Graph Neural Networks For Biological Knowledge Discovery

Antoine Toffano, Jérôme Azé, Pierre Larmande

Graph Neural Networks for Biological Knowledge Discovery

Antoine Toffano[1], Jérome Azé[1], Pierre Larmande[2]

[1]LIRMM, Univ. Montpellier, CNRS, Montpellier, France.
[2]DIADE, Univ. Montpellier, IRD, CIRAD, Montpellier, France.

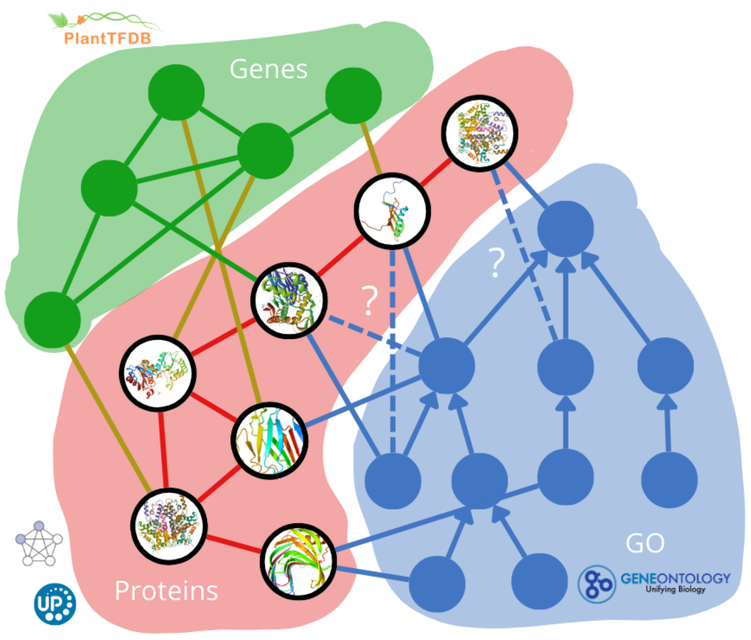
Introduction

Biological data is both abundant and heterogeneous. Consequently, it benefits from non-Euclidean structures like graphs for representation.

This study examines how Graph Neural Networks (GNNs) perform in biological knowledge discovery compared to traditional geometric models.

Knowledge Discovery

We use *Oryza Sativa* data, whose graph contains a total of **2 068 651 edges** and **264 291 nodes**, of which 149 799 are proteins, 54 954 are genes and 48 091 are GO terms.

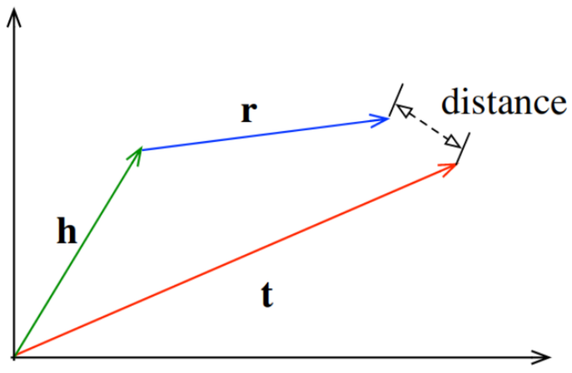


Knowledge discovery can be represented as a link prediction task between nodes in the graph. We focus on predicting new links between protein nodes and GO terms.

Geometric Models

In geometric models, data is represented as vectors in a latent space, with links as geometric transformations.

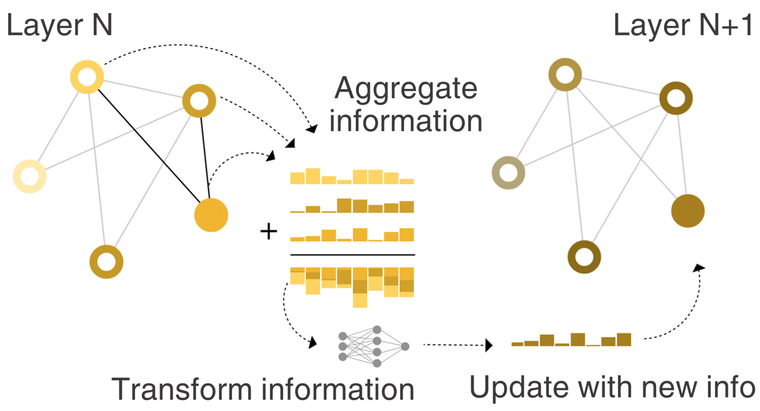
For instance, TransE[1] is trained to minimize the distance between the starting node vector h plus the relation vector r and the end node vector t (i.e., $h + r = t$).



• Figure from [2]

Graph Neural Networks

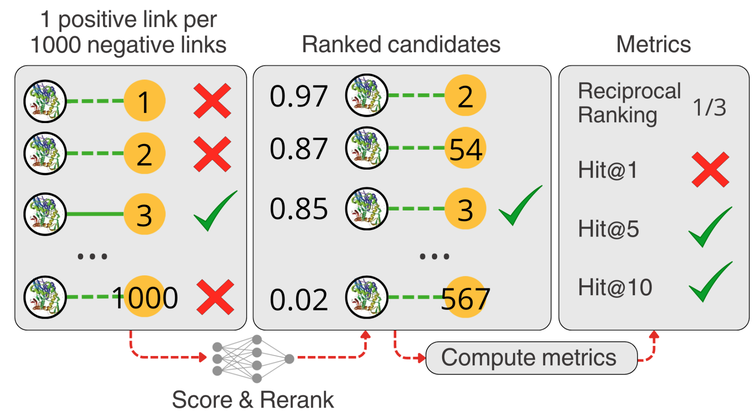
GNNs iteratively update node states by aggregating messages from neighbors through a learnable function, propagating information across the graph.



• Figure adapted from [3]

Evaluation Process

We rank the likelihood of true protein-GO term links against 1000 random negative GO terms.



Better performance is indicated by higher ranks for true links.

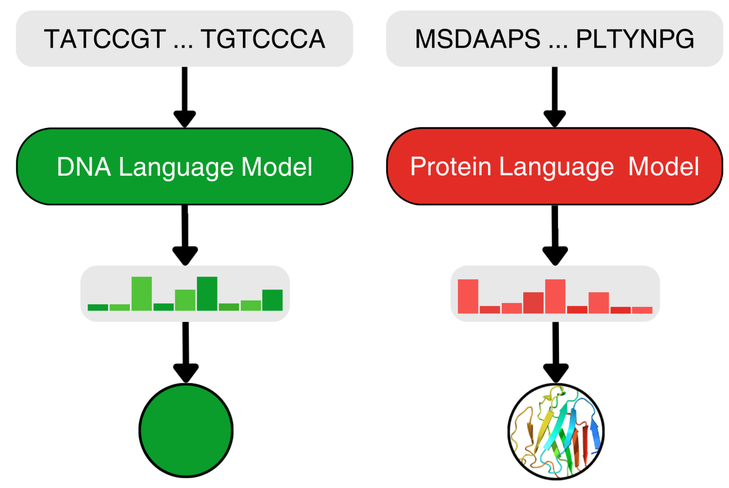
Results

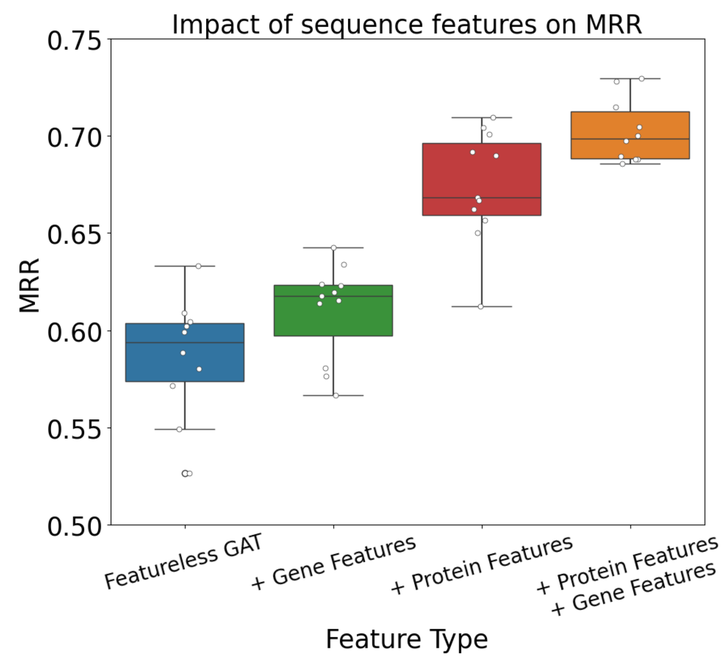
Model	MRR	Hit@1	Hit@5	Hit@10	Mem. (VRAM)
GATv2Conv	**0.58**	0.43	**0.78**	**0.89**	10.0 Gb
GeneralConv	**0.58**	0.42	**0.79**	**0.89**	14.4 Gb
TransformerConv	**0.58**	0.43	0.78	**0.89**	20.1 Gb
MFConv	**0.58**	**0.45**	0.74	0.83	3.6 Gb
ResGatedGraphConv	0.57	0.42	0.74	0.84	3.7 Gb
GraphConv	0.54	0.40	0.74	0.81	2.9 Gb
SAGEConv	0.49	0.36	0.64	0.76	5.4 Gb
LEConv	0.47	0.33	0.64	0.77	3.0 Gb
DistMult	**0.78**	0.69	**0.89**	**0.93**	29.6 Gb
ComplEx	**0.78**	**0.70**	0.87	0.90	29.4 Gb
TransE	0.46	0.37	0.4	0.64	31.2 Gb

Table 1: Link Prediction capabilities of different models.
Top: GNN models, Bottom: Geometric models

LLMs as encoders

Proteins and genes can be embedded using large language models trained on their sequences and used as GNN node features.





Protein embeddings were derived by averaging amino-acid representations from esm2-650M[4].
Gene embeddings were obtained by averaging 6-mer representations from agront-1B[5].

Conclusions & Perspectives

- Geometric models outperform GNNs, but require more memory for training.
- LLMs biological sequence embeddings enhances GNN performance.
- GNNs have potential for knowledge discovery in biological data.
- Using edge features alongside node features could improve GNN performance.

Contact information:
Antoine Toffano
antoine.toffano@lirmm.fr

References

[1] Antoine Bordes, Nicolas Usunier, Alberto GarcÃa-DurÃ¡n, J. Weston, and Oksana Yakhnenko. Translating Embeddings for Modeling Multi-relational Data. 2013.

[2] Shaoxiong Ji, Shirui Pan, E. Cambria, Pekka Marttinen, and Philip S. Yu. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33:494–514, 2020.

[3] Benjamin Sanchez-Lengeling, Emily Reif, Adam Pearce, and Alexander B. Wiltschko. A gentle introduction to graph neural networks. *Distill*, 2021. doi: 10.23915/distill.00033. https://distill.pub/2021/gnn-intro.

[4] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

[5] Javier Mendoza-Revilla, Evan Trop, Liam Gonzalez, Masa Roller, Hugo Dalla-Torre, Bernardo P de Almeida, Guillaume Richard, Jonathan Caton, Nicolas Lopez Carranza, Marcin Skwark, et al. A foundational large language model for edible plant genomes. *bioRxiv*, pages 2023–10, 2023.

LIRMM
Institut de Recherche pour le Développement
FRANCE