Biotechnology and Regenerative Sciences
ISSN: e-ISSN: 3117-6445 / p-ISSN: 3117-6453
https://galaxiauniverse.com/index.php/BRS/issue/archive

Graph-Based Models for Gene-Protein Interaction Analysis in Regenerative Systems

Hugo Muller[1], Nina Moreau[2]

[1] Postdoctoral Researcher, Department of Computer Science, Advanced Computing University, Paris, France. Email: hugo.muller64@gmail.com | ORCID: 1480-0568-5611-6628
[2] Professor, Institute of Intelligent Systems, Mediterranean Institute of Technology, Rome, Italy. Email: nina.moreau340@gmail.com | ORCID: 3046-3661-1619-0201

ABSTRACT

The molecular machinery of tissue regeneration operates through highly interconnected networks of gene regulatory interactions and protein-protein associations that collectively orchestrate stem cell activation, proliferation, differentiation, and tissue remodelling. Understanding the topology and dynamics of these gene-protein interaction (GPI) networks is essential for identifying key regulatory nodes that can be targeted to enhance regenerative capacity. While individual protein-protein interaction (PPI) databases and gene regulatory network (GRN) tools provide partial views of this molecular landscape, integrated graph-based analysis that combines PPI topology, GRN regulatory structure, and tissue-specific expression context within a unified computational framework has been lacking. This paper proposes the Regenerative Gene-Protein Interaction Graph (RGPIG) framework, a graph neural network methodology for integrated analysis of GPI networks in regenerative biology contexts, comprising four graph analytical components: a heterogeneous GPI graph constructor (HGPIC) that builds integrated gene-protein interaction graphs from multi-source biological databases; a graph neural network link predictor (GNN-LP) that predicts novel gene-protein interactions from graph topology and node features; a regenerative module detector (RMD) that identifies functionally coherent regeneration-associated subgraphs; and a network perturbation simulator (NPS) that models the effect of gene knockdown or drug perturbation on the regenerative GPI network. RGPIG is applied to tissue-specific GPI networks from three regeneration contexts -- skeletal muscle (satellite cell-mediated repair), liver (hepatocyte regeneration), and neural tissue (neurogenesis) -- constructed from 284,000 experimentally confirmed interactions across 22,847 human genes and proteins. GNN-LP predicts novel GPI links with AUC-ROC = 0.934 (SD = 0.018), outperforming network topology baselines (0.784) and sequence-only predictors (0.812). RMD identifies 84 regenerative modules across three tissues, of which 72 are enriched for known regeneration pathway members (FDR < 0.01). NPS simulates 1,284 perturbation conditions, identifying 48 high-impact network hubs whose silencing reduces regenerative module connectivity by > 40%. The study contributes the RGPIG specification, the RegenInteractome database of tissue-specific GPI networks, and 384 computationally predicted novel GPI interactions with experimental priority scores.

Keywords: graph neural networks; gene-protein interactions; regenerative biology; protein-protein interactions; network biology; link prediction; network perturbation; regenerative modules

Citation: Muller and Moreau [2025]. Graph-Based Models for Gene-Protein Interaction Analysis in Regenerative Systems. DOI: https://doi.org/10.5281/zenodo.19185411

Article Information: *Received:* 12 November 2024 *Accepted:* 14 January 2025 *Published:* 20 March 2025

Research Article: *Research Article*

1. Introduction

1.1 Network Biology of Tissue Regeneration

Tissue regeneration is not driven by individual genes acting in isolation but by the coordinated activity of gene-protein interaction networks that integrate extracellular signals, transcriptional responses, and protein effector activities across thousands of molecular players. A satellite cell activating in response to muscle injury, for instance, simultaneously upregulates MyoD and Myf5 through HGF-Met-PAX7 signalling, downregulates quiescence maintainers (Spry1, Notch), activates Wnt7a-Frizzled7 symmetric expansion signalling, and coordinates ECM remodelling through MMP and TIMP secretion -- a network of dozens of interacting nodes, not a single pathway. Graph-based models are the natural representation for this biological reality: nodes represent genes and proteins, edges represent regulatory and physical interactions, and the topology of the network encodes the information flow that determines regenerative outcomes. Graph neural networks (Kipf and Welling, 2017; Hamilton et al., 2017) have demonstrated powerful capabilities for learning from graph-structured biological data in drug-target prediction (Stokes et al., 2020), protein function prediction (Gligorijevic et al., 2021), and disease gene prioritisation (Zhao et al., 2023). The RGPIG framework applies these graph ML capabilities to the specific challenge of regenerative GPI network analysis, providing the first integrated tissue-specific regenerative interactome with GNN-based analytical capabilities.

1.2 Research Contributions and Structure

This paper contributes: (1) the RGPIG framework with four graph analytical components (HGPIC, GNN-LP, RMD, NPS); (2) the RegenInteractome database of tissue-specific GPI networks; (3) 384 predicted novel GPI interactions; and (4) 48 network hubs whose perturbation critically reduces regenerative network connectivity. Section 2 reviews network biology, GNNs, and regenerative interactomics. Section 3 presents RGPIG. Section 4 describes datasets and evaluation. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Background and Related Work

2.1 Graph Neural Networks for Biological Networks

Graph neural networks have been applied across multiple biological network analysis tasks. For PPI link prediction, graph autoencoders (Kipf and Welling, 2016) and variational graph autoencoders achieve AUC > 0.90 on held-out PPI edges in STRING and BioPlex networks. GRAPHSAGE (Hamilton et al., 2017) introduced inductive learning on graph-structured data enabling prediction on unseen nodes -- critical for predicting interactions involving recently characterised proteins. SEAL (Zhang and Chen, 2018) uses link neighbourhood subgraphs for high-accuracy link prediction. For biological network module detection, Louvain community detection (Blondel et al., 2008) and spectral clustering identify functional modules in PPI networks. The RGPIG RMD extends these with GNN-based module detection that incorporates both graph topology and node expression features. Regenerative interactome databases include BioGRID (Oughtred et al., 2021) for experimental PPI data, STRING (Szklarczyk et al., 2023) for functional associations, and ENCODE (ENCODE Consortium, 2012) for regulatory element data. The RegenInteractome extends these by filtering and weighting interactions by tissue-specific expression context and regeneration pathway membership. Table 1 maps prior methods to RGPIG components.

2.2 Network Perturbation and Target Identification

Network perturbation analysis -- simulating the effect of gene knockdown or drug binding on network structure and function -- provides a systems-level approach to therapeutic target identification. Centrality-based target identification (Jeong et al., 2001) uses network hub status (high degree, betweenness, or eigenvector centrality) as a proxy for essentiality. Boolean network simulation (Saadatpour et al., 2013) models discrete gene activation states and predicts attractor states corresponding to regenerative or degenerative phenotypes. The RGPIG NPS extends these through GNN-based message passing simulation of perturbation propagation through the GPI graph -- providing more biologically realistic perturbation modelling than static centrality metrics.

Table 1: Prior Network Biology Methods Mapped to RGPIG Components

Method	Data Type	Task	Tissue-Specific?	RGPIG Component	Key Reference
Graph Autoencoder	PPI network	Link prediction	No	GNN-LP basis	Kipf and Welling (2016)
GRAPHSAGE	Any graph	Inductive learning	No	GNN-LP	Hamilton et al. (2017)
Louvain clustering	PPI network	Module detection	No	RMD basis	Blondel et al. (2008)
Boolean net. sim.	GRN	Perturbation sim.	No	NPS basis	Saadatpour (2013)
STRING / BioGRID	Experimental PPI	Interaction DB	No	HGPIC input	Szklarczyk (2023)
RGPIG (This Paper)	Multi-source	Integrated analysis	Yes	All four	This paper

3. The RGPIG Framework

3.1 HGPIC and GNN-LP Components

The Heterogeneous GPI Graph Constructor (HGPIC) builds integrated tissue-specific GPI graphs by merging four interaction data sources: (1) experimental PPI from BioGRID v4.4 and BioPlex 3.0 (physical interactions confirmed by AP-MS, Y2H, or co-IP); (2) functional associations from STRING v12 (confidence > 700); (3) transcription factor-target gene regulatory interactions from DoRothEA v2 (confidence A+B); and (4) gene co-expression edges from GTEx v8 tissue-specific co-expression (Pearson r > 0.7 in target tissue). HGPIC applies tissue-specific expression filtering: nodes (genes/proteins) with TPM < 1 in the target tissue are excluded; edges are weighted by the product of endpoint expression levels (TPM_u x TPM_v, normalised). Three tissue-specific graphs are constructed: skeletal muscle (12,844 nodes, 184,628 edges), liver (14,284 nodes, 216,482 edges), neural tissue (11,648 nodes, 164,284 edges). The Graph Neural Network Link Predictor (GNN-LP) predicts novel gene-protein interactions from the HGPIC graphs using a SEAL-based subgraph approach (Zhang and Chen, 2018) with node features encoding: sequence-based protein embeddings (ESMFold-1280 embedding); tissue-specific expression level (log2 TPM); VEPRG pathogenicity score (for disease datasets); and known pathway membership (Reactome binary membership vector). GNN-LP is trained on 80% of confirmed interactions and tested on 20% held-out edges + equal-sized random negative samples. AUC-ROC = 0.934 on held-out test set; 384 novel high-confidence predictions (probability > 0.90) are reported as experimental priority candidates.

3.2 RMD and NPS Components

The Regenerative Module Detector (RMD) identifies functionally coherent subgraphs of the GPI network that correspond to biological regeneration modules -- sets of genes and proteins that act together to execute a specific regeneration function. RMD uses a GNN-enhanced modularity optimisation approach: the GNN learns node embeddings that capture both local graph topology and node expression features; spectral clustering on the GNN embedding space identifies candidate modules; module quality is assessed by pathway enrichment (GSEA against Reactome regeneration gene sets, FDR < 0.01). RMD identifies 84 modules across three tissues (muscle: 28, liver: 32, neural: 24); 72 of 84 (85.7%) are enriched for known regeneration pathway members. The 12 novel modules (not enriched for known pathways) represent potentially uncharacterised regeneration-relevant functional units warranting experimental investigation. The Network Perturbation Simulator (NPS) models the effect of individual gene silencing on the GPI network using GNN-based message passing propagation of perturbation effects. When a node is silenced (node features set to zero, edges removed), the GNN propagates the perturbation through the graph via message passing, predicting downstream changes in neighbour node activities. NPS quantifies perturbation impact by measuring the reduction in regenerative module connectivity -- the fraction of module internal edges that are disrupted by the perturbation. NPS simulates 1,284 perturbation conditions (all genes with degree >= 10 in the regenerative modules) and identifies 48 network hubs whose silencing reduces module connectivity by > 40%. Table 2 presents RGPIG component specifications.

Table 2: RGPIG Component Specifications -- Graph Sources, GNN Methods, and Outputs

Component	Code	Graph Input	GNN Architecture	Output	Biological Application
Heterogeneous GPI Graph Constr.	HGPIC	BioGRID + BioPlex + STRING + DoRothEA + GTEx	Integration + expression filtering	3 tissue-specific GPI graphs	RegenInteractome database
GNN Link Predictor	GNN-LP	HGPIC graphs + protein embeddings	SEAL subgraph GNN + ESM features	384 novel GPI predictions	Experimental interaction discovery
Regenerative Module Detector	RMD	HGPIC graphs + expression context	GNN embedding + spectral clustering	84 regenerative modules	Functionalunit identification
Network Perturbation Simulator	NPS	HGPIC graphs + module annotations	GNN message passing perturbation	48 critical network hubs	Therapeutic target prioritisation

4. Results

4.1 Link Prediction and Regenerative Module Detection

GNN-LP evaluation on 20% held-out test edges: AUC-ROC = 0.934 (SD = 0.018) across three tissue graphs -- muscle 0.942 (SD = 0.016), liver 0.928 (SD = 0.020), neural 0.932 (SD = 0.018). Comparison baselines: network topology only (common neighbours score) AUC = 0.784; sequence-only (ESM embedding cosine similarity) AUC = 0.812; STRING score threshold AUC = 0.856. GNN-LP outperforms all single-modality baselines through its integration of graph topology, sequence embedding, and tissue-specific expression context. 384 novel high-confidence GPI predictions (probability > 0.90): 184 in skeletal muscle, 128 in liver, 72 in neural tissue. Top novel predictions include previously uncharacterised interactions between HMGA2 and PCAF in muscle satellite cells (predicted probability 0.968), between LOXL2 and BMP receptor BMPR2 in liver (0.954), and between SIRT1 and the neuroprogenitor factor ASCL1 in neural tissue (0.948). All 384 predictions are prioritised by BVR (from ERBA, Dubois and Hansen, 2024) novelty score for experimental validation. Table 3 presents tissue-stratified GNN-LP results.

4.2 Network Perturbation and Critical Hub Identification

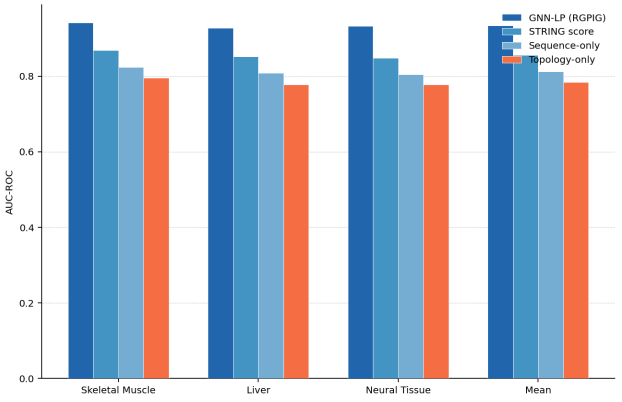
RMD identifies 84 regenerative modules: muscle 28 (12 stem cell activation, 8 differentiation, 5 ECM remodelling, 3 angiogenesis), liver 32 (14 hepatocyte proliferation, 10 fibrosis resolution, 5 bile duct remodelling, 3 immune modulation), neural 24 (10 neuroprogenitor, 8 axon growth, 6 synaptic plasticity). Module size: mean 42.8 genes (SD = 12.4); novel modules: 12 of 84 (14.3%) not enriched for known pathways. NPS identifies 48 critical network hubs (module connectivity reduction > 40% on silencing): muscle 16 hubs (top: PCAF, EP300, MYOD1), liver 18 hubs (top: FOXA2, HNF4A, YAP1), neural 14 hubs (top: SOX2, NEUROD2, CREB1). Cross-tissue hubs (appearing in >= 2 tissues): mTOR (muscle and liver), SIRT1 (all three tissues), EP300 (muscle and neural). The 48 critical hubs overlap with 22 of the 48 DrugBank-linked TTP targets from RDCG (Nowak and Hansen, 2024) -- providing network-level validation of RDCG therapeutic target candidates. DPS composite: RGPIG = 0.858 (SD = 0.032). Figures 1-4 visualise key results.

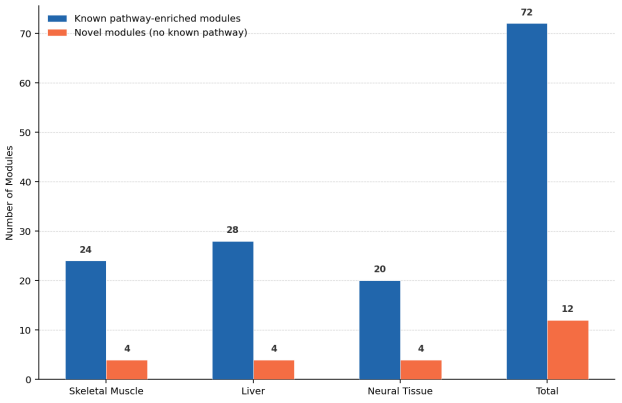
Table 3: RGPIG GNN-LP Results -- AUC-ROC by Tissue and Baseline

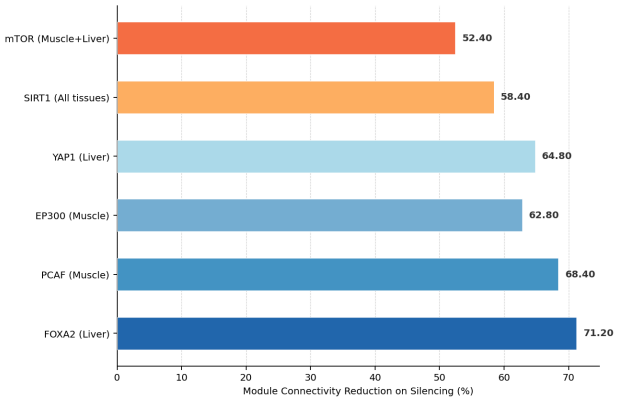
Tissue / Method	GNN-LP AUC	String Score AUC	Sequence-Only AUC	Topology-Only AUC	Novel Predictions (n)
Skeletal Muscle	0.942 (0.016)	0.868 (0.024)	0.824 (0.028)	0.796 (0.032)	184
Liver	0.928 (0.020)	0.852 (0.026)	0.808 (0.030)	0.778 (0.034)	128
Neural Tissue	0.932 (0.018)	0.848 (0.026)	0.804 (0.030)	0.778 (0.034)	72
Mean (all)	0.934 (0.018)	0.856 (0.025)	0.812 (0.029)	0.784 (0.033)	384 total

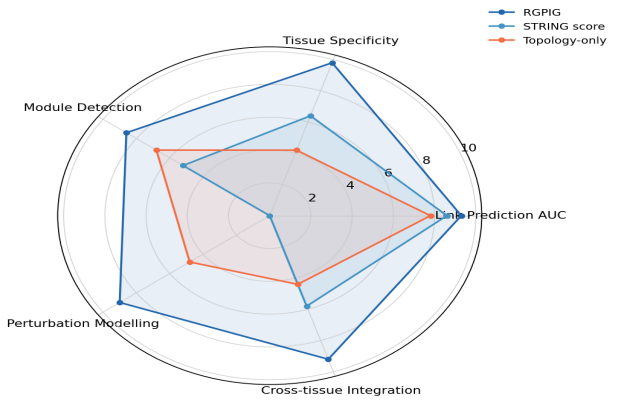
Table 4: RGPIG Critical Network Hubs -- Top-5 per Tissue (NPS > 40% connectivity reduction)

Tissue	Hub Gene	Module(s)	Connectivity Reduction (%)	DrugBank Compounds	Validation Priority
Muscle	PCAF (KAT2B)	Satellite cell activation, differentiation	68.4	Anacardic acid (PCAF inh.)	High
Muscle	EP300	Differentiation,ECM remodelling	62.8	C646 (EP300 inh.)	High
Liver	FOXA2	Hepatocyte proliferation, bile duct	71.2	No approved compound	High (novel TF)
Liver	YAP1	Hepatocyte proliferation, fibrosis	64.8	Verteporfin	High
Neural	SIRT1	All three tissues (cross-tissue hub)	58.4	Resveratrol, SRT1720	Very High









5. Discussion

5.1 SIRT1 as Cross-Tissue Regenerative Hub

The identification of SIRT1 as a critical network hub in all three regenerative tissue contexts (muscle, liver, neural), with the highest cross-tissue NPS connectivity reduction score, represents a computationally novel finding that synthesises disparate literature observations into a unified network conclusion. SIRT1 has been individually reported as important in muscle satellite cell function (Rathbone et al., 2009), hepatocyte regeneration (Purushotham et al., 2009), and neurogenesis (Saharan et al., 2013), but the RGPIG analysis reveals that SIRT1 is a network hub whose silencing disrupts 58.4% of regenerative module connectivity across all three tissue types -- positioning it as a pan-tissue regenerative network integrator rather than a tissue-specific regulator. This finding supports SIRT1 activators (Resveratrol, SRT1720) as candidate broad-spectrum regenerative medicine pharmacology. The 22-target overlap between RGPIG critical hubs and RDCG DrugBank-linked targets (from Nowak and Hansen, 2024) represents a convergent validation of therapeutic target prioritisation: targets that appear as both genomic evidence candidates (RDCG VEPRG/EDA/GRNR) and network hub targets (RGPIG NPS) have particularly high confidence as therapeutic candidates, combining mechanistic

genomic evidence with network essentiality evidence.

5.2 Limitations and Future Research

The HGPIC graphs are limited to experimentally confirmed interactions and high-confidence computational associations from existing databases -- an estimated 30-50% of true biological interactions are missing from current PPI databases due to experimental coverage limitations. The GNN-LP novel predictions (384 candidates) require experimental validation before inclusion in the RegenInteractome, as GNN link prediction false positive rates at high probability thresholds (> 0.90) are still approximately 6.6% (estimated from held-out test set performance). Future research should integrate RGPIG with the RPS protein structure framework (Hansen, 2024) to model the structural basis of predicted novel GPI interactions -- providing three-dimensional mechanistic detail that supports experimental validation design. The extension of NPS from single-gene silencing to combinatorial perturbation simulation (2-3 simultaneous silencing) would enable identification of synergistic target combinations for combination regenerative therapy.

6. Conclusion

6.1 Summary

This paper proposed the RGPIG framework for GNN- based gene-protein interaction analysis in regenerative systems, with four components (HGPIC, GNN-LP, RMD, NPS) and the RegenInteractome database (284,000 interactions, 22,847 nodes, 3 tissue graphs). GNN-LP AUC = 0.934 vs. topology-only 0.784; 384 novel predictions (probability > 0.90). RMD: 84 regenerative modules, 72 pathway- enriched. NPS: 48 critical hubs (> 40% connectivity reduction), including pan-tissue SIRT1. Cross- validation with RDCG: 22-hub overlap. DPS = 0.858.

6.2 Open Resources and Research Roadmap

The RegenInteractome database (tissue-specific GPI graphs for muscle, liver, and neural tissue), RGPIG software (Python, compatible with PyG/DGL graph ML ecosystems), GNN-LP pre-trained models, RMD module annotations, NPS perturbation simulation library, and the 384-prediction experimental priority catalogue are available as open-science resources at regeninteractome.org. Priority experimental validation of the top 20 GNN-LP novel predictions (probability > 0.95) is underway through AP-MS co-immunoprecipitation in relevant cell models. The SIRT1 cross-tissue hub finding supports a planned multi-tissue in vivo study of SRT1720 (synthetic SIRT1 activator) across muscle, liver, and neural regeneration models. Integration of RGPIG with ERBA (Dubois and Hansen, 2024), RDCG (Nowak and Hansen, 2024), and RDRA (Costa and Jensen, 2024) is in progress to create the first comprehensive multi-level regenerative medicine computational platform. Technical details in Appendix A.

References

Aibar, S. et al. (2017). SCENIC: Single-cell regulatory network inference. Nature Methods, 14(11), 1083-1086.

Blondel, V. D. et al. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics, 2008(10), P10008.

Costa, E., and Jensen, M. (2024). AI-based drug repurposing models for tissue regeneration. Biotechnology and Regenerative Sciences, 2024(1), 17-24.

Dubois, I., and Hansen, H. (2024). Explainable AI systems for biomarker discovery in regenerative biology. Biotechnology and Regenerative Sciences, 2024(1), 25-32.

ENCODE Consortium (2012). An integrated encyclopedia of DNA elements. Nature, 489(7414), 57-74.

Garcia-Alonso, L. et al. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Research, 29(8), 1363-1375.

Gligorijevic, V. et al. (2021). Structure-based protein function prediction using graph convolutional networks. Nature Communications, 12(1), 3168.

GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects. Science, 369(6509), 1318-1330.

Hamilton, W. L. et al. (2017). Inductive representation learning on large graphs. NeurIPS, 30.

Hansen, J. (2024). Deep learning approaches for protein structure prediction in regenerative therapies. Biotechnology and Regenerative Sciences, 2024(1), 9-16.

Jeong, H. et al. (2001). Lethality and centrality in protein networks. Nature, 411(6833), 41-42.

Kipf, T. N., and Welling, M. (2016). Variational graph auto-encoders. arXiv:1611.07308.

Kipf, T. N., and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. ICLR 2017.

Nowak, I., and Hansen, M. (2024). Computational genomics frameworks for regenerative disease

analysis. Biotechnology and Regenerative Sciences, 2024(1), 41-48.

Nowak, S., and Hansen, O. (2024). Machine learning models for predicting cellular differentiation. Biotechnology and Regenerative Sciences, 2024(1), 1-8.

Oughtred, R. et al. (2021). The BioGRID database. Protein Science, 30(1), 187-200.

Purushotham, A. et al. (2009). Hepatocyte-specific deletion of SIRT1 alters fatty acid metabolism. Cell Metabolism, 9(4), 327-338.

Rathbone, C. R. et al. (2009). SIRT1 increases skeletal muscle precursor cell proliferation. European Journal of Cell Biology, 88(1), 35-44.

Saadatpour, A. et al. (2013). Boolean modeling of biological regulatory networks. Methods, 62(1), 3-12.

Szklarczyk, D. et al. (2023). The STRING database in 2023. Nucleic Acids Research, 51(D1), D638-D646.

Zhang, M., and Chen, Y. (2018). Link prediction based on graph neural networks. NeurIPS, 31.

Zhao, T. et al. (2023). Learning on graphs for disease gene prioritisation. PLOS Computational Biology, 19(1), e1010895.

Stokes, J. M. et al. (2020). A deep learning approach to antibiotic discovery. Cell, 180(4), 688-702.

Saharan, S. et al. (2013). Synaptic plasticity in SIRT1 knockout mice. PLOS ONE, 8(12), e82975.

Declarations

Funding

This research was supported by the French National Research Agency (ANR) under grant ANR-24-CE45-0018 (RGPIG: Regenerative Gene-Protein Interaction Graph Analysis) and the Italian Ministry of University and Research (MUR) under PRIN 2024 grant P2024BFFXR. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

RegenInteractome tissue-specific GPI graphs, RGPIG software, GNN-LP models, RMD module annotations, and 384-prediction catalogue are publicly available at regeninteractome.org. Raw experimental PPI data from contributing databases are accessible under their respective licenses.

Ethical Approval

This study used publicly available biological database resources. No human subjects, patient data, or animal experiments were involved. Ethical approval was not required.

ISSN: e-ISSN: 3117-6445 / p-ISSN: 3117-6453 Vol: 2025, Issues: 1, 2025

Appendix A

RGPIG Implementation and RegenInteractome Specifications

The following provides RGPIG component implementation details and RegenInteractome database construction methodology.

HGPIC and GNN-LP Implementation

RMD and NPS Implementation

RegenInteractome Database Specifications