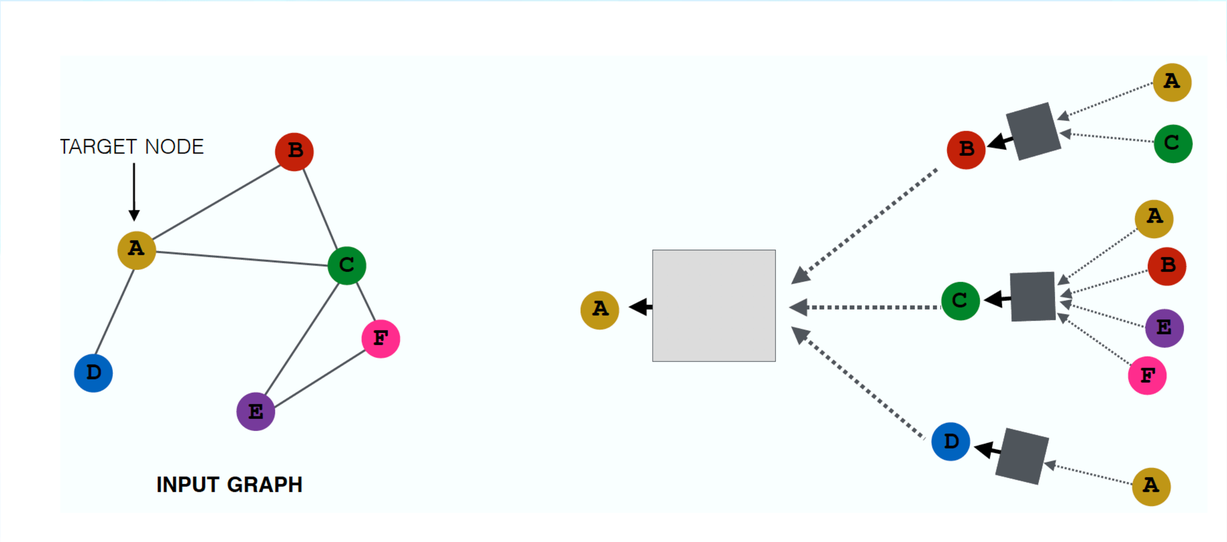
Graph neural network

- GNNs yield to many improvements in different tasks like: graph classification, link prediction and node classification…
- GNN suffer from many limitations, what are these limitations and their causes ?

GNNs

Limitation and causes

Expressive power issue:



GNNs relies on message passing:

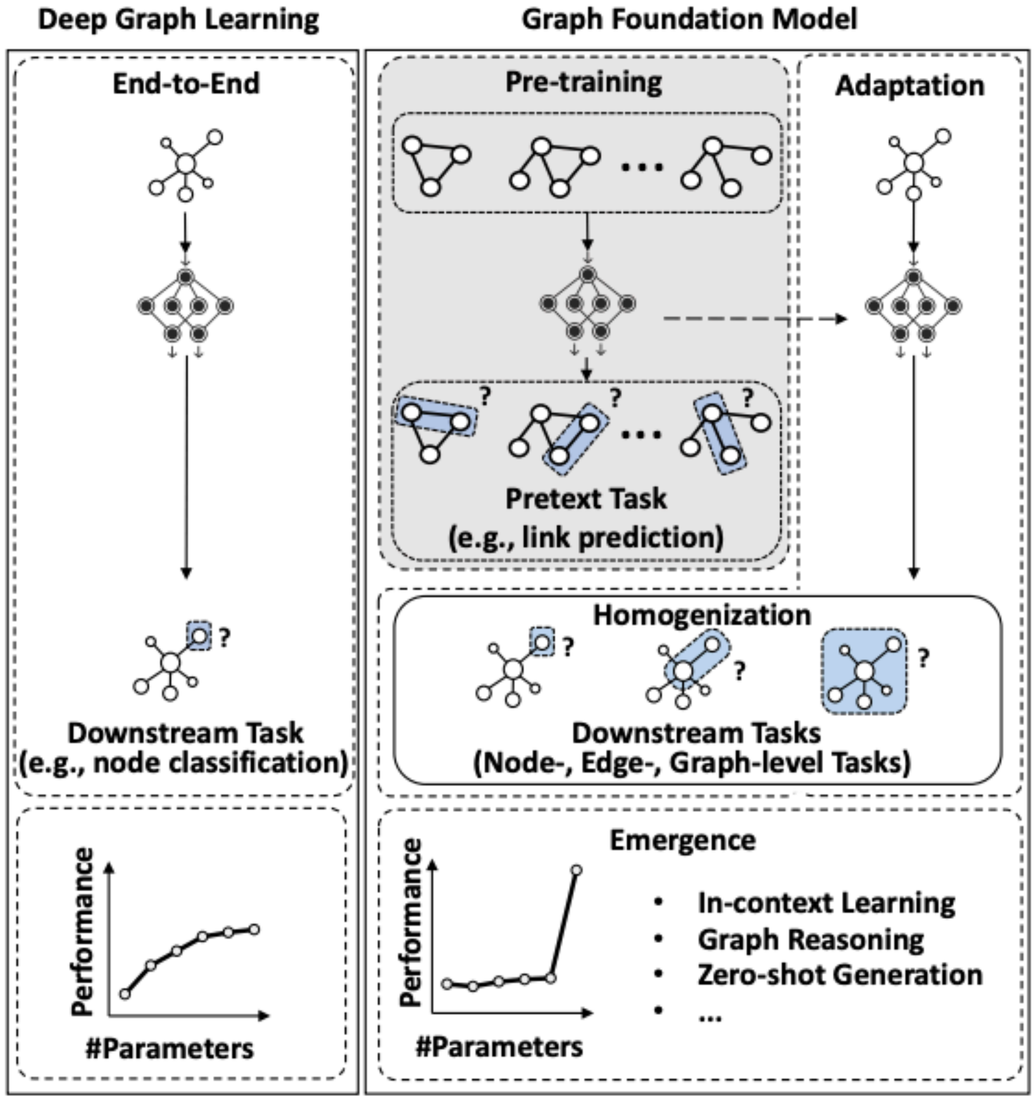
Difficulties to distinguish certain types of non-isomorphic graphs

GNNs

- Inability to Distinguish Certain Graphs:
    a. GNNs can fail to differentiate between **structurally similar but non-isomorphic graphs**
- Over-Smoothing:
    b. As layers increase, node representations become **too similar**
- GNNs struggle to capture **higher-order relationships or motifs in the graph**

Could graph foundations models represent the next frontier in graph machine learning ?

Graph foundation model



Can we easily achieve and build a graph foundation model ? What are the challenges ?

Language foundation models

- LLMs are trained on **extensive and diverse datasets**
- Trained using **self-supervised learning**
- Tackle a **broad spectrum** of downstream tasks

How these LLMs achieve such performance ?
What are the key components ?

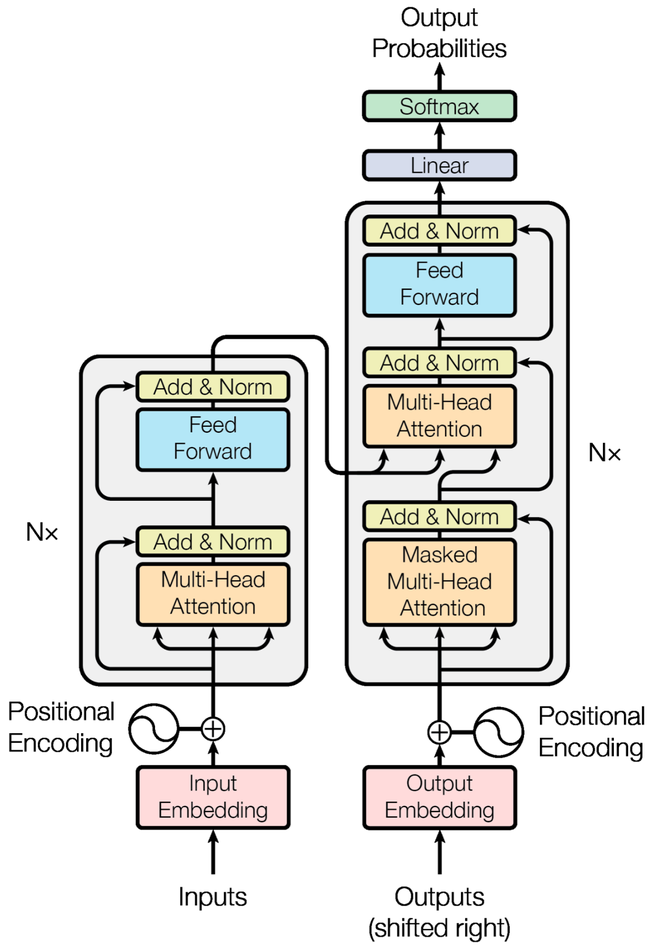
LLMs

Language data

- Language data is Euclidean, hence **easy** to model
- Rich of **semantic information**
- The quality and the quantity of this type of data enhance **knowledge transferability**

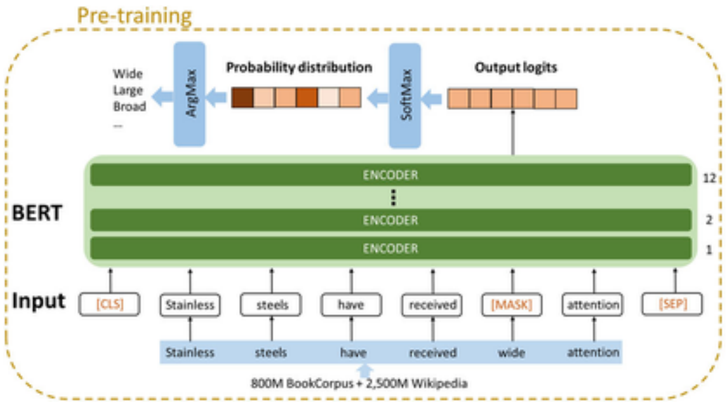
LLMs

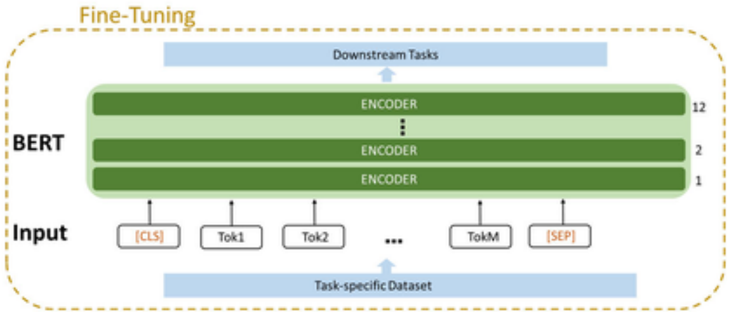
Backbone architecture



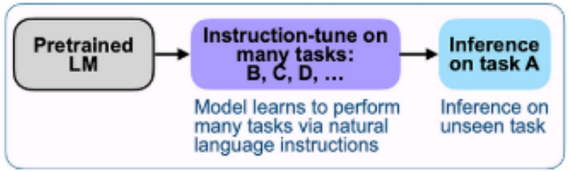
Unified learning paradigms

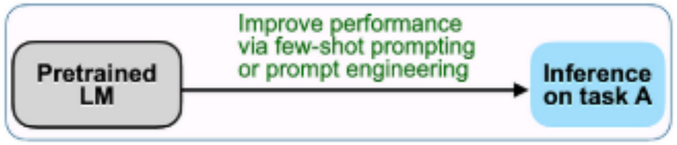
Pretrain and fine tune





Pretrain, prompt and predict





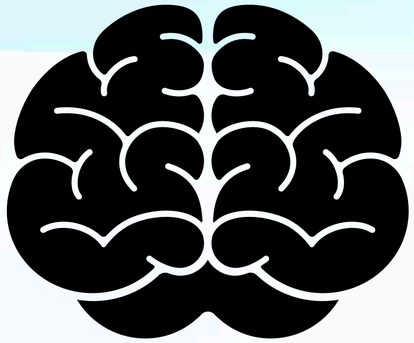
Graph Foundation models

Can you give based on the aforementioned key components, what are the essential abilities that we want to have in GFM ? A definition of GFM ?

GFM

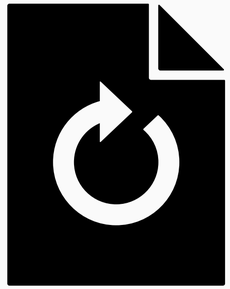
Definition

A graph foundation model (GFM) is a model that is expected to benefit from the pre-training of broad graph data, and can be adapted to a wide range of downstream graph tasks.



Emergence:

Manifesting novel capabilities



Homogenisation

Deployment across diverse application

Challenges

Impact from graph data

Graph type:

- **Homogeneous and heterogeneous** graphs (difficulties to define a unified backbone)
- **Dynamic graph** that poses additional challenges

Graph scale:

- **Large graph** impose higher demands on the capacities of GFM (long range dependency)

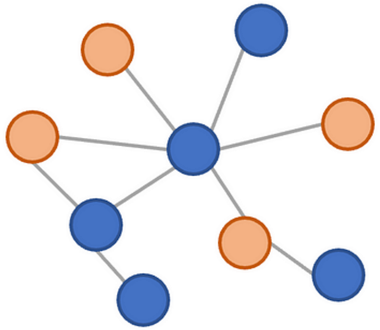
Graph diversity:

- **Same domain** graph or **cross domain** graphs

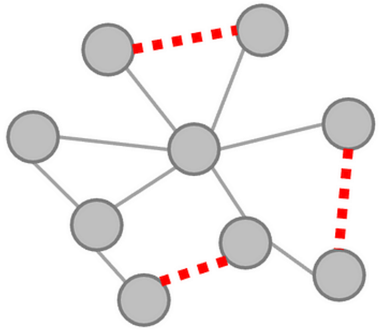
Challenges

Impact from graph tasks

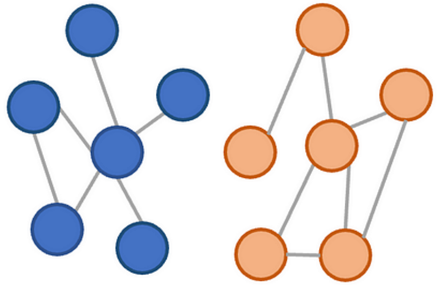
Node Classification



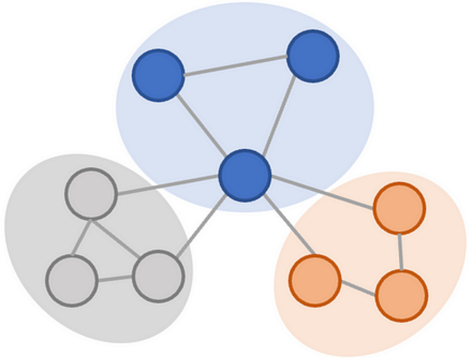
Link Prediction



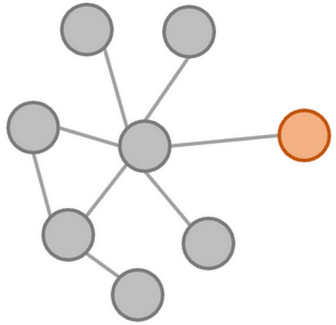
Graph Classification



Community Detection



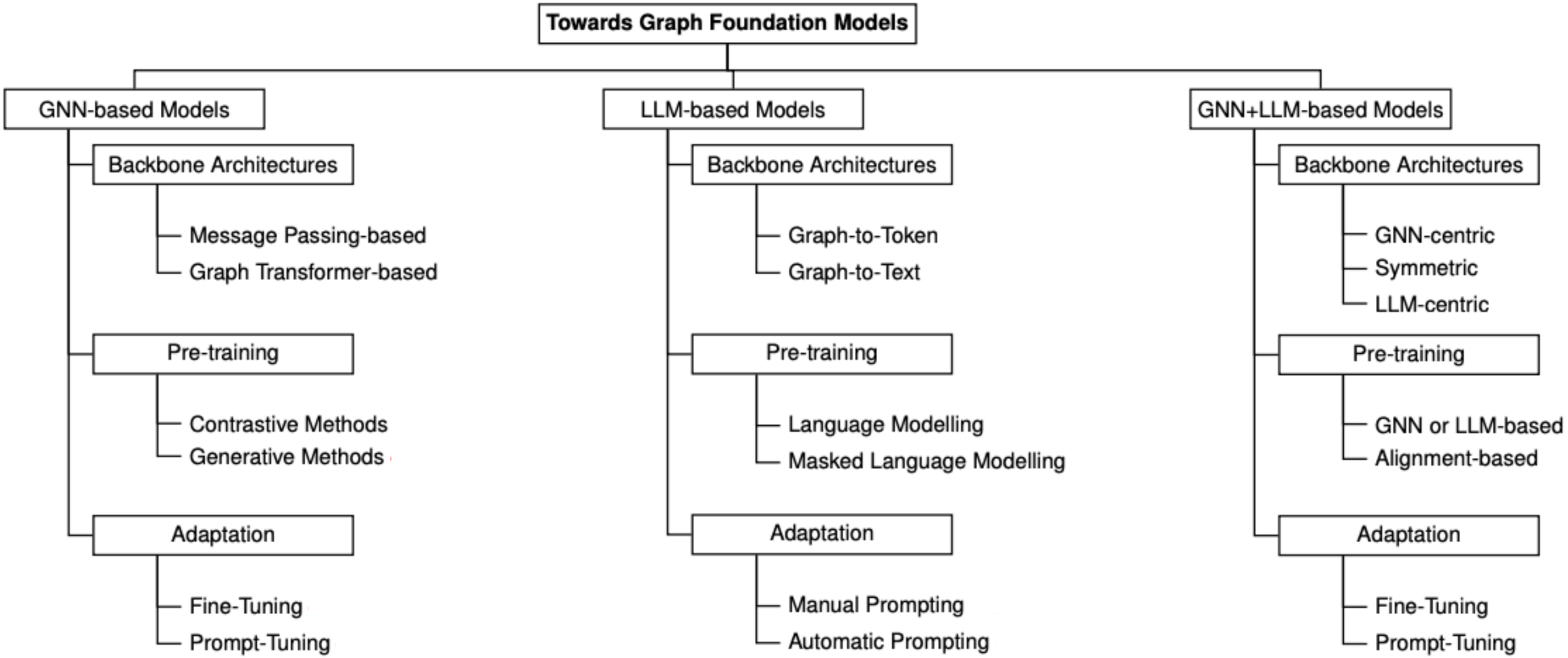
Anomaly Detection



LLMs Vs GFMs

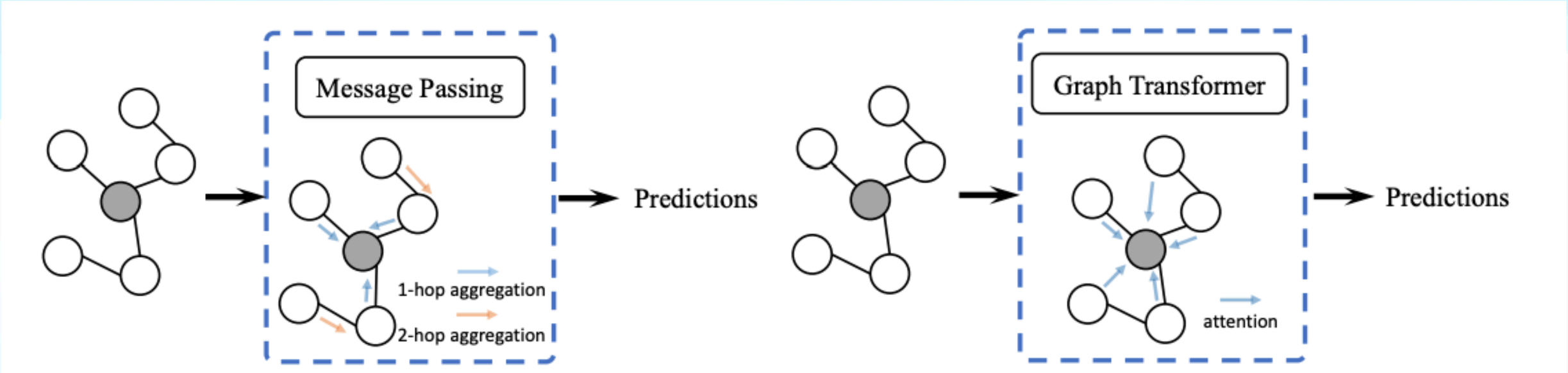
		Language Foundation Model	Graph Foundation Model
Similarities	Goal	Enhancing the model's expressive power and its generalization across various tasks	
	Paradigm	Pre-training and Adaptation	
Intrinsic differences	Data	Euclidean data (text)	Non-Euclidean data (graphs) or a mixture of Euclidean (e.g., graph attributes) and non-Euclidean data
	Task	Many tasks, similar formats	Limited number of tasks, diverse formats
Extrinsic differences	Backbone Architectures	Mostly based on Transformer	No unified architecture
	Homogenization	Easy to homogenize	Difficult to homogenize
	Domain Generalization	Strong generalization capability	Weak generalization across datasets
	Emergence	Has demonstrated emergent abilities	No/unclear emergent abilities as of the time of writing

GFMs


Towards Graph Foundation Models
GNN-based Models
Backbone Architectures
Message Passing-based
Graph Transformer-based
Pre-training
Contrastive Methods
Generative Methods
Adaptation
Fine-Tuning
Prompt-Tuning
LLM-based Models
Backbone Architectures
Graph-to-Token
Graph-to-Text
Pre-training
Language Modelling
Masked Language Modelling
Adaptation
Manual Prompting
Automatic Prompting
GNN+LLM-based Models
Backbone Architectures
GNN-centric
Symmetric
LLM-centric
Pre-training
GNN or LLM-based
Alignment-based
Adaptation
Fine-Tuning
Prompt-Tuning

GNN based models

Backbone architecture



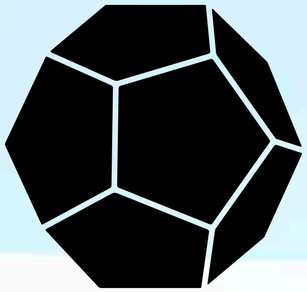
GNN based models

Pre-training

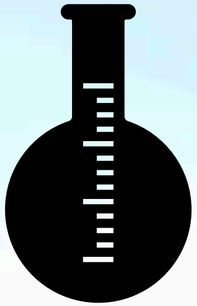
- Contrastive methods: aim to **maximise mutual information** between different views
  - Same scale contrastive learning: Consider different subgraphs of the same nodes as **positive examples**
  - Cross scale contrastive learning: compares two graph views at different levels (node and graph embeddings)
- Generative methods; graph reconstruction that aim to reconstruct specific parts of given graphs

GNN based models

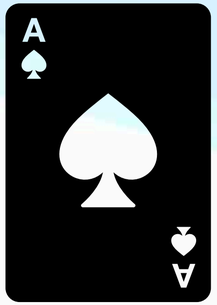
Adaptation



Fine tuning



Pre-prompt

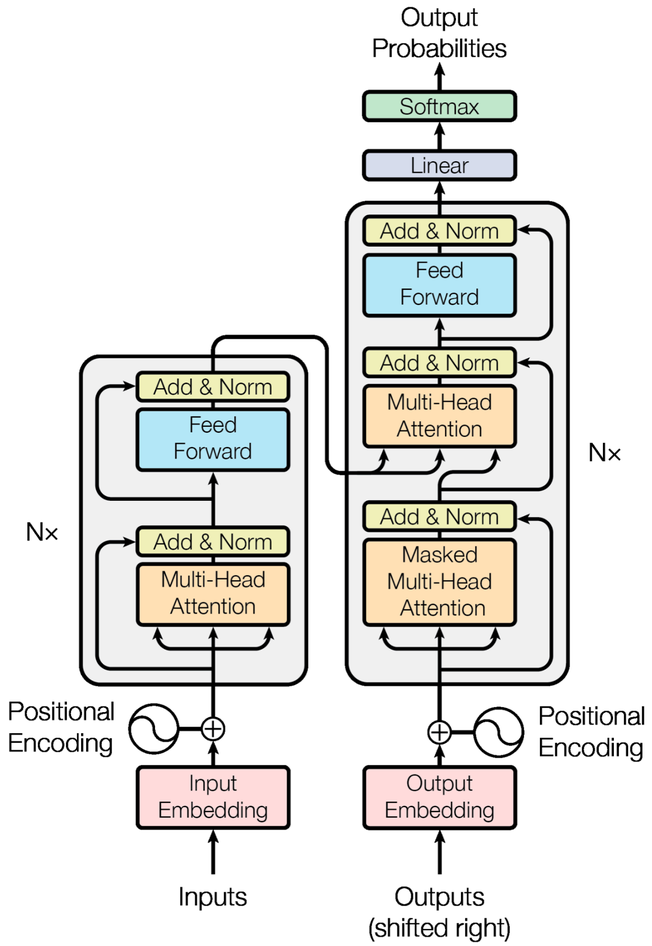


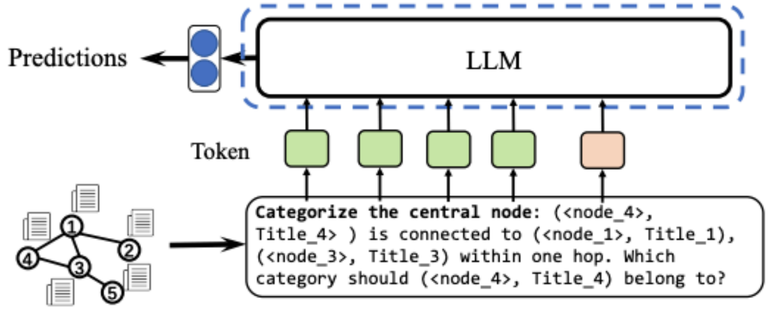
Post-prompt

What are the limitations of these approaches ?

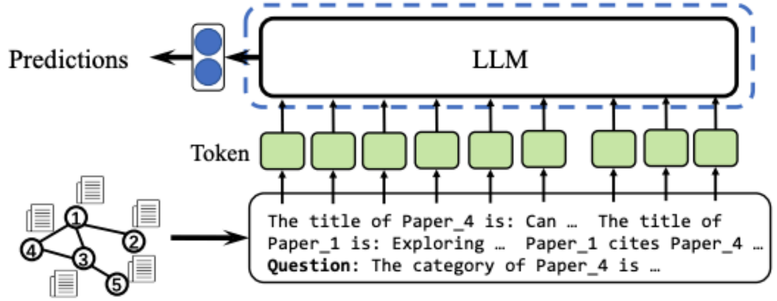
LLMs based models

Backbone architecture





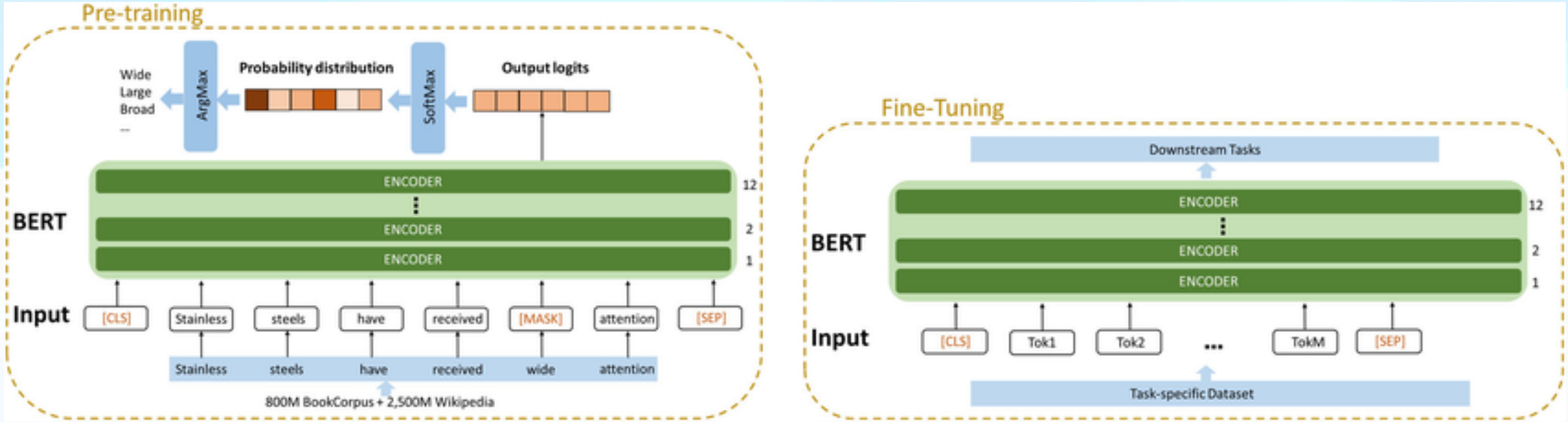
(a) Graph-to-token.



(b) Graph-to-text.

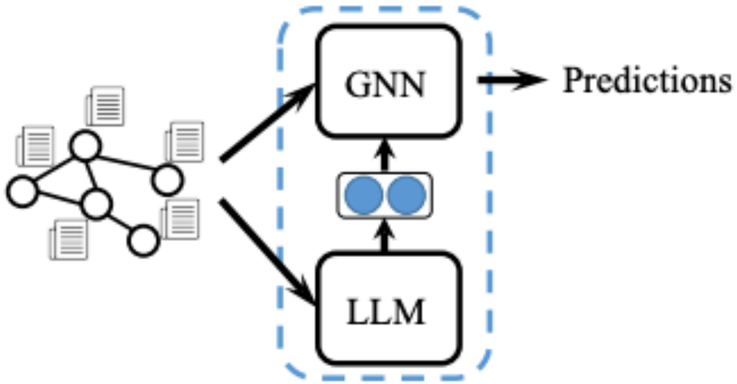
LLMs based models

Pre-training and fine tuning

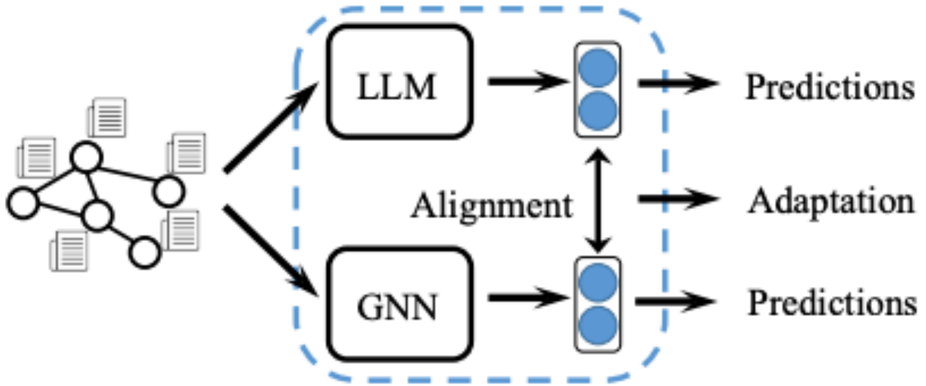


What are the limitations of these approaches ?

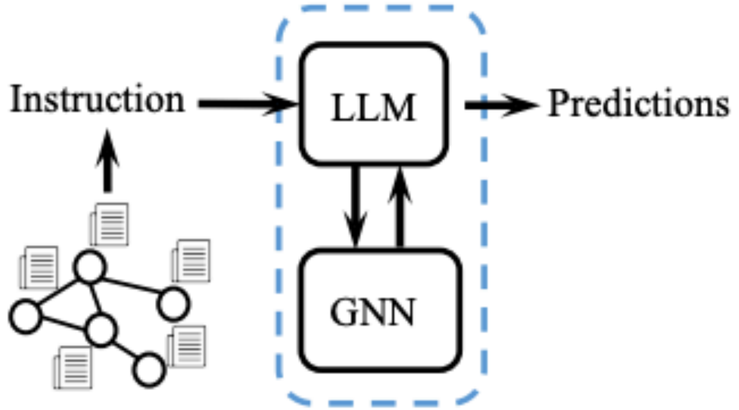
GNN+ LLMs based models



(a) GNN-centric methods.



(b) Symmetric methods, where the aligned embeddings can be further utilized for downstream tasks.

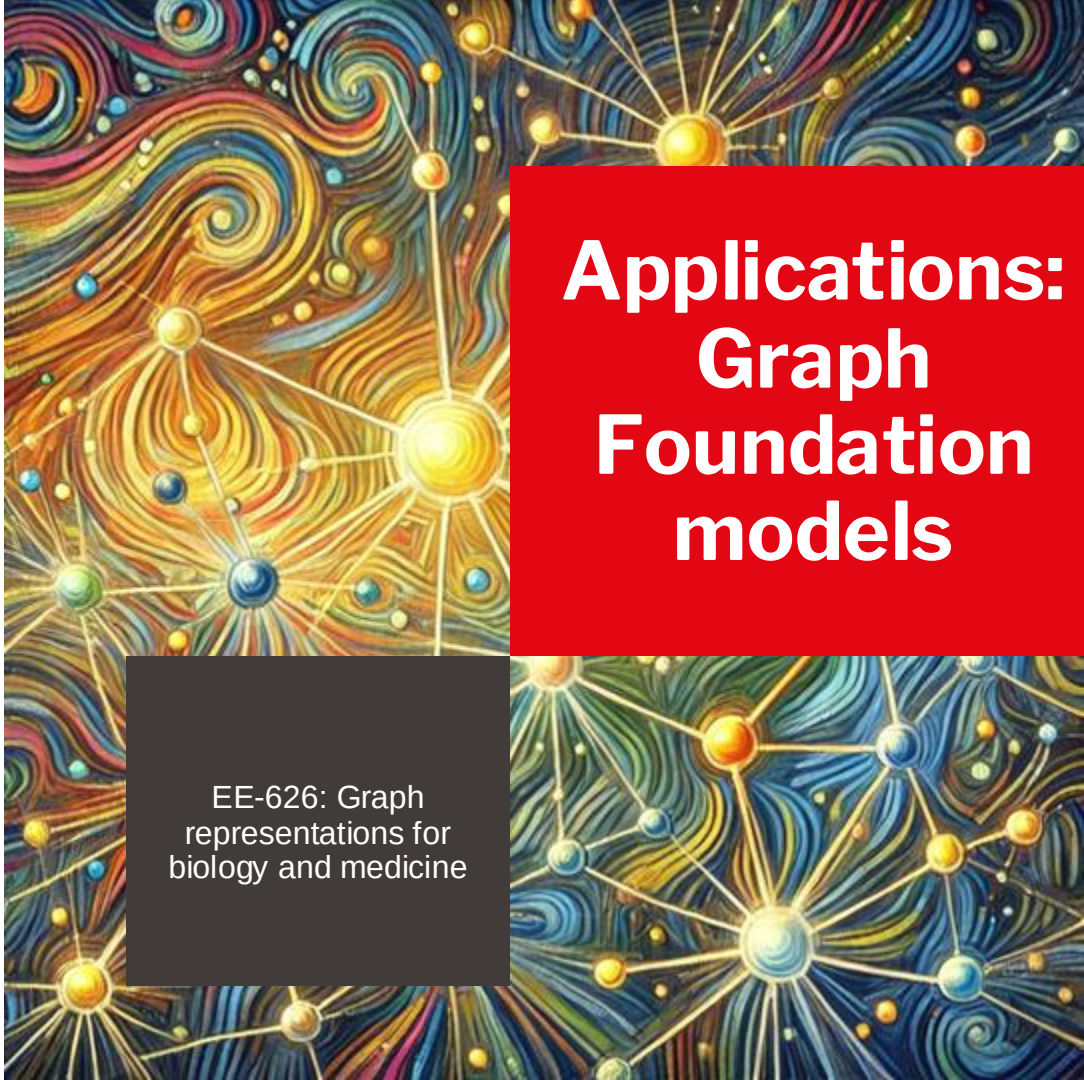


(c) LLM-centric methods, which take an instruction as input and output an answer.

Conclusion

- GFMs target to build model that can handle different tasks
- Multiple challenges are present
- Graph data is challenging
- Graph tasks are very different
- Can you think about some biomedical application where you can apply one of the aforementioned technics ? If not why is it not possible ?

EPFL



Applications: Graph Foundation models

EE-626: Graph representations for biology and medicine

Vincent Jung

My research

Self-supervised learning on RNA sequences …

A C G T A C G T A C G T

Masking

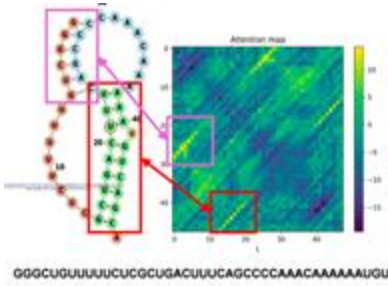
A C G _ A C G T _ C G T

RNA Language Model

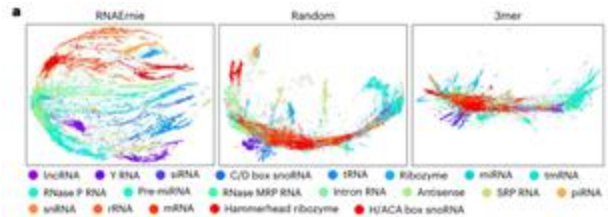
A C G T A C G T G C G T

… can create complex representations with structural information …

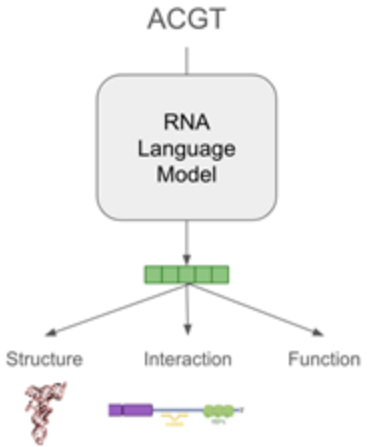
1



2



…which can then be used for many downstream tasks.



1. Yin, Weijie, et al. "ERNIE-RNA: An RNA Language Model with Structure-enhanced Representations." bioRxiv (2024): 2024-03.
2. Wang, N., Bian, J., Li, Y. *et al.* Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning. *Nat Mach Intell* **6**, 548–557 (2024). https://doi.org/10.1038/s42256-024-00836-4

A foundation model for clinician-centered drug repurposing

Kexin Huang, Payal Chandak, Qianwen Wang, Shreyas Havaldar, Akhil Vaid, Jure Leskovec, Girish N. Nadkarni, Benjamin S. Glicksberg, Nils Gehlenborg & Marinka Zitnik

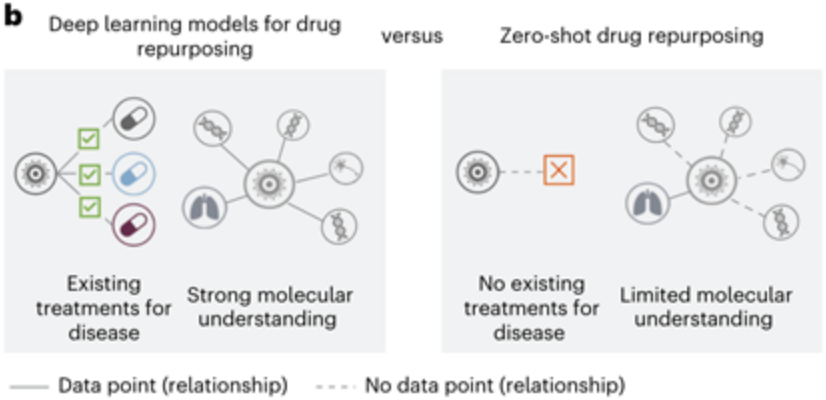
Goal: Drug repurposing

- Find new use for an already approved drug since drugs can have a pleiotropic effect
- ~30% of FDA-approved drugs are issued a new indication post-approval
- Most of these new purposes are found semi-randomly, through observation by clinicians or reported patient experience
- Why:
  - Lower costs of development (drug is already tested for safety)
  - Potential to find new use for existing drugs on rare diseases (7000 rare diseases, 5-7% have a FDA-approved drug)

- ✓ Existing preclinical knowledge
- ✓ Lower development costs
- ✓ Improved efficacy across patient groups
- ✓ Faster clinical translation

How: Drug repurposing

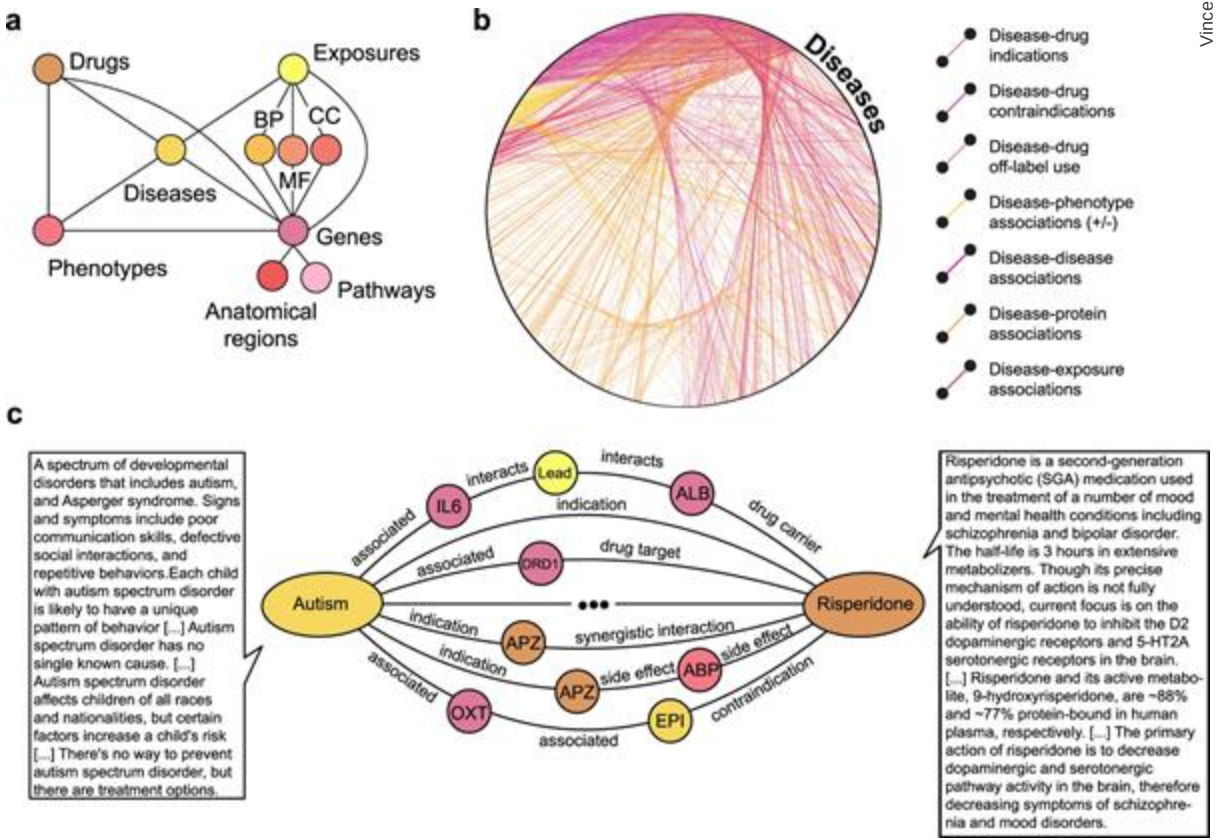
- Previous methods assume that we have either a strong understanding of the disease, and/or existing treatments for the disease.
- This may not be the case for rare diseases



Data : Knowledge graph

Disease and drug nodes have multiple features associated, all in text, that include:

- **Disease**: definitions, prevalence, epidemiology, clinical descriptions and management/treatment, symptoms, causes, risk factors, complications, and prevention
- **Drug**: description, indication, mechanism of action, Anatomical Therapeutic Chemical (ATC) code, pharmacodynamics, half-life, protein binding information, and pathways

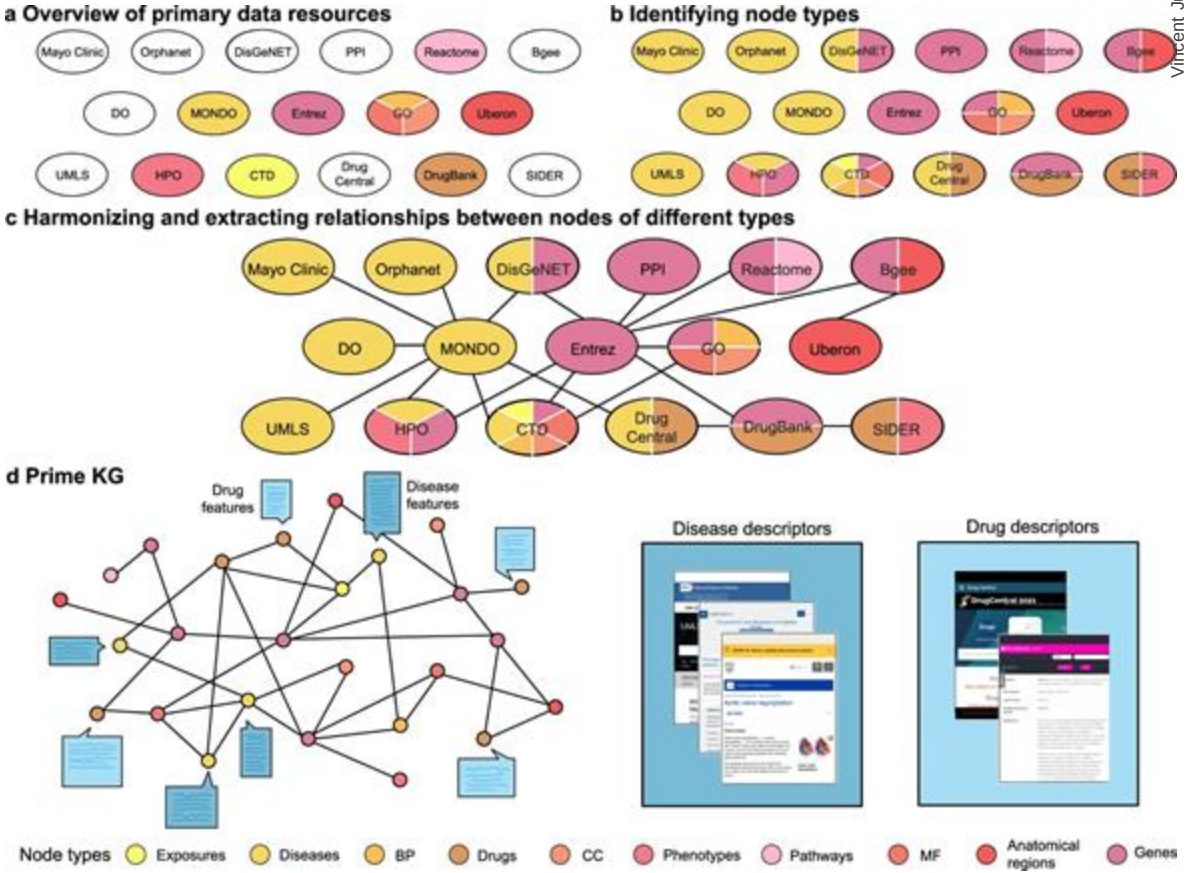


1. Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision medicine. *Sci Data* 10, 67 (2023). https://doi.org/10.1038/s41597-023-01960-3

Data : Knowledge graph

This work leverages PrimeKG[1] made by the same group.

MF: molecular function
BP: biological process
CC: cellular component PPI: protein-protein interactions
DO: disease ontology,
MONDO: MONDO disease ontology
Entrez: Entrez gene
GO: gene ontology
UMLS: unified medical language system
HPO: human phenotype ontology
CTD: comparative toxicogenomics database
SIDER: side effect resource.



1. Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision medicine. *Sci Data* 10, 67 (2023). https://doi.org/10.1038/s41597-023-01960-3

Problem definition

- Heterogeneous KG: $G=(V,E,T_R)$
  - nodes $i \in V$, edges $e_{i,j} = (i,r,j)$
  - $r \in T_R$, relationship type
  - each node $v \in T_V$, node type set

- Given a disease *i* and a drug *j*, we want to predict the likelihood of drug *j* being indicated and contraindicated for disease *i*

Vincent Ju ng

TxGNN Framework

- Heterogenous GNN encoder
- Disease similarity metric learning
- Pretraining followed by drug-disease centric, full-graph fine-tuning
- Graph explanation module to retain sparse set of Edges relevant for a given prediction

TxGNN: GNN Encoder

- TxGNN uses a RGCN[2] architecture, which updates node representations at each layer by multiplying the neighbors' previous representations using relationship-specific weights.

- Given a node embedding at layer l $\mathbf{h}_i^{(l)}$ for node *i* and its neighborhood with relations r $N_{i,r}$:
  - Message from neighbor: $\mathbf{m}_{r,i}^{(l)} = \sum_{j \in N_{i,r}} W_{r,M}^{(l)} \mathbf{h}_j^{(l-1)}$
  - Update node embedding**:** $\mathbf{h}_i^{(l)} = \mathbf{h}_i^{(l-1)} + \sum_{T_R} \mathbf{m}_{r,i}^{(l)}$

2. Schlichtkrull, Michael, et al. "Modeling relational data with graph convolutional networks." The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, proceedings 15. Springer International Publishing, 2018.

TxGNN: GNN Encoder

- Each drug-disease (*i*,*j*) pair is given the likelihood of a (contra)indication by the following equation:

$$p_{i,j,r} = \frac{1}{1 + \exp\left(-\text{sum}\left(\mathbf{h}_i \times \mathbf{w}_r \times \mathbf{h}_j\right)\right)}.$$

2. Schlichtkrull, Michael, et al. "Modeling relational data with graph convolutional networks." The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, proceedings 15. Springer International Publishing, 2018.

TxGNN: Training

Vincent Jung

- TxGNN is first pre-trained on predicting the presence of a relationship *r* between two entities *i* and *j* to which we assign the probability $p_{i,r,j}$. Positive pairs comprise all existing pairs with a connecting edge, negative pairs are sampled from non-connected pairs. The model maximizes $p_{i,r,j}$ for positive pairs and minimizes it for negative ones.

- It is then fine-tuned via the same training principle but only focusing on drug-disease pairs.

TxGNN: Disease distance metric learning

- In the KG, rare diseases have significantly less relevant nodes and edges -> low quality embeddings

- Their solution:
  - add an auxiliary embedding (different from the one learned by the GNN) which they call “disease signature vector”
  - aggregate it with original embedding
  - add gating mechanism to modulate between original and auxiliary embedding

TxGNN: Disease distance metric learning

- For disease *i*, signature vector is defined as:

$$\mathbf{p}_i = [\mathrm{p}_1 \cdots \mathrm{p}_{|\mathcal{V}_\mathrm{P}|} \mathrm{ep}_1 \cdots \mathrm{ep}_{|\mathcal{V}_\mathrm{EP}|} \mathrm{ex}_1 \cdots \mathrm{ex}_{|\mathcal{V}_\mathrm{EX}|} \mathrm{d}_1 \cdots \mathrm{d}_{|\mathcal{V}_\mathrm{D}|}]$$

where

$$\mathrm{p}_j = \begin{cases} 1 & if\ j \in \mathcal{N}_i^{\mathcal{P}} \\ 0 & \text{otherwise} \end{cases}, \mathrm{ep}_j = \begin{cases} 1 & if\ j \in \mathcal{N}_i^{\mathcal{EP}} \\ 0 & \text{otherwise} \end{cases}, \mathrm{ex}_j = \begin{cases} 1 & if\ j \in \mathcal{N}_i^{\mathcal{EX}} \\ 0 & \text{otherwise} \end{cases}, \mathrm{d}_j$$

$$= \begin{cases} 1 & if\ j \in \mathcal{N}_i^{\mathcal{D}} \\ 0 & \text{otherwise} \end{cases},$$

$\mathcal{N}_i^{\mathcal{P}}$: set of gene/protein

$\mathcal{N}_i^{\mathcal{EP}}$: set of effect/phenotype

$\mathcal{N}_i^{\mathcal{D}}$: set of disease node

$\mathcal{N}_i^{\mathcal{EX}}$: set of exposure

In the 1-hop neighborhood of node i

TxGNN: Disease distance metric learning

- Similarity sim(i,j) is defined as the dot product between p_i and p_j
- The top k most similar diseases are taken and their GNN embeddings are averaged, using the normalized similarity scores as weights :

$$\mathcal{D}_{\text{sim},i} = \text{argmax}_{j \in \mathcal{V}_{\mathcal{D}}} \text{sim}(i,j).$$

$$\mathbf{h}_i^{\text{sim}} = \sum_{j \in \mathcal{D}_{\text{sim}}} \frac{\text{sim}(i,j)}{\sum_{k \in \mathcal{D}_{\text{sim}}} \text{sim}(i,k)} \times \mathbf{h}_j.$$

TxGNN: Disease distance metric learning

- The final embedding is the weighted average between the GNN embedding and the similarity embeddings, with weights defined by a variable c dependent on the degree of node *i*

$$c_i = 0.7 \times \exp(-0.7 \times |\mathcal{N}_i^r|) + 0.2.$$

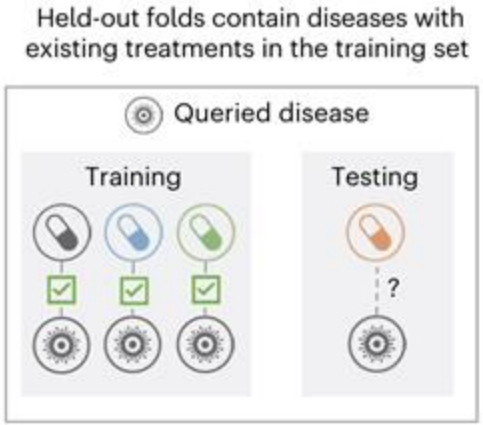
$$\widehat{\mathbf{h}}_i = c_i \times \mathbf{h}_i^{\text{sim}} + (1 - c_i) \times \mathbf{h}_i.$$

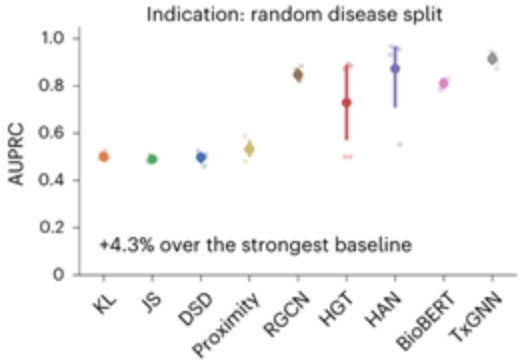
- The rationale is that a node with a higher degree has more information and thus is not required to rely on the similarity embedding as much.

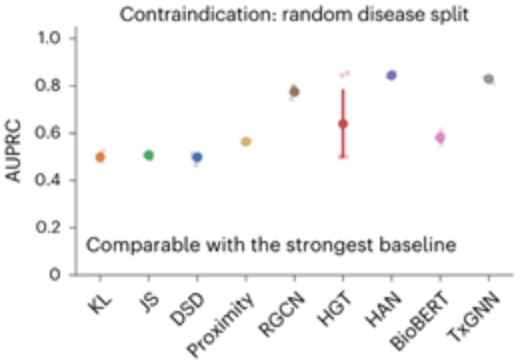
Evaluation : Random split

Vincent Ju ng

- Test set with already seen drug and disease



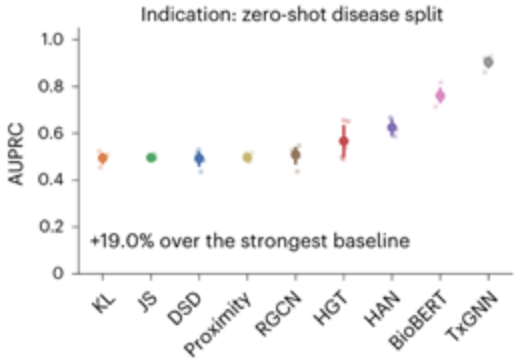


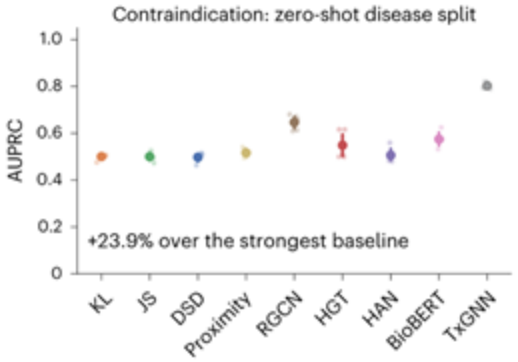


Evaluation : Zero-shot prediction

Vincent Ju ng

- Test set with disease with no known drug

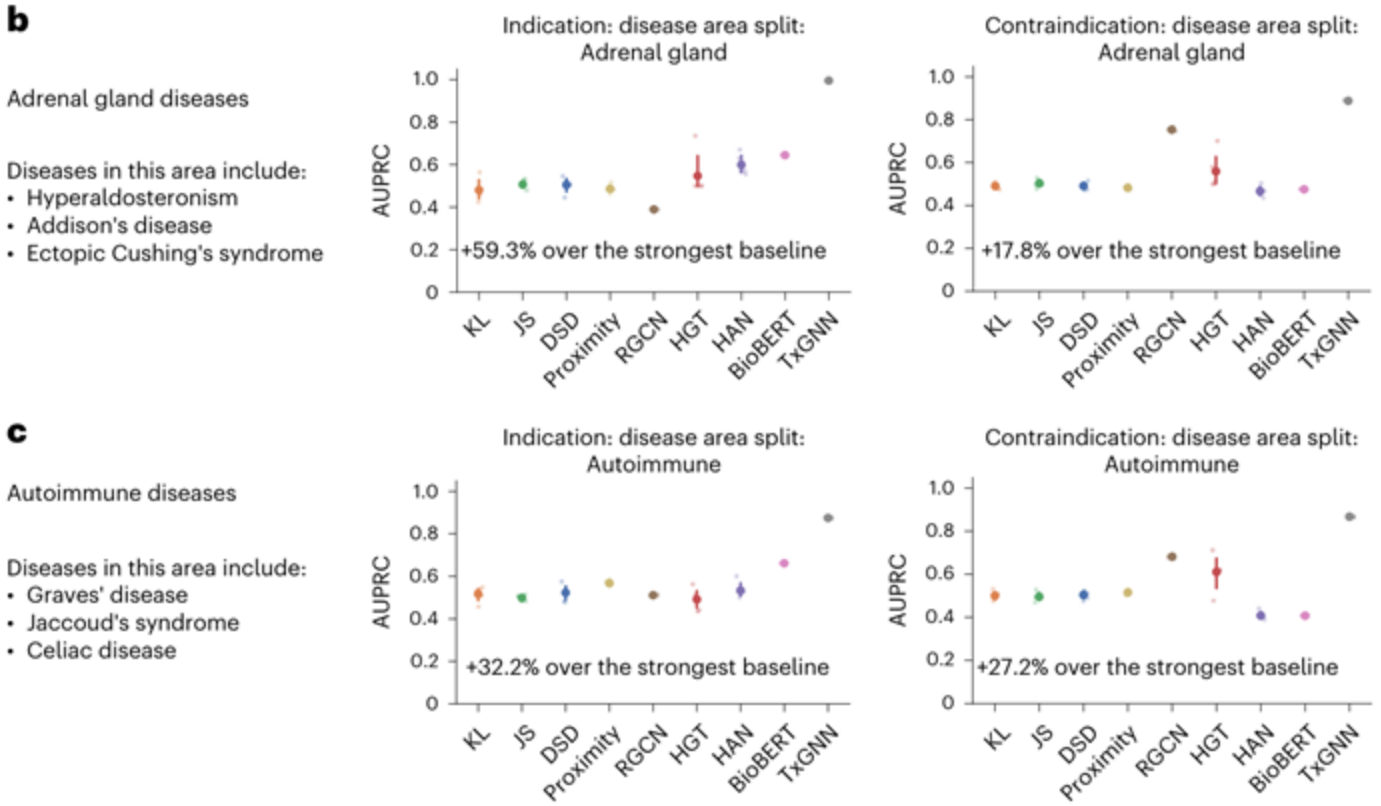




Evaluation : Held-out entire groups of diseases

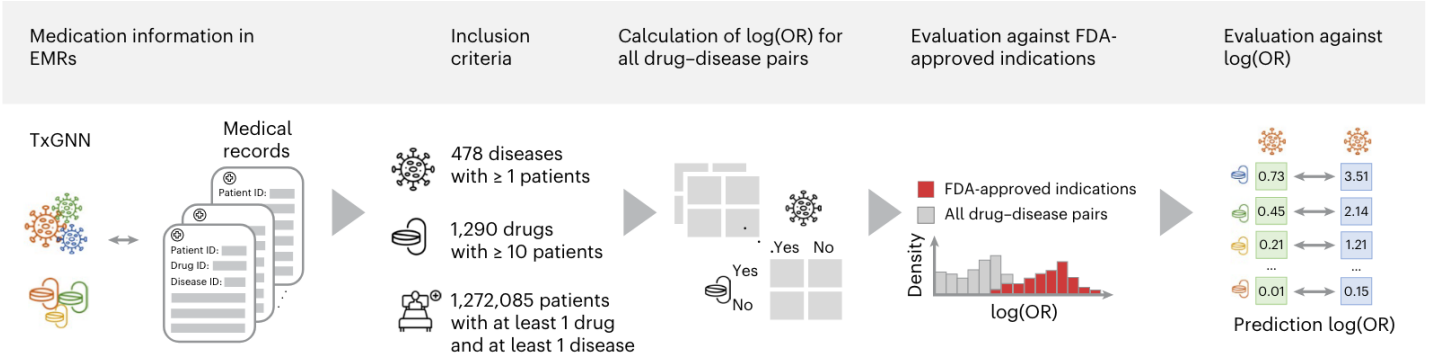
- Shortcut learning can happen: even if a disease does not have any associated drugs during training, if it has a very similar disease in the training set, the model can simply output the drugs for that disease
- Holding out entire disease groups to evaluate true(r) generalisation performance
- Disease groups considered:
  - Diabetes-related
  - Adrenal gland diseases
  - Autoimmune disease
  - Anemia
  - Neurodegenerative
  - Mental health disorders
  - Metabolic disorders
  - Cardiovascular diseases
  - Cancerous diseases

Evaluation : Held-out entire groups of diseases



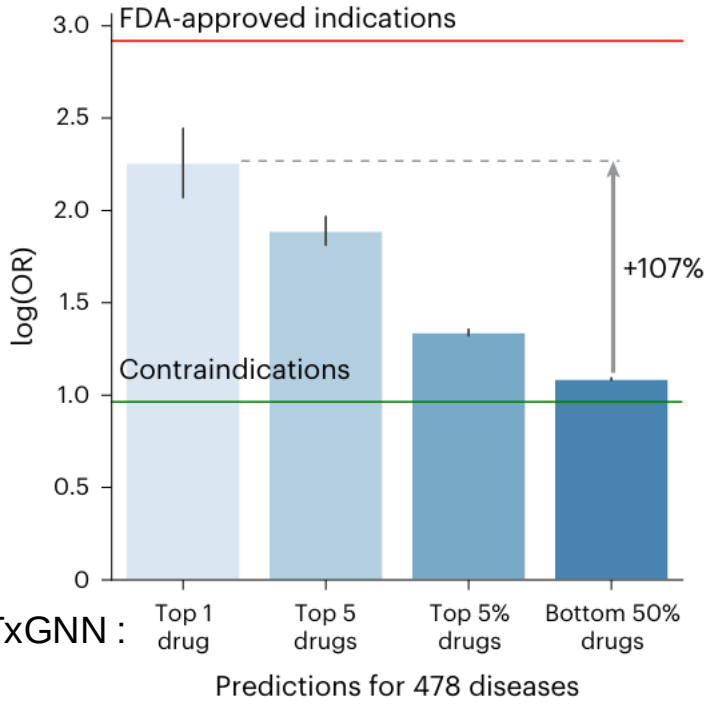
Evaluation : Are new predicted drug-disease combo relevant?

- The KG contains official drug-disease indication and contraindication
- Electronic medical records contain disease information and prescribed treatment, which contains off-label use information
- Log(odds_ratio) is calculated for drug-disease pairs and is evaluated against predicted drug-disease combos



Evaluation : Are new predicted drug-disease combo relevant?

-

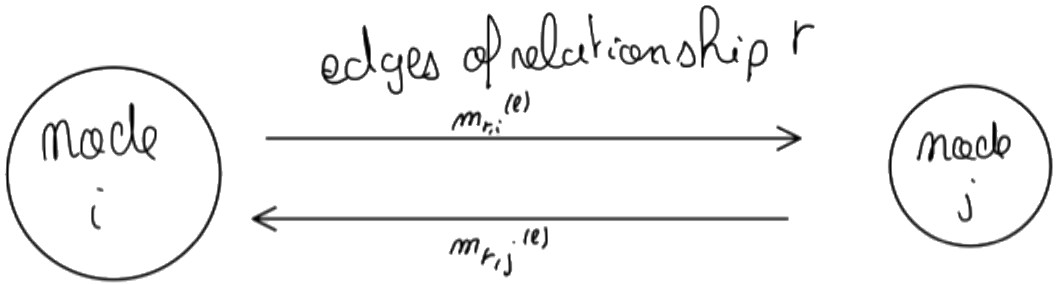


As ranked by TxGNN :

TxGNN: Interpretability

Vincent Ju ng

- Post-training edge dropout to find relevant subgraph for a prediction



Messages are gated:

$$z_{i,j,r}^{(l)} = \mathbb{I}_{\mathbb{R}>0.5}\left(\text{sigmoid}\left(W_{g,r}^{(l)}\left(\mathbf{m}_{r,i}^{(l)} \| \mathbf{m}_{r,j}^{(l)}\right)\right)\right)$$

TxGNN: Interpretability

- Gating mechanism is trained to minimize discrepancy in predicted probabilities and to maximize the number of opened gates.

$$\max_{\lambda} \min_{W_g} \sum_{k=1}^{L} \sum_{(i,r,j)\in\mathcal{D}_+\cup\mathcal{D}_-} \mathbb{I}_{[\mathbb{R}\neq 0]} z_{i,j,r}^{(k)} + \lambda(||\hat{p}_{i,j,r} - p_{i,j,r}||_2^2 - \beta),$$

- After this training, edges where z = 0 are dropped. We are left with a subgraph meant to explain TxGNN's predictions.

$$z_{i,j,r}^{(l)} = \mathbb{I}_{\mathbb{R}>0.5}\left(\underbrace{\text{sigmoid}\left(W_{g,r}^{(l)}\left(\mathbf{m}_{r,i}^{(l)}||\mathbf{m}_{r,j}^{(l)}\right)\right)}_{\text{Can be used to rank edges}}\right)$$

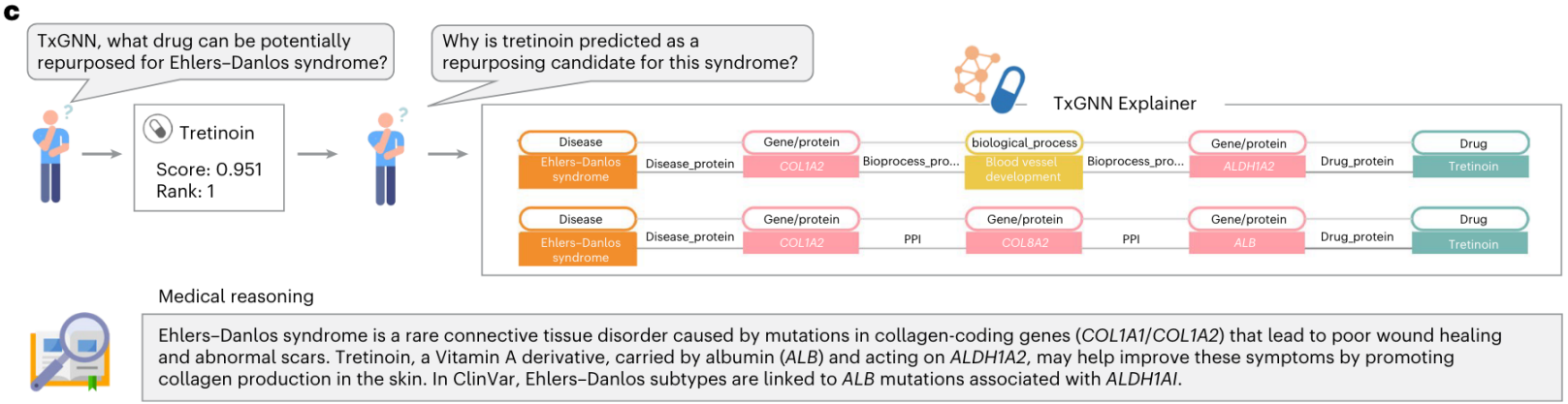
TxGNN: Interpretability

- Keeping those subgraphs instead of the whole KG minimally reduces performance (AUPRC=0.890 -> 0.886).
- Excluding edges deemed important (importance score > 0.5), performance drops significantly (AUPRC=0.890 -> 0.628)

TxGNN: Interpretability results

Vincent Jung

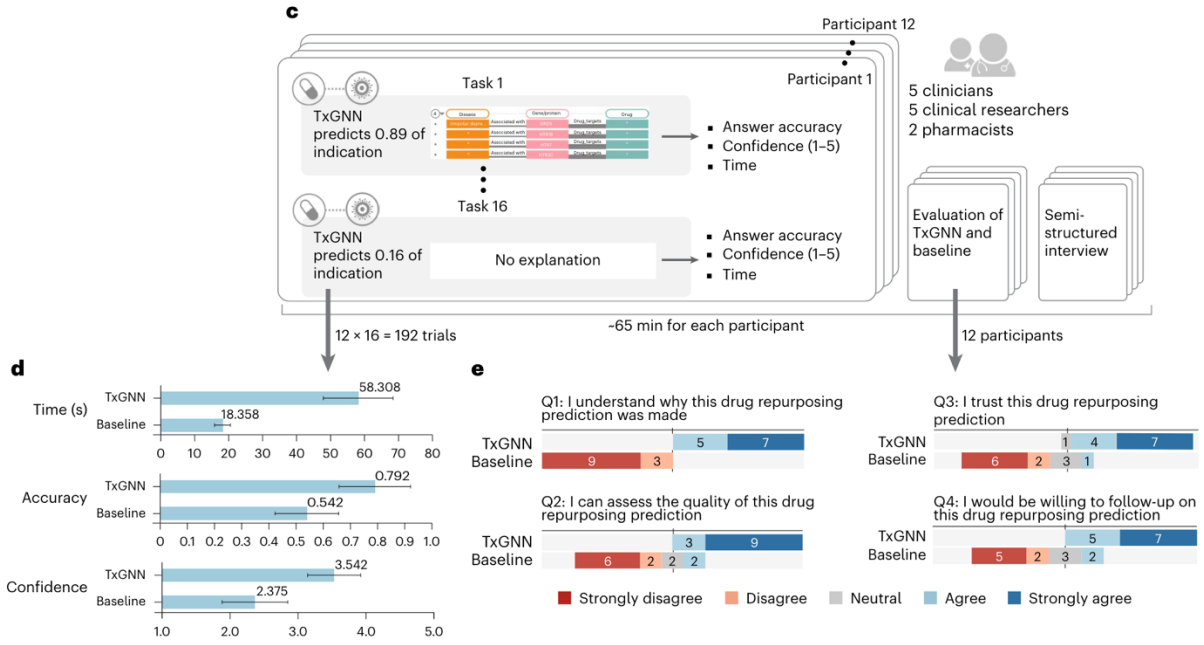
- Subgraphs for predicted drug-disease are medically relevant



TxGNN: Interpretability results

Vincent Ju ng

- Subgraphs for predicted drug-disease serve as good explanations to experts



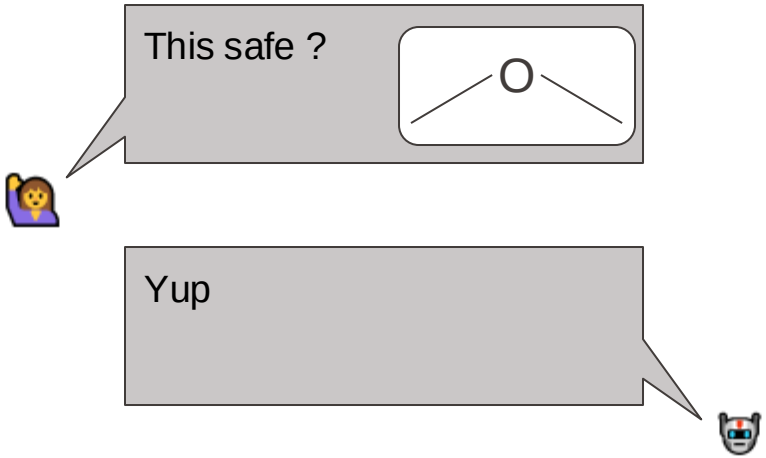
Conclusions

To summarize:

- TxGNN pre-trains on link prediction in a KG, then fine-tunes on predicting drug-disease relationships
- Disease metric learning improves performance for rare diseases
- It beats many other models, especially when tested on unseen disease groups
- TxGNN explainer gives relevant subgraphs for predictions

My opinions:

- Framework could easily be expanded to other uses (PPI, disease understanding)
- Interpretability method was a good showcase for GNNs



GIMLET: A Unified Graph-Text Model for Instruction-Based Molecule Zero-Shot Learning

Haiteng Zhao, Shengchao Liu, Chang Ma, Hannan Xu, Jie Fu, Zhi-Hong Deng, Lingpeng Kong, Qi Liu

Goal: Zero-Shot molecular property prediction

- Molecular datasets are limited because experiments can be expensive, thus supervised setting is not desirable, especially for tasks with very small labeled datasets

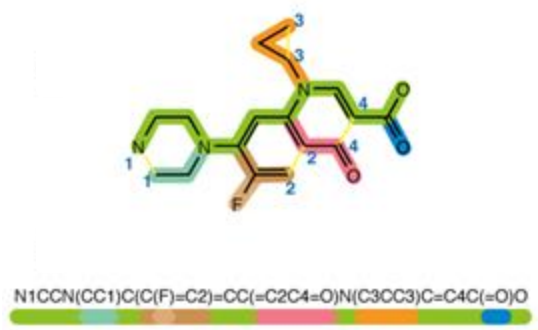
- Additional information provided in text form often not taken into account.

-> Embed molecule and text together, can prompt specific tasks

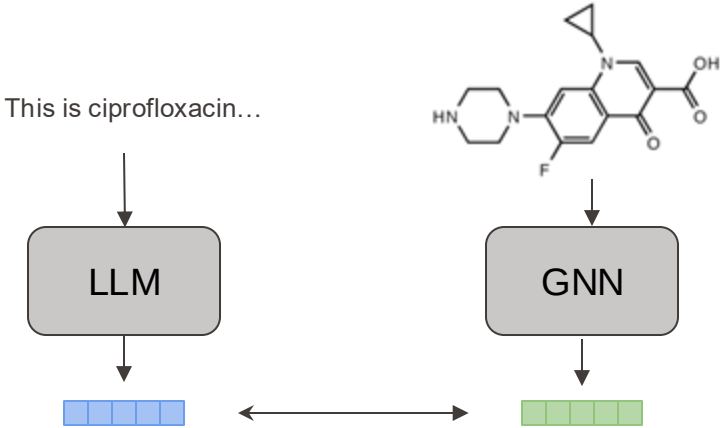
Vincent Jung

How: Zero-Shot molecular property prediction

- Previous methods generally used either SMILES representation or a GNN to embed molecular graph.



SMILES representation



GIMLET: Unified Graph-Text Transformer

- GIMLET uses the full graph and text as input to one single transformer-based model.

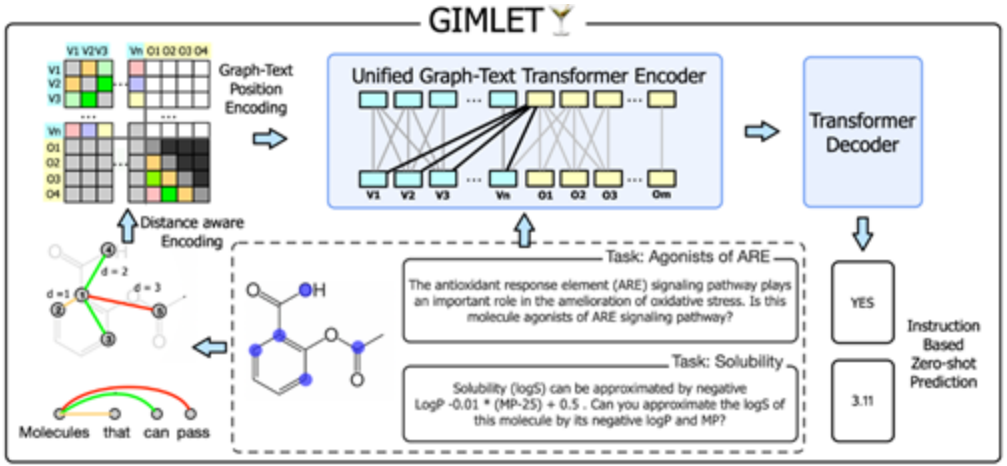
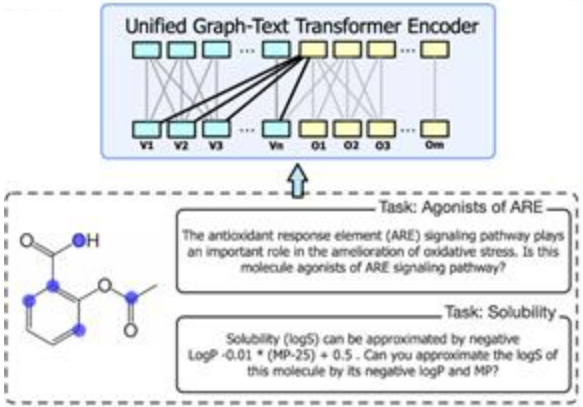


Figure 1: Our framework handles molecule tasks in the zero-shot fashion by natural language instruction. Within GIMLET, we employ distance-based joint position embedding to encode graphs and instruction texts. Additionally, we utilize attention masks to decouple the graph encoding process.

GIMLET: Unified Graph-Text Transformer

- Given a graph *G* with n nodes and a text input *T* with m tokens, graph nodes and text tokens are represented as tokens. This results in hidden state:

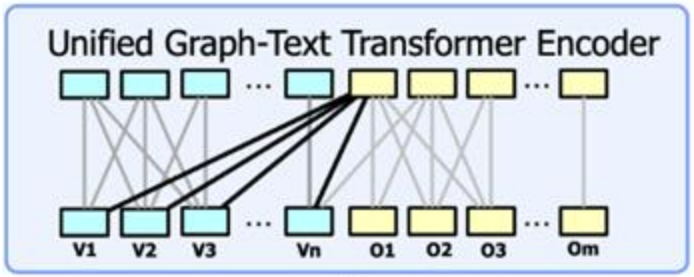
$$H = [h_1,\ldots,h_n,h_{n+1},\ldots,h_{n+m}]$$



GIMLET: Unified Graph-Text Transformer

Vincent Ju ng

- Attention is modified to let text tokens attend to graph tokens, but graph tokens can only attend to other graph tokens



GIMLET: Unified Graph-Text Transformer

- Token embeddings: $H = [h_1,\ldots,h_n,h_{n+1},\ldots,h_{n+m}]$
- Attention coefficient between two tokens:

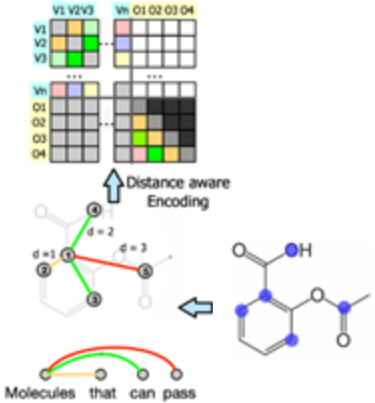
$$\hat{A}_{ij} = \underbrace{\frac{\left(h_i W^Q\right)\left(h_j W^K\right)^T}{\sqrt{d_k}}}_{\text{Scaled Dot-Product self-attention (unmodified)}} + \underbrace{b\left(i,j\right)}_{\textbf{Bias (modified)}}$$

GIMLET: Unified Graph-Text Transformer

- Bias: $b(i,j) = b^D_{\text{POS}(i,j)} + b^M_{i,j} + \underset{k \in \text{SP}(i,j)}{\text{Mean}} b^E_{e_k}$,

- $b^D_{\text{POS}(i,j)}$:

$$\begin{cases} i-j & \text{if } n+1 \le i,j \le n+m \\ \text{GRAPH SHORTEST DISTANCE}(i,j) & \text{if } 1 \le i,j \le n \\ <\text{CROSS}> & \text{otherwise} \end{cases},$$



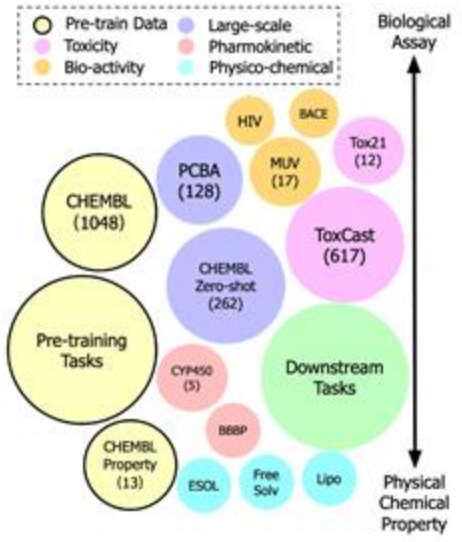
GIMLET: Unified Graph-Text Transformer

- Bias: $b(i,j) = b^D_{\text{POS}(i,j)} + b^M_{i,j} + \underset{k \in \text{SP}(i,j)}{\text{Mean}} b^E_{e_k}$,
- $b^M_{i,j}$: $-\infty$ if $i \leq n$ and $j > n$ otherwise 0
- $\underset{k \in \text{SP}(i,j)}{\text{Mean}} b^E_{e_k}$: Mean pooling of edge features of Shortest path between i and j

Vincent Jung

Data: Paired Graph and text

Vincent Jung



Heavy atoms counting (From Chembl property)

"Heavy atom refers to any atom that is not hydrogen. How many heavy atoms do the molecule have?"

Inhibitors of Schistosoma Mansoni Peroxiredoxins (From Chembl)

"The functional assay is named qHTS Assay for the Inhibitors of Schistosoma Mansoni Peroxiredoxins. It is related to two other pubchem assays, namely Confirmation Concentration-Response Assay for Inhibitors of the Schistosoma mansoni Redox Cascade and Schistosoma Mansoni Peroxiredoxins (Prx2) and thioredoxin glutathione reductase (TGR) coupled assay. The assay category is also confirmatory and it pertains to the Schistosoma mansoni organism. Is this molecule effective to the assay?"

Inhibition of receptor SF-1 (From MUV)

"The nuclear receptor SF-1 (steroidogenic factor-1) is expressed in the pituitary, testes, ovaries, and adrenal gland and regulates steroid hormone production at many levels, including direct regulation of expression of major P450 enzymes involved in steroid hormone synthesis. Is this molecule inhibitor of SF-1?"

Toxicity to ARE signaling pathway (From Tox21)

"Oxidative stress has been implicated in the pathogenesis of a variety of diseases ranging from cancer to neurodegeneration. The antioxidant response element (ARE) signaling pathway is important in the amelioration of oxidative stress. Is this molecule agonists of antioxidant response element (ARE) signaling pathway?"

Figure 2: (Left) Illustration of datasets. Circle size corresponds to task number. Tasks are organized by category. Tasks on the top are more related to biological assay, on the bottom need more chemical and physical properties. GIMLET is trained on pretraining tasks, then tested on downstream tasks in the zero-shot setting. (Right) Our task instructions contain task explanations and questions.

Data: Paired Graph and text Pre-training dataset

Chembl

"The assay is PUBCHEM_BIOASSAY: qHTS Assay for Activators of Human Muscle isoform 2 Pyruvate Kinase. (Class of assay: confirmatory) , and it is Direct single protein target assigned . The assay has properties: assay category is confirmatory ; assay organism is Homo sapiens ; assay type description is Functional . Is the molecule effective to this assay?"

Chembl property

"The partition coefficient, abbreviated P, is defined as a particular ratio of the concentrations of a solute between the two solvents (a biphase of liquid phases), specifically for un-ionized solutes, and the logarithm of the ratio is thus Log P. When one of the solvents is water and the other is a non-polar solvent, then the log P value is a measure of lipophilicity or hydrophobicity. The defined precedent is for the lipophilic and hydrophilic phase types to always be in the numerator and denominator respectively. What is the logarithm of the partition coefficient of this molecule?"

Data: Paired Graph and text Downstream tasks, classification

Some labeled datasets are transformed into instruction prompts.

BACE

"BACE1 is an aspartic-acid protease important in the pathogenesis of Alzheimer's disease, and in the formation of myelin sheaths. BACE1 is a member of family of aspartic proteases. Same as other aspartic proteases, BACE1 is a bilobal enzyme, each lobe contributing a catalytic Asp residue, with an extended active site cleft localized between the two lobes of the molecule. The assay tests whether the molecule can bind to the BACE1 protein. Is this molecule effective to the assay?"

HIV

"Human immunodeficiency viruses (HIV) are a type of retrovirus, which induces acquired immune deficiency syndrome (AIDs). Now there are six main classes of antiretroviral drugs for treating AIDs patients approved by FDA, which are the nucleoside reverse transcriptase inhibitors (NRTIs), the non-nucleoside reverse transcriptase inhibitors (NNRTIs), the protease inhibitors, the integrase inhibitor, the fusion inhibitor, and the chemokine receptor CCR5 antagonist. Is this molecule effective to this assay?"

Data: Paired Graph and text Downstream tasks, regression

ESOL

```
"Solubility (logS) can be approximated by negative LogP -0.01
* (MPt \u2013 25) + 0.5 . Can you approximate the logS of this
molecule by its negative logP and MPt?"
```

FreeSolv

```
"The free energy of hydration can be approximated by
\u0394G_hyd = \u0394G_solv,soln - \u0394G_solv,gas + RT ln
(10^(-pKa)). Can you tell me the free energy of hydration (by
using the negative pka) of this molecule, predicted by using
\u0394G_solv and negative pka?"
```

Results: Better than other ZS methods

Table 1: Zero-shot performance (ROC-AUC) over Bio-activity, Toxicity, and Pharmacokinetic tasks.

Method	#Param	Type	bace	hiv	muv	Avg. bio	tox21	toxcast	Avg. tox	bbbp	cyp450	Avg. pha
KVPLM	110M		0.5126	0.6120	0.6172	0.5806	0.4917	0.5096	0.5007	0.6020	0.5922	0.5971
MoMu	113M		0.6656	0.5026	0.6051	0.5911	0.5757	0.5238	0.5498	0.4981	0.5798	0.5390
Galactica-125M	125M	Zero Shot	0.4451	0.3671	0.4986	0.4369	0.4964	0.5106	0.5035	**0.6052**	0.5369	0.5711
Galactica-1.3B	1.3B		0.5648	0.3385	0.5715	0.4916	0.4946	0.5123	0.5035	0.5394	0.4686	0.5040
GIMLET (Ours)	64M		**0.6957**	**0.6624**	**0.6439**	**0.6673**	**0.6119**	**0.5904**	**0.6011**	0.5939	**0.7125**	**0.6532**
GCN	0.5M		*0.736*	*0.757*	*0.732*	*0.742*	*0.749*	*0.633*	*0.691*	*0.649*	0.8041	0.7266
GAT	1.0M		*0.697*	*0.729*	*0.666*	*0.697*	*0.754*	*0.646*	*0.700*	*0.662*	0.8281	0.7451
GIN	1.8M	Supervised	*0.701*	*0.753*	*0.718*	*0.724*	*0.740*	*0.634*	*0.687*	*0.658*	0.8205	0.7392
Graphormer	48M		0.7760	0.7452	0.7061	0.7424	0.7589	0.6470	0.7029	0.7015	0.8436	0.7725
Graphormer-p	48M		0.8575	0.7788	0.7480	0.7948	0.7729	0.6649	0.7189	0.7163	0.8877	0.8020

Table 2: Zero-shot performance (ROC-AUC) over large scale molecule tasks.

Method	Chembl Zero-Shot	PCBA
KVPLM	0.4155	0.4811
MoMu	0.5002	0.5150
Galactica-125M	0.6461	0.4800
Galactica-1.3B	0.4818	0.5202
GIMLET (Ours)	**0.7860**	**0.6211**

Table 3: Zero-Shot performance (RMSE) on Physical-chemical datasets.

Method	Type	ESOL	Lipophilicity	FreeSolv	Avg. phy
KVPLM		-	-	-	-
MoMu	Zero Shot	-	-	-	-
GIMLET (Ours)		1.132	1.345	5.103	2.527
GCN		1.331	0.760	2.119	1.403
GAT		1.253	0.770	2.493	1.505
GIN	Supervised	1.243	0.781	2.871	1.632
Graphormer		0.901	0.740	2.210	1.284
Graphormer-p		0.804	0.675	1.850	1.110

Results: Effectivness of framework

Table 4: Ablation study on GIMLET module.

Method	bace	hiv	muv	Avg. bio	tox21	toxcast	Avg. tox	bbbp	cyp450	Avg. pha
w.o. unifying	0.4319	0.6133	0.6067	0.5506	0.5922	0.5537	0.5730	0.5309	0.6206	0.5758
w.o. decoupling	0.6458	0.6406	0.5421	0.6095	**0.6306**	**0.5954**	**0.6130**	0.5666	0.6320	0.5993
GIMLET	**0.6957**	**0.6624**	**0.6439**	**0.6673**	0.6119	0.5904	0.6011	**0.5939**	**0.7125**	**0.6532**

Results: Few-Shot fine-tuning

Examples are given to the model in the prompt, and it is fine-tuned.

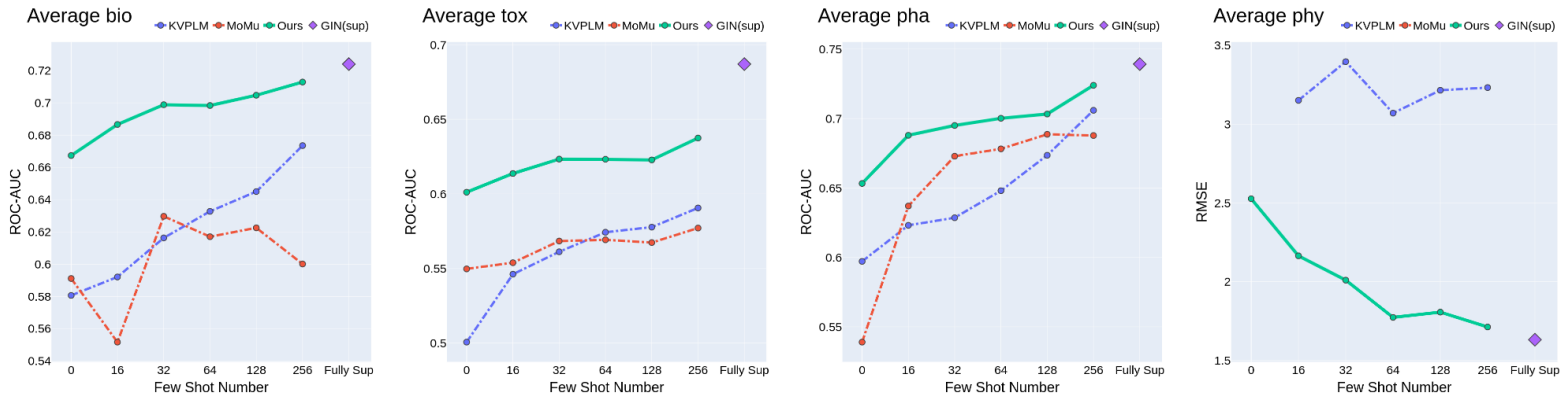


Figure 3: Few shot performance. Higher is better for bio, tox, and pha, and lower is better for phy.

Conclusions

To summarize:

- GIMLET uses a unified graph-text module.
- Attention is decoupled and masked
- ZS performance is better than other methods

My opinions:

- Few-Shot fine-tuning is odd to me
- Wonder how perofrmance would be with newer LLMs
- Agentic behavior ? CoT