geneDRAGNN: Gene Disease Prioritization using Graph Neural Networks — Supplementary Material

Awni Altabaa, David Huang, Ciaran Byles-Ho, Hani Khatib, Fabian Sosa, Ting Hu

Last Updated: July 16, 2022

1 Modeling Results

Table 1: Model Results

Model	Features Used	Accuracy (avg)	Accuracy (std)	Recall (avg)	Recall (std)	Precision (avg)	Precision (std)	F1 (avg)	F1 (std)	# of Trials
Baseline Models										
RF	node features	0.707	0.027	0.728	0.037	0.700	0.030	0.707	0.027	100
MLP	node features	0.699	0.025	0.668	0.029	0.714	0.035	0.699	0.025	100
SVM	node features	0.693	0.022	0.506	0.040	0.809	0.039	0.681	0.024	100
KNN	node features	0.645	0.024	0.451	0.042	0.737	0.040	0.630	0.026	100
Graph-Based Models										
RF	node2vec features	0.765	0.022	0.666	0.039	0.831	0.030	0.762	0.022	100
RF	node features, node2vec features	0.766	0.022	0.705	0.037	0.803	0.027	0.765	0.022	100
MLP	node2vec features	0.731	0.027	0.736	0.051	0.730	0.032	0.730	0.027	100
MLP	node features, node2vec features	0.744	0.026	0.735	0.038	0.749	0.031	0.744	0.026	100
SVM	node2vec features	0.780	0.022	0.758	0.036	0.794	0.028	0.780	0.022	100
SVM	node features, node2vec features	0.705	0.021	0.530	0.036	0.815	0.033	0.695	0.022	100
KNN	node2vec features	0.564	0.017	0.949	0.023	0.537	0.010	0.488	0.029	100
KNN	node features, node2vec features	0.645	0.026	0.752	0.041	0.620	0.025	0.640	0.027	100
SGCN	node features, LS-GIN network, edge features	0.750	0.019	0.774	0.080	0.743	0.034	0.749	0.019	11
SGCN	node features, LS-GIN network	0.743	0.028	0.750	0.084	0.746	0.050	0.741	0.028	100
GraphSAGE	node features, LS-GIN network	0.714	0.015	0.674	0.040	0.733	0.020	0.713	0.016	10
GraphSAGE	node features, LS-GIN network, node2vec features	0.745	0.020	0.721	0.047	0.759	0.027	0.745	0.021	16
TAGCN	node features, LS-GIN network	0.749	0.024	0.706	0.067	0.778	0.049	0.747	0.024	10
TAGCN	node features, LS-GIN network, node2vec features	0.741	0.035	0.726	0.047	0.750	0.042	0.741	0.035	10
Cluster-GCN	node features, LS-GIN network, edge features	0.726	0.020	0.671	0.068	0.757	0.031	0.724	0.021	11

2 Literature-based Validation of Top Prioritized Genes

Table 2: A detailed literature analysis of some of the top predicted genes associated with LUAD from the SGConv model. * indicates that the result was predicted with the MLP model.

Rank	Gene	Gene functional description and Literature review
2	CDC42	This gene encodes a member of the Rho subfamily of small GTP-binding proteins and plays a key role in cancer cell migration and metastasis. [1, 2] found decreased levels of StarD13, a surpressor of CDC42, in lung tumor tissue and A549 cells subsequently leading to increased CDC42 activation thus increasing formation of invadopodia, a unique hallmark of cancer, and matrix degradation.
3	PTPRC	This gene encodes a tyrosine-protein phosphatase required for T-cell activation. Upon T-cell activation, PTPRC recruits and dephosphorylates FYN which has been shown to be correlated with LUAD by [1]. [3] directly supports PTPRC association with LUAD by demonstrating that PTPRC was a key gene in affecting the immune state of the tumor microenvironment and was ultimately correlated with a variety of tumor-infiltrating immune cells.
4	LRRK2	In an analysis of TCGA LUAD RNA-seq data, [4] identified that decreased LRRK2 expression is associated with LUAD. In [5], reduced LRRK2 expression was found to promote LUAD tumorigenesis and was associated with poor survival outcomes. This study also found overactivity in LRRK2 contributes to Parkinson's disease, which suggests pathological links between neurodegenerative disease and cancer are emerging.
5	CREB1	This gene encodes a phosphorylation-dependent transcription factor that stimulates transcription upon binding to DNA cAMP response element (CRE) [1]. [6] used protein expression assays to understand the underlying mechanism of ferroptosis, a new form of regulated cell death associated with cancer, in LUAD. They found that CREB was highly expressed in LUAD and knockdown of CREB inhibited cell viability and growth by promoting apoptosis- and ferroptosis-like cell death.
8	ISG15	ISG15 acts as a cytokine, modulating immune responses, and can delay tumor cell growth by inhibiting tumor cell proliferation and angiogenesis. [7] found that high expression of ISG15 serves as a positive prognostic marker for long-term survival in LUAD patients. ISG15 has a broad network of protein targets, and [8] concludes that covalent ISG15 conjugation enhances the tumor-suppressive activity of the carboxyl terminus of Hsp70-interacting protein (CHIP), thereby showing an antitumor effect of Type 1 interferon.
9	FYN	FYN encodes a tyrosine-protein kinase essential in cell motility and adhension and plays an important role in the PI3K/AKT pathway responsible for regulating the cell cycle [1]. [9] demonstrated that overexpression of FYN accelerated cell apoptosis and reduced both angiogensis capacity and PI3K/AKT expression levels in lung carcinoma A549 cells. Conversely, FYN expression was shown to be correlated with LUAD prognosis as FYN expression levels were shown to be down-regulated in LUAD tissues and cells .
10	ITGB1	ITGB1 encodes the Integrin beta-1 subunit which when associated with Integrin alpha-3 provides a docking site for FAP, a serine protease involved in extracellular matrix degradation and tumor growth, at invadopodia plasma membranes. Hence ITGB1 may participate in formation of invadopodia, matrix degradation, and can promote cell invasion [1]. Immune infiltration analysis revealed that the ITGB1-DT/ARNTL2 axis may effect the progression of LUAD and the immune microenvironment. ITGB1-DT/ARTL2 [10]

4800*	FTL	A study published in 2021 demonstrated a negative correlation between high expression of FTL or LUAD malignancy and a decrease in lipid peroxidation and labile. [11] In an RNA-sequencing study, FTL and FTH1 expression levels were significantly positively correlated with tumor infiltration by tumor-associated macrophages and T regulatory cells. [12] Ferritin light chain (FTL) and ferritin heavy chain (FTH1) were compared and FTL subunits had higher expression levels for LUAD than its heavy chain iron counterpart. FTL also had higher expression for tumorous tissues in most types of cancer, and LUAD showed higher levels of expression and tumor tissue (in comparison to other subtypes of cancer) for the FTL experiment [12]
12626*	TFF3	It is involved in the maintenance and repair of the intestinal mucosa. It also promotes the mobility of epithelial cells during healing [1] The highest classified gene from the MLP node-only model is TFF3. This gene was excluded from the training set because it had a GDA score of 0.01, but an association has already been detected by DisGeNet. In 2021, for the first time it was shown that TFF3 was associated with sub diseases of LUAD and was specifically associated with LUAD invasiveness [13]. In 2019, TFF3 was identified as a biomarker to be able to distinguish between LUAD and lung squamous-cell carcinoma (SCC) [14]. In [14], over 90% of LUAD was observed to be TFF3 positive, whereas no TFF3 expression was observed in cases of SCC, suggesting TFF3 is an insightful biomarker that enables experts to distinguish and diagnose subtypes of non-small cell lung cancer (NSCLC).

3 Functional Enrichment Analysis

Table 3: Functional Enrichment Top 10 genes

Category	Term	Count	P-value
GOTERM_MF_DIRECT	Identical protein binding	6	1.8×10^{-5}
GOTERM_BP_DIRECT	vascular endothelial growth factor receptor signaling pathway	3	6.4×10^{-4}
GOTERM_CC_DIRECT	plasma membrane	8	7.0×10^{-4}
GOTERM_CC_DIRECT	focal adhesion	4	7.5×10^{-4}
GOTERM_BP_DIRECT	ephrin receptor signaling pathway	3	9.1×10^{-4}
GOTERM_BP_DIRECT	Fc-gamma receptor signaling pathway involved in phagocytosis	3	2.0×10^{-3}
GOTERM_BP_DIRECT	negative regulation of gene expression	3	2.3×10^{-3}
GOTERM_BP_DIRECT	positive regulation of GTPase activity	4	2.7×10^{-3}
GOTERM_BP_DIRECT	axon guidance	3	3.1×10^{-3}
GOTERM_CC_DIRECT	membrane raft	3	4.3×10^{-3}
GOTERM_CC_DIRECT	external side of plasma membrane	3	4.6×10^{-3}
GOTERM_CC_DIRECT	extracellular exosome	6	6.4×10^{-3}
GOTERM_BP_DIRECT	MAPK cascade	3	8.1×10^{-3}
GOTERM_BP_DIRECT	cellular response to platelet-derived growth factor stimulus	2	9.6×10^{-3}
GOTERM_CC_DIRECT	cytosol	6	1.3×10^{-2}
GOTERM_BP_DIRECT	sarcomere organization	2	1.5×10^{-2}
GOTERM_CC_DIRECT	membrane	5	1.6×10^{-2}

Table 4: Functional Enrichment Top Decile genes

Category	Term	Count	P-value
GOTERM_MF_DIRECT	protein binding	1076	1.7×10^{-102}
GOTERM_CC_DIRECT	cytosol	562	1.3×10^{-94}
GOTERM_CC_DIRECT	extracellular exosome	451	5.5×10^{-64}
GOTERM_CC_DIRECT	nucleoplasm	410	7.5×10^{-47}
GOTERM_CC_DIRECT	extracellular matrix	104	1.6×10^{-42}
GOTERM_BP_DIRECT	T cell receptor signaling pathway	71	2.7×10^{-37}
GOTERM_CC_DIRECT	membrane	320	3.8×10^{-34}
GOTERM_CC_DIRECT	cell surface	131	4.7×10^{-34}
GOTERM_BP_DIRECT	inflammatory response	111	5.9×10^{-34}
GOTERM_BP_DIRECT	NIK/NF-kappaB signaling	46	6.4×10^{-34}

Table 5: Functional Enrichment MLP

Category	Term	Count	Percent	P-value
GOTERM_CC_DIRECT	intracellular ferritin complex	2	20	9.87×10^{-4}
GOTERM_CC_DIRECT	cell	3	30	0.001 047
GOTERM_MF_DIRECT	iron ion binding	3	30	0.002 205
GOTERM_BP_DIRECT	intracellular sequestering of iron ion	2	20	0.003 212
GOTERM_CC_DIRECT	autolysosome	2	20	0.003 945
GOTERM_MF_DIRECT	ferric iron binding	2	20	0.004 73
GOTERM_CC_DIRECT	extracellular region	5	50	0.005 32
GOTERM_BP_DIRECT	iron ion transport	2	20	0.006 415
GOTERM_CC_DIRECT	endocytic vesicle lumen	2	20	0.007 876
GOTERM_BP_DIRECT	translational elongation	2	20	0.009 608
GOTERM_CC_DIRECT	cytosol	6	60	0.013 053
GOTERM_BP_DIRECT	cellular iron ion homeostasis	2	20	0.023 342
GOTERM_MF_DIRECT	protein binding	8	80	0.045 004
GOTERM_BP_DIRECT	platelet degranulation	2	20	0.053 882
GOTERM_CC_DIRECT	blood microparticle	2	20	0.072 625
GOTERM_CC_DIRECT	cytoplasm	6	60	0.083 113
GOTERM_BP_DIRECT	receptor-mediated endocytosis	2	20	0.095 407

4 KEGG Pathway Enrichment

Table 6: KEGG Pathway Enrichment Top 10

Category	path	Count	p-value
KEGG_PATHWAY	Pathogenic Escherichia coli infection	4	3.1×10^{-5}
KEGG_PATHWAY	Focal adhesion	4	1.9×10^{-3}
KEGG_PATHWAY	Shigellosis	3	2.9×10^{-3}
KEGG_PATHWAY	Adherens junction	3	3.6×10^{-3}
KEGG_PATHWAY	Bacterial invasion of epithelial cells	3	4.3×10^{-3}
KEGG_PATHWAY	Salmonella infection	3	4.9×10^{-3}
KEGG_PATHWAY	T cell receptor signaling pathway	3	7.0×10^{-3}
KEGG_PATHWAY	Leukocyte transendothelial migration	3	9.2×10^{-3}
KEGG_PATHWAY	Axon guidance	3	1.1×10^{-2}
KEGG_PATHWAY	Platelet activation	3	1.2×10^{-2}

Table 7: KEGG Pathway Enrichment Top Decile

Category	path	Count	p-value
KEGG_PATHWAY	Epstein-Barr virus infection	58	2.2×10^{-17}
KEGG_PATHWAY	Pathways in cancer	120	6.4×10^{-16}
KEGG_PATHWAY	Measles	59	6.8×10^{-16}
KEGG_PATHWAY	PI3K-Akt signaling pathway	106	2.8×10^{-14}
KEGG_PATHWAY	Osteoclast differentiation	56	3.0×10^{-14}
KEGG_PATHWAY	Toxoplasmosis	49	2.3×10^{-13}
KEGG_PATHWAY	TNF signaling pathway	48	3.1×10^{-13}
KEGG_PATHWAY	Proteasome	28	1.7×10^{-12}
KEGG_PATHWAY	Chagas disease (American trypanosomiasis)	44	3.7×10^{-11}
KEGG_PATHWAY	HTLV-I infection	79	4.5×10^{-11}

5 Model Parameters

The MLPs used a simple architecture with 2 hidden layers of 128 units each and ReLU activation, implemented with PyTorch [15]. The Random Forests, K-Nearest Neighbours, and Support Vector Machine Classifiers used the default Sci-Kit Learn parameters. The number of estimators of the random forests was set to 100 with the Gini impurity criterion and bootstrapping enabled for the individual trees. The K-Nearest Neighbors classifier used k=5 and the standard euclidean metric. The Support Vector Machine classifier used the Radial Basis Function kernel. Random Forests, K-Nearest Neighbours, and Support Vector Machine classifiers were implemented with Sci-Kit Learn [16].

The GNN models were implemented with the ‘PyTorch Geometric’ library [17]. First, we converted LS-GIN to a format readable by PyTorch Geometric by mapping our Ensembl gene identifiers to a 0-based indexing system and creating an edge list of shape `[2, n_edges]`. The edge features (not used by all GNN models) are represented by`[n_edges, n_edge_feats]` tensors. The Adam optimizer is used [18] with the cross-entropy loss and models are trained for 250 epochs. Throughout the training process, the model is evaluated on the training and validation sets, and model checkpoints are saved along the way. At the end of the 250 epochs, the model with the highest validation accuracy is restored and evaluated on the test set. This completes one trial. As explained in the main paper, for each model we evaluate on 100 different trials.

Table 8: GNN Model Hyperparameters

Model	Hyperparameters

SGCN	`conv hidden channels = [128, 256, 256, 128]`
	`dense hidden layer = 128`
	`number of hops = 1`
	`add self loops = True`
	`bias = True`

TAGCN	`conv hidden channels = [128, 256, 128]`
	`dense hidden layer = 256`
	`number of hops = 3`
	`apply symmetric normalization = True`
	`bias = True`

GraphSAGE	`conv hidden channels = [256, 256, 256]`
	`jumping knowledge mode = ’max’`

Cluster-GCN	`conv hidden channels = [128, 256, 256, 128]`
	`dense hidden layer = 128`
	`diagonal enhancement value = 0`
	`add self loops = True`
	`bias = True`

References

[1] T. U. Consortium, “UniProt: the universal protein knowledgebase in 2021,” *Nucleic Acids Research*, vol. 49, no. D1, pp. D480–D489, 11 2020. [Online]. Available: https://doi.org/10.1093/nar/gkaa1100

[2] M. Al Haddad, R. El-Rif, S. Hanna, L. Jaafar, R. Dennaoui, S. Abdellatef, V. Miskolci, D. Cox, L. Hodgson, M. El-Sibai, and et al., “Differential regulation of rho gtpases during lung adenocarcinoma migration and invasion reveals a novel role of the tumor suppressor stard13 in invadopodia regulation,” *Cell Communication and Signaling*, vol. 18, no. 1, 2020.

[3] J. Wei, D. Fang, and W. Zhou, “Ccr2 and ptprc are regulators of tumor microenvironment and potential prognostic biomarkers of lung adenocarcinoma,” *Annals of Translational Medicine*, vol. 9, no. 18, p. 1419–1419, 2021.

[4] C. Lebovitz, N. Wretham, M. Osooly, K. Milne, T. Dash, S. Thornton, B. Tessier-Cloutier, P. Sathiyaseelan, S. Bortnik, N. E. Go, and et al., “Loss of parkinson’s susceptibility gene lrrk2 promotes carcinogen-induced lung tumorigenesis,” *Scientific Reports*, vol. 11, no. 1, 2021.

[5] C. B. Lebovitz, A. G. Robertson, R. Goya, S. J. Jones, R. D. Morin, M. A. Marra, and S. M. Gorski, “Cross-cancer profiling of molecular alterations within the human autophagy interaction network,” *Autophagy*, vol. 11, no. 9, p. 1668–1687, 2015.

[6] Y. Wang, S. Qiu, H. Wang, J. Cui, X. Tian, Y. Miao, C. Zhang, L. Cao, L. Ma, X. Xu, and et al., “Transcriptional repression of ferritin light chain increases ferroptosis sensitivity in lung adenocarcinoma,” *Frontiers in Cell and Developmental Biology*, vol. 9, 2021.

[7] T. Qu, W. Zhang, L. Qi, L. Cao, C. Liu, Q. Huang, G. Li, L. Li, Y. Wang, Q. Guo, and et al., “Isg15 induces esrp1 to inhibit lung adenocarcinoma progression,” *Cell Death & Disease*, vol. 11, no. 7, 2020.

[8] L. Yoo, A.-R. Yoon, C.-O. Yun, and K. C. Chung, “Covalent isg15 conjugation to chip promotes its ubiquitin e3 ligase activity and inhibits lung cancer cell growth in response to type i interferon,” *Cell Death & Disease*, vol. 9, no. 2, 2018.

[9] F. Xue, Y. Jia, and J. Zhao, “Overexpression of fyn suppresses the epithelial-to-mesenchymal transition through down-regulating pi3k/akt pathway in lung adenocarcinoma,” *Surgical Oncology*, vol. 33, p. 108–117, 2020.

[10] B.-Q. Qiu, X.-H. Lin, S.-Q. Lai, F. Lu, K. Lin, X. Long, S.-Q. Zhu, H.-X. Zou, J.-J. Xu, J.-C. Liu, and et al., “Itgb1-dt/arntl2 axis may be a novel biomarker in lung adenocarcinoma: A bioinformatics analysis and experimental validation,” *Cancer Cell International*, vol. 21, no. 1, 2021.

[11] Y. Wang, S. Qiu, H. Wang, J. Cui, X. Tian, Y. Miao, C. Zhang, L. Cao, L. Ma, X. Xu, and et al., “Transcriptional repression of ferritin light chain increases ferroptosis sensitivity in lung adenocarcinoma,” *Frontiers in Cell and Developmental Biology*, vol. 9, 2021.

[12] M. Lucchetta, I. da Piedade, M. Mounir, M. Vabistsevits, T. Terkelsen, and E. Papaleo, “Distinct signatures of lung cancer types: Aberrant mucin o-glycosylation and compromised immune response,” *BMC Cancer*, vol. 19, no. 1, 2019.

[13] W. Luo, S. Tahara, K. Kawasaki, A. Kobayashi, S. Nojima, and E. Morii, “The expression of trefoil factor 3 is related to histologic subtypes and invasiveness in lung adenocarcinoma,” *Oncology Letters*, vol. 21, no. 1, 2020.

[14] X.-N. Wang, S.-J. Wang, V. Pandey, P. Chen, Q. Li, Z.-S. Wu, Q. Wu, and P. E. Lobie, “Trefoil factor 3 as a novel biomarker to distinguish between adenocarcinoma and squamous cell carcinoma,” *Medicine*, vol. 94, no. 20, 2015.

[15] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[17] M. Fey and J. E. Lenssen, “Fast graph representation learning with PyTorch Geometric,” in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[18] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2017.