Unveiling Protein Connections through Graph Neural Networks and ProtT5 Embeddings

Graduand: Lorenzo Marinelli - 2043092
Supervisor: Paola Paci

Applied Computer Science and Artificial Intelligence
Sapienza University of Rome

Table of contents

1 | Introduction, Background & Related Work

2 | Pre-processing

3 | Methods

4 | Results

5 | NOTCH-2 Case study

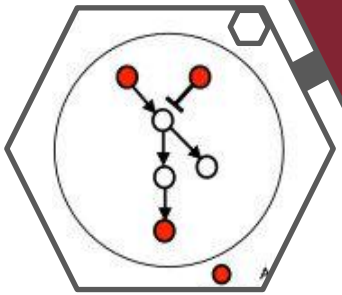
6 | Conclusion and Future Work

1)Biological Importance of PPIs

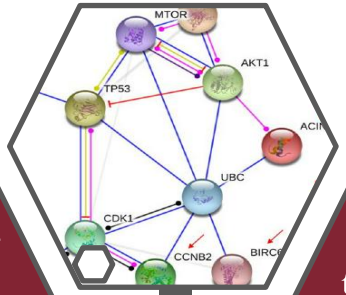
Proteins Operate in Networks, Not Isolation

- Proteins form dynamic complexes; interactions enable new catalytic or signaling functions.
- Millions of distinct PPIs exist in a typical human cell.
- Interaction networks translate genetic information into biological function.

PPIs, Health, and Disease

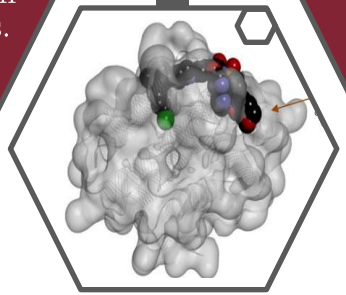


Disrupted PPIs propagate dysfunction through pathways

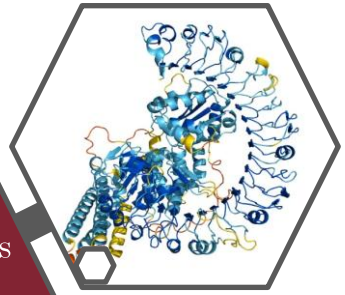


Certain diseases arise from broken interaction links.

PPIs as drug targets: new therapies aim to block or enhance specific protein interactions



Synthetic biology leverages predicted PPIs to engineer new cellular functions

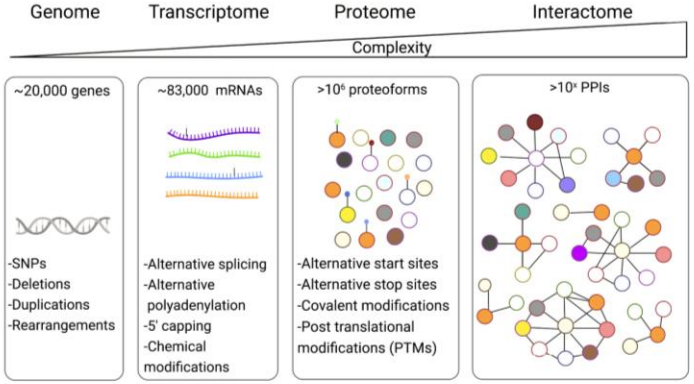


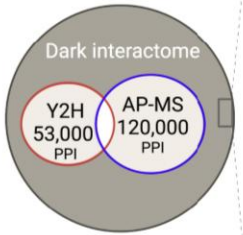
Click on small hexagons on images for image sources

The Challenge – Mapping the Interactome

Many Interactions, Limited by Experiments

- Millions of potential PPIs; large portions of the human interactome remain uncharted.
- Experimental screens (yeast two-hybrid, etc.) are resource-intensive and can miss condition-specific interactions.
- Different cell contexts -> different subsets of interactions (difficult to capture exhaustively in the lab).



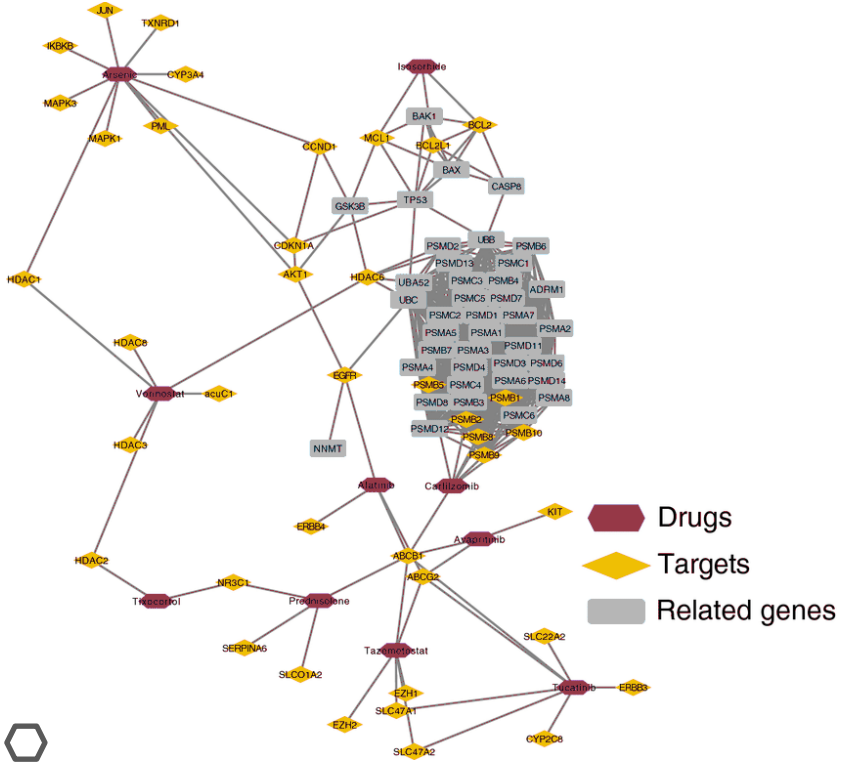


Context	Example
-Organelle-specific	-Transmembrane proteins can be insoluble during cell lysis
-Transient PPIs	-Enzyme-substrate interactions
-Cell-type-specific	-Oligodendrocytes vs dopaminergic neurons
-Development/ differentiation	-OCT4 during embryonic development

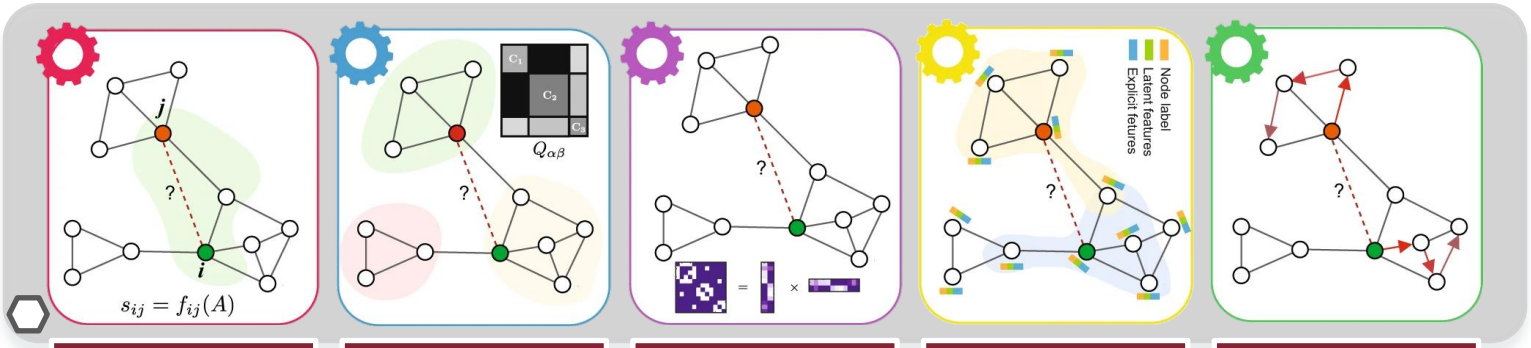
Role of Computational PPI Prediction

Why Computational Predictions?

- Generate hypotheses: propose new interaction partners to test in the lab (saves time/resources).
- Network analysis: use predicted edges for clustering, centrality, module detection, to reveal hidden functional modules.
- Drug and treatment discovery: identify novel protein-protein interfaces for therapeutic targeting.



Existing Approaches to PPI Prediction

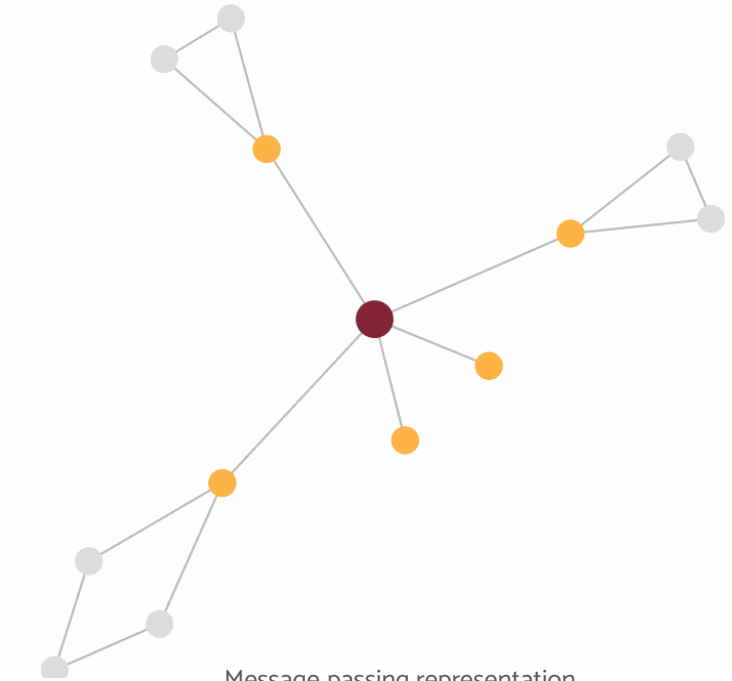


Topology-based scores (network proximity): e.g. Common neighbors, Jaccard index.	Probabilistic graphical models: find community-driven interaction likelihood.	Matrix/Tensor factorization: low-rank embeddings of adjacency matrix.	Graph-Neural-Networks (deep learning): learn from sequence + network context.	Diffusion methods: propagate signals through network (e.g. random walk).

Graph Neural Networks in Bioinformatics

Why Use Graph Neural Networks for PPIs?

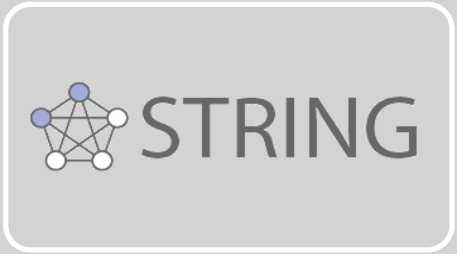
- Naturally models the PPI network (nodes = proteins, edges = interactions).
- Message passing: aggregates information from a protein's neighbours (learns context).
- Combines protein biological features with network structure for improved predictions.
- Key tasks: Node classification, graph classification, link prediction, generative modelling.



Message passing representation

Public PPI data sets

STRING



Integrates diverse evidence into a single confidence score (0–1000).

Release 12: >20 billion interactions across ~15 000 organisms.

Continuous scores enable easy thresholding, class rebalancing, and weighted losses in ML workflows.

BioGRID

Entirely manually curated from primary literature.

Records experimental method, throughput category, and PubMed ID for each interaction.

Yields high-precision but relatively sparse networks.

IntAct

Adopts the PSI-MI standard to provide rich metadata (host organism, tag, affinity matrix, detection instrument...).

Part of the IMEx consortium, sharing curation to reduce redundancy and boost coverage.

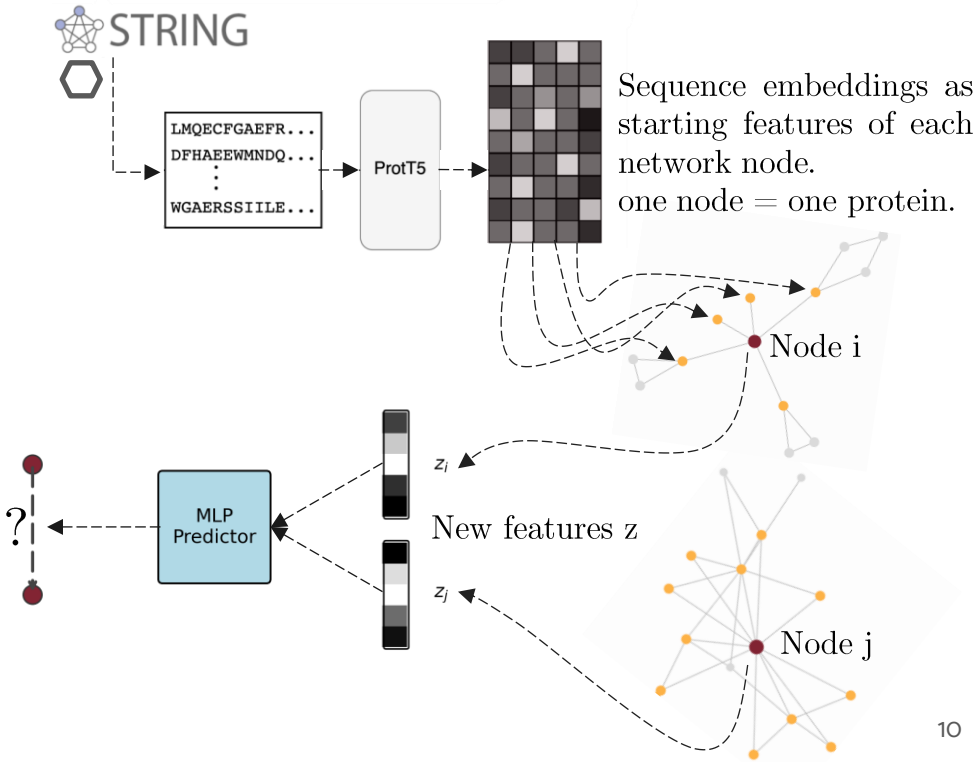
GNN + ProtT5 Embeddings

Proposed Solution Overview

1. Protein Sequences -> ProtT5 Embeddings (1024-dimensional feature per protein)

2. Graph Neural Network: message passing on the PPI network

3. Link Predictor: outputs probability of interaction for any pair

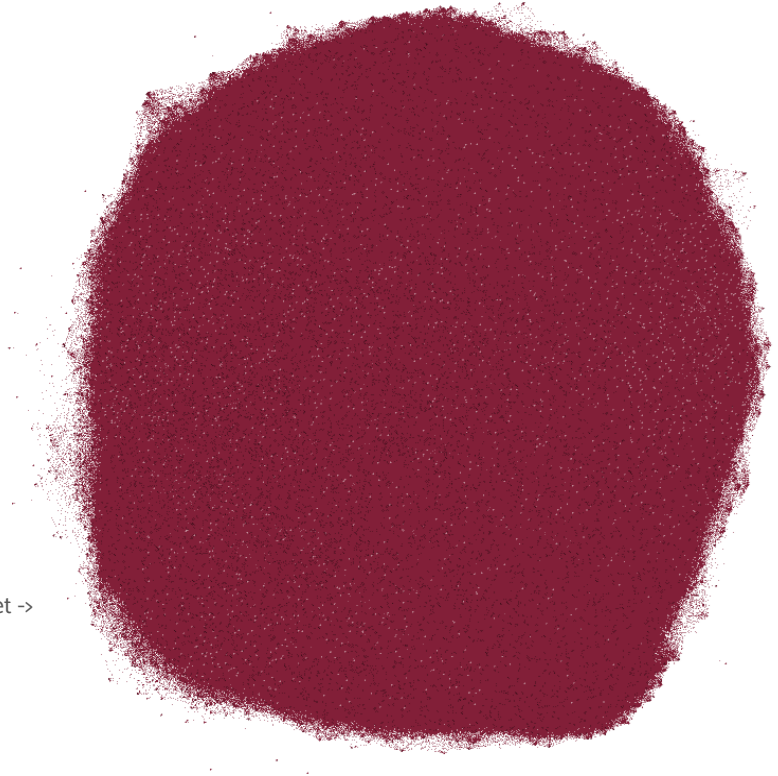


Data – STRING v12 (human) Database

Dataset overview

- Source: STRING v12, curated human PPI network.
- Included edges: confidence ≥ 950 (very high reliability, Removed lower-confidence edges to avoid noise in training).
- Resulting graph: 10,430 proteins, ~0.12 million interactions.

Full vs filtered dataset ->



Training/Validation/Test Split

Edge partitions & label tensors used by the GNN

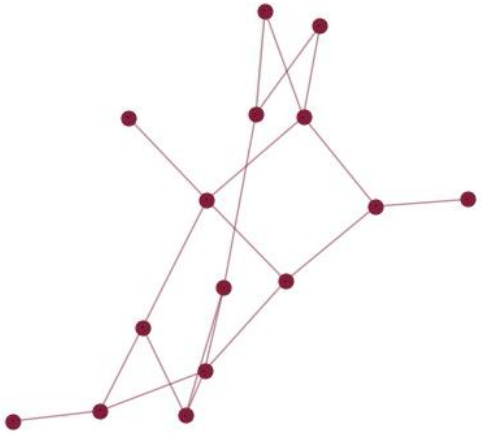
	MP edges	Pos. label edges	Neg. label edges
Training ($\mathcal{E}^{\text{train}}$)	93 698	46 849	on–the–fly at 1:10 ratio
Validation ($\mathcal{Q}^{\text{val}}$)	93 698	5 855	58 550 (fixed, 1:10)
Test ($\mathcal{Q}^{\text{test}}$)	105 408	5 855	58 550 (fixed, 1:10)

Overview of the Dataset split

Negative Sampling Strategy

Handling Class Imbalance

- Only ~0.12M positive edges vs. ~50+ million possible pairs, very sparse network.
- For each positive edge, sample 10 random non-interacting pairs as negatives.
- No loss up-weighting for positives (treated negatives at 10× frequency) in the binary cross entropy loss, bias model towards lower recall and higher precision.



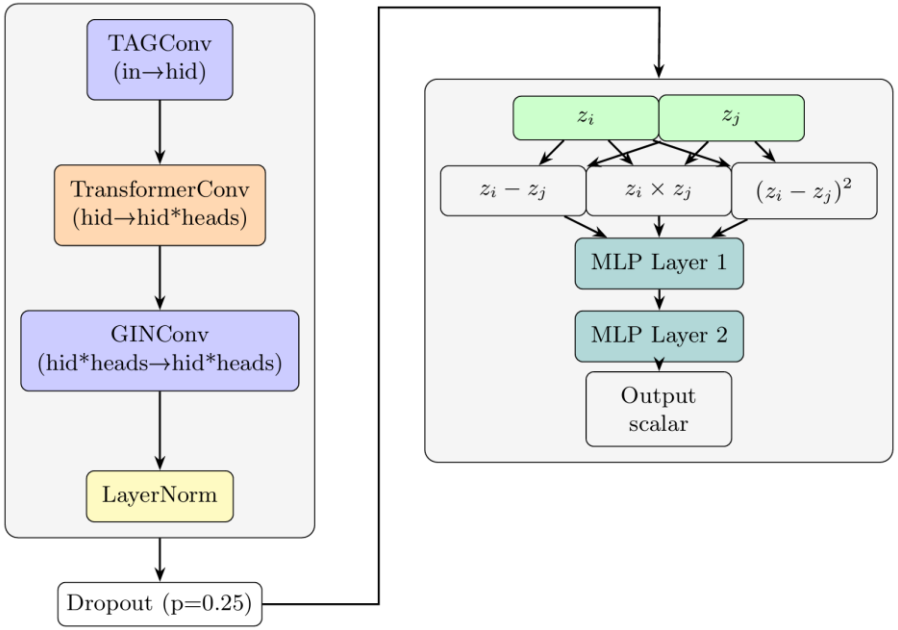
MODEL

GNN Architecture Overview

Encoder–Decoder Model

- Encoder GNN:
  inputs = ProtT5 vectors + graph edges;
  output = new embedding z_n for each node.

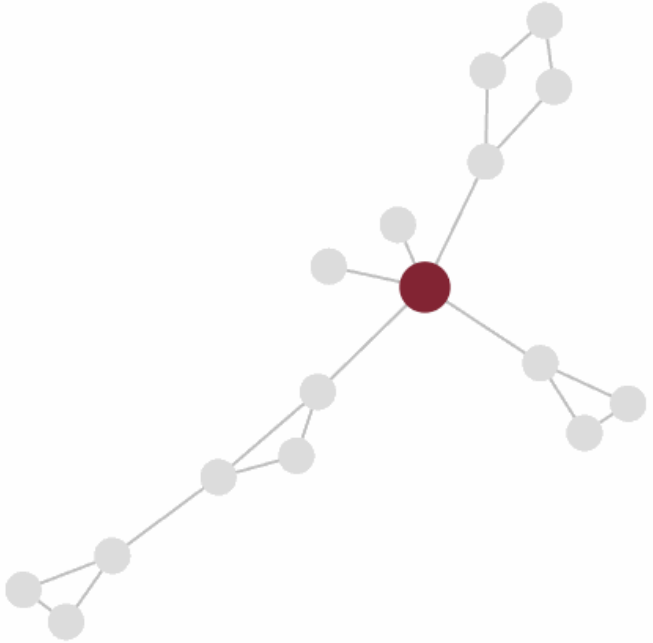
- Pair Decoder:
  inputs = embedding of node u and v;
  output = interaction score ℓ_{uv} (logit).



Encoder – TAGConv Layer (Multi-Hop)

Encoder Layer 1 – TAGConv (K=3)

- Aggregates over 1-hop, 2-hop, 3-hop neighbours in one layer.
- Captures most biologically relevant interaction paths (signal decays beyond 3 hops).
- Prevents need for many stacked layers -> avoids over-smoothing of node features.
- Balances information breadth vs. model depth (3 hops in one convolution).

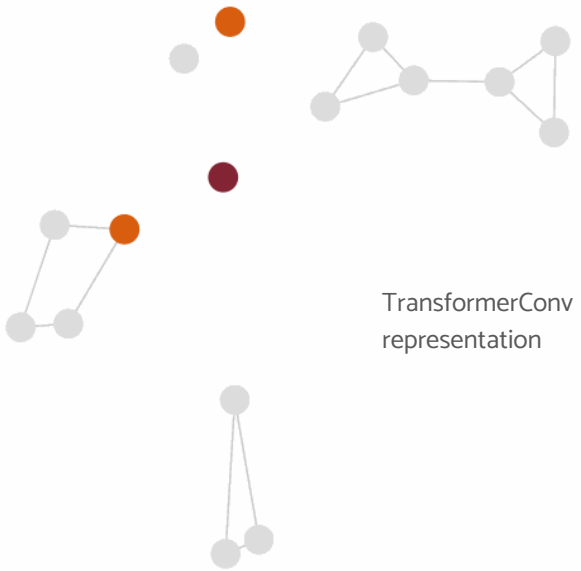


Encoder – Transformer and GIN

Encoder Layers 2 & 3

- TransformerConv (H=6): Graph attention mechanism inspired by the key/query/value mechanism of the Transformer, each of 6 heads learns to focus on different aspects/neighbours. Key point: modulates neighbour influence.

- GINConv layer: Graph Isomorphism Network layer, applies a simple learnable MLP after summing neighbor info. Key point: increases representational power; refines final embedding with a nonlinear transformation.

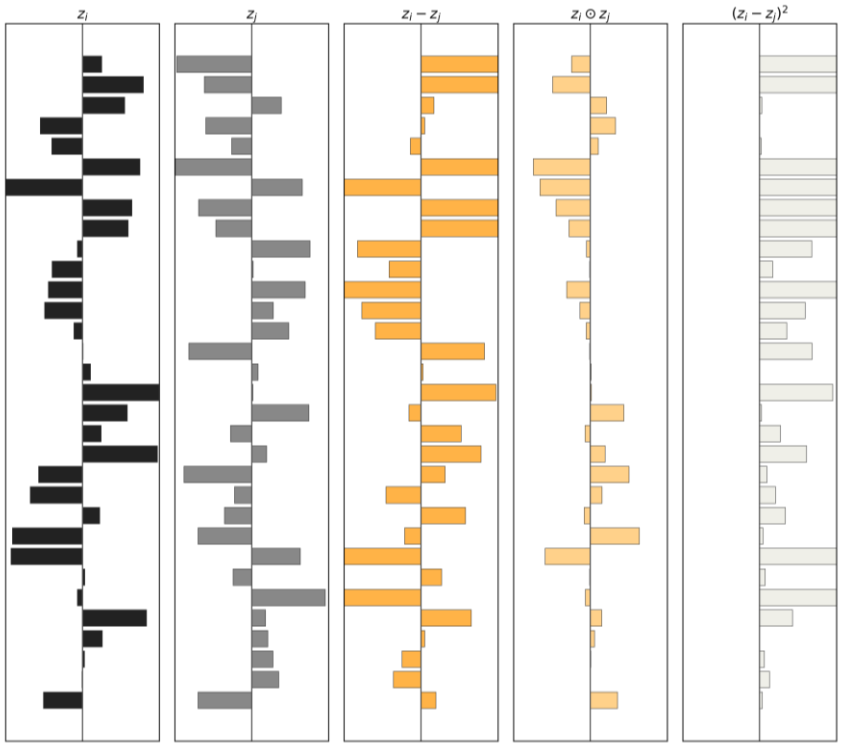


TransformerConv representation

Pair decoder overview

MLP for Edge Scoring

- Build a comparison vector with element-wise operations – the difference $(z_u - z_v)$, the Hadamard product $(z_u \odot z_v)$, and the squared difference $(z_u - z_v)^2$.
- Two-layer MLP: a dense layer (BatchNorm + ReLU) compresses and denoises; a second MLP laye r outputs the logit $\ell_{(uv)}$.
- Because the decoder is edge-local – it needs only the two embeddings, not the graph – we can score any pair after a single graph inference, producing a ranked list of novel interactions.



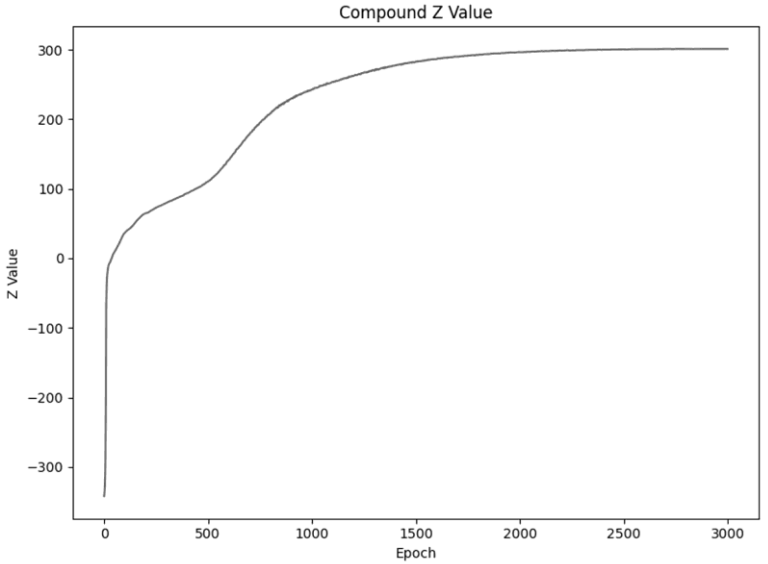
Parameter count of the model

Breakdown of the parameters

Component	Formula	# params
TAGConv (K+1 = 4)	(K+1) 1024 × 192+192	786624
TransformerConv (H=6 heads)	(H × 192 × 192 + 1 152) × 4	889344
GINConv	1152 × 1152	1327105
LayerNorm (affine)	2 × 1152	2304
Encoder subtotal		3005377
Pair MLP (3456 → 192 → 1)	3456 × 192 + 192 × 1	663744
Total model		3669121

Training Configuration

- Optimizer & epochs: AdamW, 3000 epochs; cosine annealing LR schedule.
- Batching: full-batch training on 10k-node graph (fits in 16 GB GPU).
- Regularization: edge dropout (randomly omit ~60% of message passing edges per epoch) to avoid overfitting topology.
- Checkpointing criterion: Compound Z Value made of standardized AUPRC + P@500 (focus on early precision).



$$z_{\mathrm{AUPRC},t} = \frac{\widehat{\mathrm{AUPRC}}_t - \mu_{\mathrm{AUPRC}}}{\sigma_{\mathrm{AUPRC}}}, \qquad z_{P@500,t} = \frac{\widehat{P@500}_t - \mu_{P@500}}{\sigma_{P@500}},$$

$$\boxed{Z_t = z_{\mathrm{AUPRC},t} + z_{P@500,t}}.$$

Overall Performance

Test Set Performance

- **Global separability**: AUROC 0.96 means the model ranks the true interaction above a negative 96% of the time.
- **Imbalance robustness**: AUPRC 0.89 is roughly 10× the 0.09 baseline from the 1:10 skew.
- **Practical hit rate**: All top-500 predictions are true positives (P@500 = 1.00).
- **Thresholded operation**: At a 0.5 cutoff, precision = 82% and recall = 81%, showing strong performance also on the general task.

Metric	Score
AUROC	0.966
AUPRC	0.885
Accuracy @ 0.50 threshold	0.967
Precision @ 0.50 threshold	0.825
Recall @ 0.50 threshold	0.809
P@500	1.000
NDCG@500	0.998

Early Precision Performance

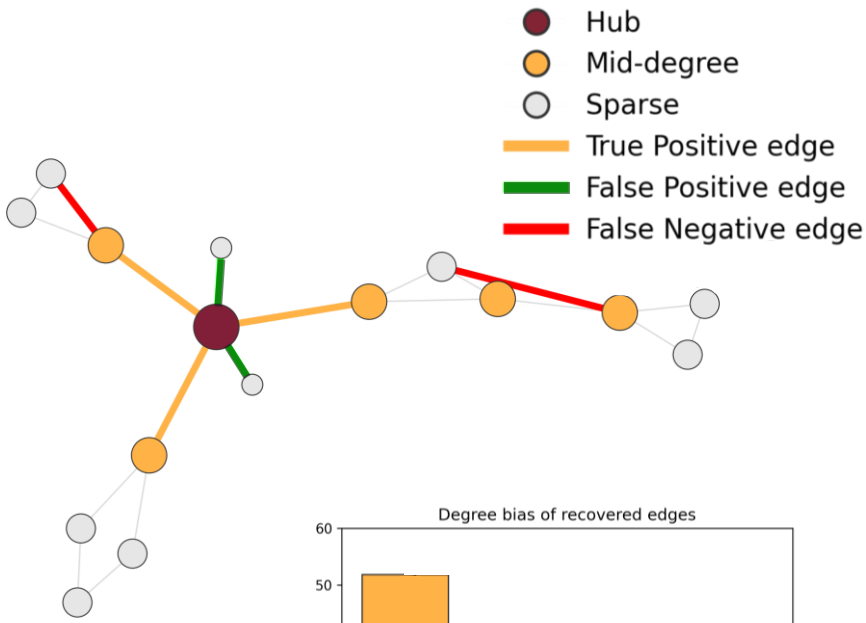
	K=50	K=100	K=500	K=1000	K=3000
$P@K$	1.000	1.000	1.000	1.000	0.984
NDCG@K	1.000	1.000	1.000	1.000	0.986

Early precision performance across different cutoffs. Both Precision@K and NDCG@K remain perfect (1.000) for K = 50, 100, 500, and 1 000, with a slight decrease to 0.984 (Precision) and 0.986 (NDCG) at K = 3 000, showing the strong model performance in regard to the top predictions.

Analysing Model Bias

Topological Bias in Predictions

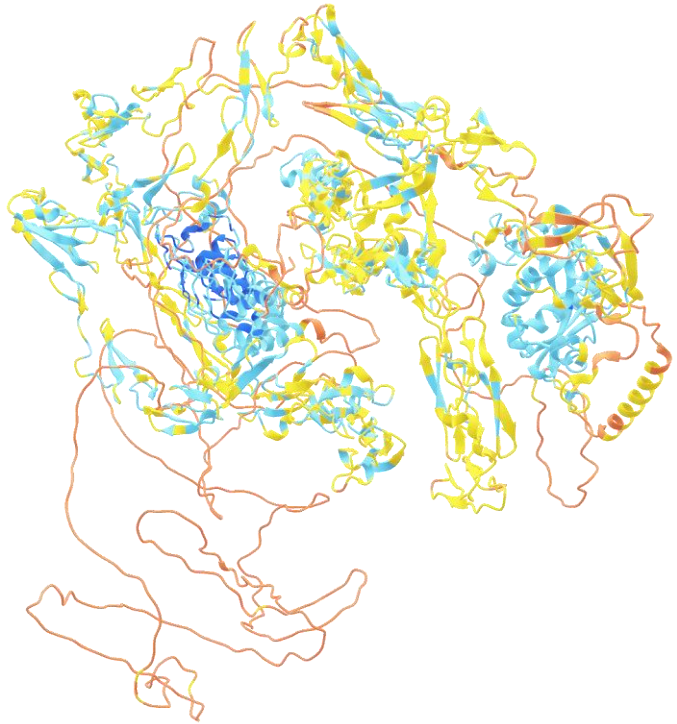
- Recovered test edges involved proteins with ~18% higher degree than average of all positive edges.
- Model confidence concentrates in dense network areas (hubs, clusters), as the number of shared neighbours is much higher for TP predictions.
- Low-degree proteins interactions are under-predicted (possible blind spot).
- Bias source: sparse neighborhoods provide less signal for GNN, causing cautious scores.



Case Study – NOTCH2 Protein

Topological Bias in Predictions

- NOTCH2: a signalling receptor (Notch family) with many known interactions. 2,294 partners in full STRING, 781 partners in LC-250 and 14 partners in HC-950.
- Dense scenario (LC-250 graph): includes lower-confidence edges (≥250), new edges and recovered edges study in different graph context.
- Sparse scenario (HC-950 graph): only new edges study in the same graph context used for the training, only higher-confidence edges (≥950) .

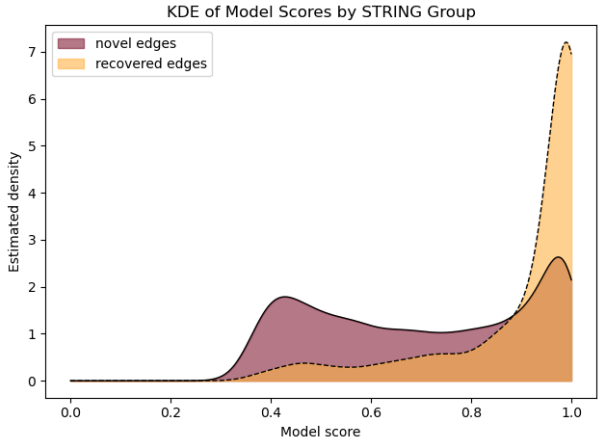


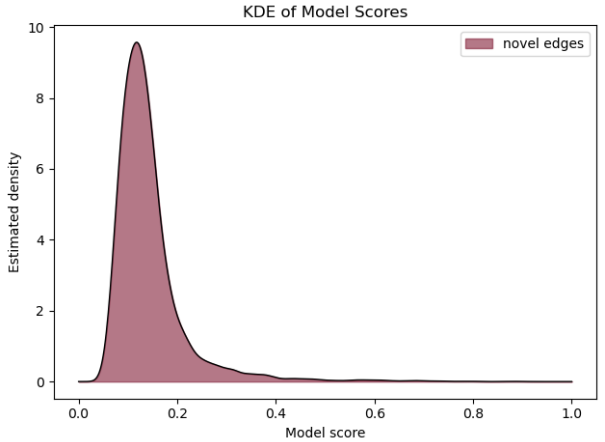
Statistical Results

Dense network (LC-250) vs Sparse network (HC-950)

- LC-250 graph: Yellow recovered edges distribution has a sharp spike near p=0.95; Red novel edges distribution has a broad shoulder ~0.45 and a smaller peak overlapping the high region at ~0.9.

- HC-950 graph: no recovered edges study, Red novel edges distribution peaks only around p=0.12, falls off sharply; no scores above ~0.9.





Biological Results

Case study takeaways

- In a dense context, model is bold – assigns very high scores (≈1.0) to likely partners (e.g. predicted NOTCH2–LRRK2, which has strong literature support, and even novel ones like CALML6).
- In a sparse context, model is cautious – few scores >0.8; still, top hits (e.g. PATJ, FBN1) seem biologically plausible.
- Demonstrates GNN's strength and dependency: can rediscover hidden interactions but needs rich topology for high confidence and easily gets context specific.

Graph setting	Predicted partner (score)	Biological plausibility
Dense (≥250)	**LRRK2 (1.0)**	Modulates Notch signalling; acts via HERC2/NEURL4 in neurons (Parkinson context).
	PLEKHA4 (1.0)	No direct NOTCH2 link, but a polarity/Wnt regulator. Low PLEKHA4 expression correlates with Notch up-regulation in breast cancer.
	CALML4/6 (1.0)	Ca^{2+}/calmodulin family; NOTCH1 signalling is CaM-dependent, suggesting a similar mechanism for NOTCH2.
	AKIRIN2 (1.0)	No known Notch-related studies; represents a genuinely novel candidate for validation.
Sparse (≥950)	**PATJ (0.85)**	Tight-junction scaffold; PATJ knock-out reduces Notch pathway activity in brain endothelial cells.
	FBN1 (0.83)	Extracellular-matrix fibrillin; matrix proteins can bind or prime Notch receptors, making interaction plausible.
	JUP (0.81)	Plakoglobin at adherens junctions; potential Wnt/Notch cross-talk, though no direct binding shown yet.

Limitations

- Topology bias: Underperforms on low-degree proteins; reliable mainly in dense network regions.
- Single data source: Trained solely on STRING@950, so may not generalize to other databases or confidence thresholds.
- No external baseline: Strong internal metrics but difficult to compare against other studies due to differing resources, methods and scopes.
- Scalability limits: Graphs over 5 million edges exhaust GPU memory; not yet optimized for very large networks.

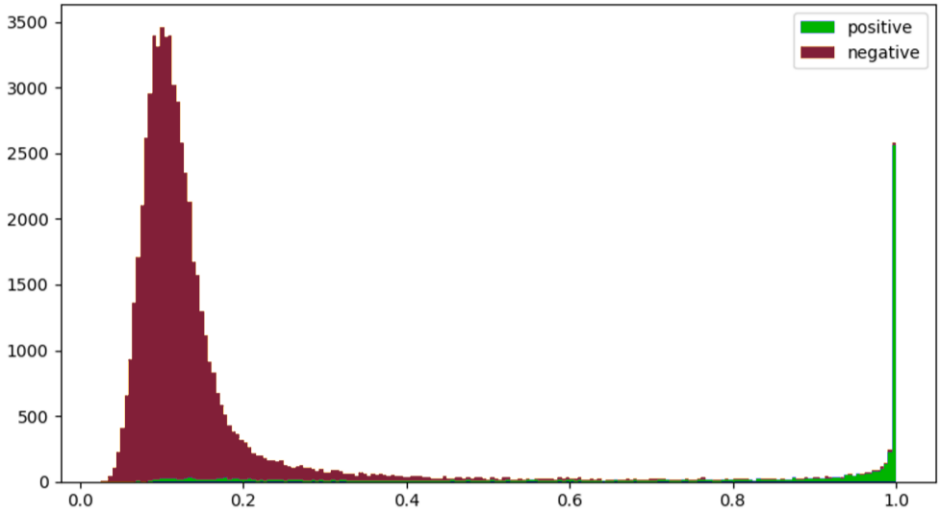
Future Work

- Generalisation: learn from—and be evaluated on—multiple PPI sources of differing density and reliability.
- Explainability: expose which sequence features, paths or local motifs drive a prediction, enabling trust and biological insight.
- Scalability: handle million-edge graphs via subgraph batching and curriculum schedules without degrading calibration

Conclusions – Contributions

Takeaways

- Developed a lightweight GNN + ProtT5 pipeline for PPI prediction.
- Achieved good general accuracy and precision ~100% in top-ranked predictions.
- Validated model's hypotheses with a case study (NOTCH2).
- Can significantly narrow experimental search space by providing a reliable short-list of candidate interactions.



Distribution of model scores on test set

Thank you for the attention

Questions?