MATTERS ARISING
https://doi.org/10.1038/s42256-022-00454-y

nature machine intelligence

Reusability report: Capturing properties of biological objects and their relationships using graph neural networks

Chenyang Hong[1,6], Qin Cao[1,2,6] ✉, Zhenghao Zhang[1], Stephen Kwok-Wing Tsui[2,3] and Kevin Y. Yip[1,3,4,5] ✉

ARISING FROM R. Schulte-Sasse et al. *Nature Machine Intelligence* https://doi.org/10.1038/s42256-021-00325-y (2021).

Graph neural networks (GNNs)[1], especially graph convolutional networks (GCNs)[2], have been increasingly used to model complex interactions. A basic idea behind GNNs is that some properties of an object, represented by a node in a graph, are reflected by properties of the objects that it directly and indirectly interacts with, where direct interactions are represented by edges in the graph. In biomedicine, GNNs have been used in a variety of applications, such as predicting protein functions and drug–disease associations[3]. Schulte-Sasse et al. have recently proposed a new use for GCNs in biomedicine: identifying cancer genes[4]. Their method, EMOGI (explainable multiomics graph integration), integrates multiomic data by aggregating information over a protein–protein interaction (PPI) network. The integrated information was shown to predict cancer genes better than having multiomic data alone or PPI connections alone.

Here we evaluate the reproducibility of the results reported by Schulte-Sasse et al. We also show that other biological networks can be used in place of the PPI network and demonstrate that the GCN approach can be used in yet another biomedical application: prediction of essential genes.

Reproducing the reported results

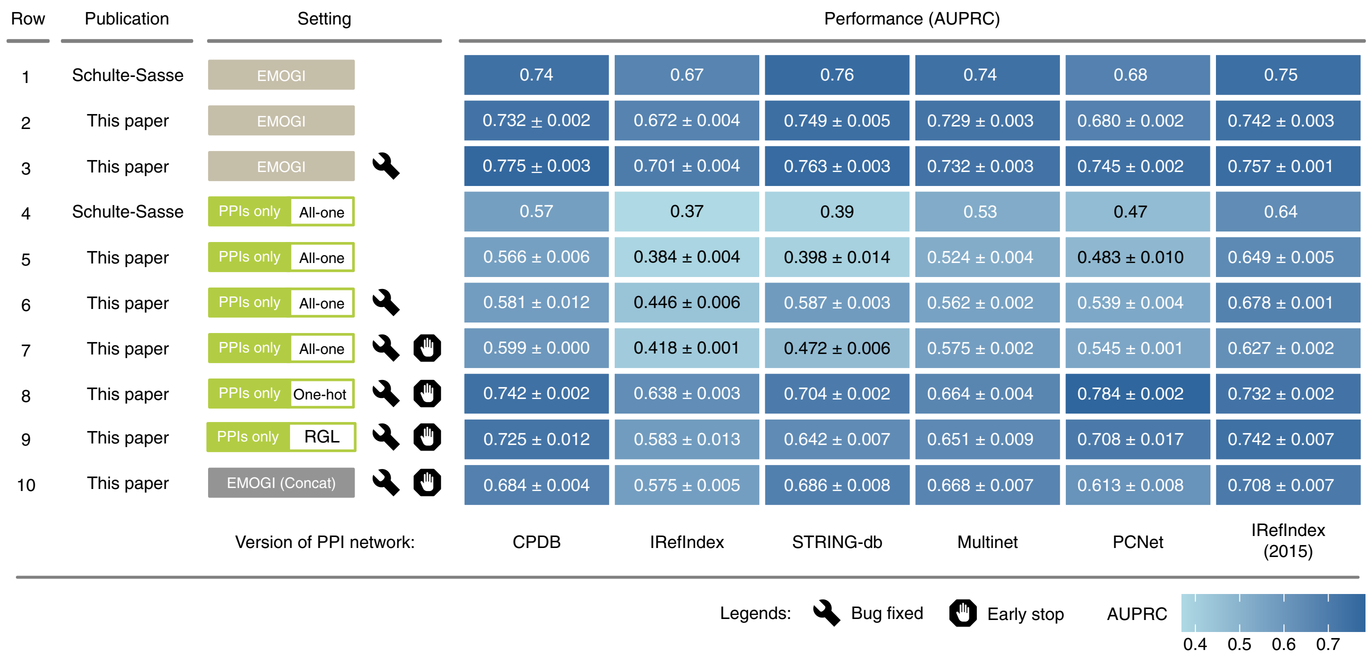
We downloaded the source code of EMOGI from https://github.com/schulter/EMOGI (git commit 5670b81) and re-trained the models on the given data sets according to the procedure described by Schulte-Sasse et al.[4]. For each experiment, we performed five independent runs with different random seeds and reported both the average performance and the standard deviation. For all six versions of the human PPI network, our average AUPRC values were close to those reported in the original paper, deviating by −1.486% to 0.298% of their reported values (Fig. 1, rows 1 and 2). The deviations could be due to the lack of fixed random seeds in the EMOGI code and the non-deterministic implementations of some graphics processing unit (GPU) functions in TensorFlow. We then inspected the downloaded code carefully and found a redundant sigmoid transformation added to the output layer. We confirmed with the original authors that this was indeed a bug. After fixing this issue, the performance of EMOGI was generally improved (Fig. 1, rows 2 and 3).

We next tested the prediction performance when only PPI information was available without the multiomic features. In the original code, this was achieved by shrinking the feature dimension to one and filling all the entries with the constant value of one. Our reproduced results were similar to those reported in the original paper (Fig. 1, rows 4 and 5). We also tested multiple technical variations of this setting, namely having the bug fix and/or using the early stopping strategy, which takes the best intermediate model during optimization based on the performance of a validation set separate from the final testing set (Fig. 1, rows 6 and 7). In general, the performance was improved in some of these variations, but not to the level when multiomic features were also used.

We argued that implementing the PPI-only setting by using an all-one feature matrix was not ideal because all nodes would become indistinguishable on the basis of their features and thus they would differ from each other only by their numbers of direct and indirect interaction partners. Therefore, we followed the original implementation of GCN to use a square identify matrix (that is, one-hot encoding) as the feature matrix instead[5], which gives a unique ID to each node directly. To avoid over-fitting, we applied the early-stopping strategy[6]. The results (Fig. 1, row 8) were substantially better than those with the all-one matrix (Fig. 1, rows 4–7), suggesting that some additional graph structural information contained in the PPI network other than interaction counts is useful for identifying cancer genes. Surprisingly, the performance of the PPI-only setting with the one-hot matrix implementation based on the PPI network from PCNet (AUPRC = 0.784) was even better than the EMOGI setting that involved both PPI and multiomic information (AUPRC = 0.745). We found that this version of the PPI network was 5 to 18 times denser than the other five versions (Supplementary Table 1), which might have made it outstandingly informative.

An alternative way to allow the nodes to be distinguishable is to create a high-dimensional random node feature matrix, such as one with values sampled from a standard Gaussian distribution, which is conceptually related to the one-hot encoding[7]. We tested this random Gaussian initialization approach, with the feature matrix either fixed (RGF) or learnable (RGL) during the process of learning the

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. [2]School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. [3]Hong Kong Bioinformatics Centre, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. [4]Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. [5]Present address: Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, USA. [6]These authors contributed equally: Chenyang Hong, Qin Cao. ✉e-mail: qcao@cuhk.edu.hk; kevinyip@cse.cuhk.edu.hk



Row	Publication	Setting			Performance (AUPRC)					
		Version of PPI network:			CPDB	IRefIndex	STRING-db	Multinet	PCNet	IRefIndex (2015)
1	Schulte-Sasse	EMOGI			0.74	0.67	0.76	0.74	0.68	0.75
2	This paper	EMOGI			0.732 ± 0.002	0.672 ± 0.004	0.749 ± 0.005	0.729 ± 0.003	0.680 ± 0.002	0.742 ± 0.003
3	This paper	EMOGI	🔧		0.775 ± 0.003	0.701 ± 0.004	0.763 ± 0.003	0.732 ± 0.003	0.745 ± 0.002	0.757 ± 0.001
4	Schulte-Sasse	PPIs only All-one			0.57	0.37	0.39	0.53	0.47	0.64
5	This paper	PPIs only All-one			0.566 ± 0.006	0.384 ± 0.004	0.398 ± 0.014	0.524 ± 0.004	0.483 ± 0.010	0.649 ± 0.005
6	This paper	PPIs only All-one	🔧		0.581 ± 0.012	0.446 ± 0.006	0.587 ± 0.003	0.562 ± 0.002	0.539 ± 0.004	0.678 ± 0.001
7	This paper	PPIs only All-one	🔧	✋	0.599 ± 0.000	0.418 ± 0.001	0.472 ± 0.006	0.575 ± 0.002	0.545 ± 0.001	0.627 ± 0.002
8	This paper	PPIs only One-hot	🔧	✋	0.742 ± 0.002	0.638 ± 0.003	0.704 ± 0.002	0.664 ± 0.004	0.784 ± 0.002	0.732 ± 0.002
9	This paper	PPIs only RGL	🔧	✋	0.725 ± 0.012	0.583 ± 0.013	0.642 ± 0.007	0.651 ± 0.009	0.708 ± 0.017	0.742 ± 0.007
10	This paper	EMOGI (Concat)	🔧	✋	0.684 ± 0.004	0.575 ± 0.005	0.686 ± 0.008	0.668 ± 0.007	0.613 ± 0.008	0.708 ± 0.007

Fig. 1 | Reproducing the main reported performance results of EMOGI. The area under the precision-recall curve (AUPRC) values under different settings are reported. For each experiment conducted in this study, the average value among different runs and the standard deviation are shown. 'Bug fixed' is the code version with the redundant sigmoid transformation removed. 'Early stop' indicates that training is stopped when the performance is not improved in the evaluation for ten successive epochs; otherwise 5,000 epochs are run during training when the early stop option is not chosen. 'All-one' means all values in the feature matrix are constant values of one. 'One-hot' means the feature matrix is the identity matrix, that is, each node is one-hot encoded. 'RGL' indicates that the feature matrix is a trainable matrix with 512 dimensions initialized by sampling from the Gaussian distribution. EMOGI(Concat) is a setting with the multiomic features concatenated to the identity matrix to become the final feature matrix.

node embeddings, with different lengths of the feature vectors (Fig. 1, row 9, and Supplementary Fig. 1). Cancer gene prediction performance with RGL was consistently higher than that with RGF. Among the six versions of the PPI network, the prediction performance with the one-hot encoding was higher than that with RGL in five of them (Fig. 1, rows 8 and 9), but the performance of RGL shows an increasing trend when the length of the feature vector increases (Supplementary Fig. 1).

Considering the better performance of our new implementation of the PPI-only setting using the one-hot feature matrix, we next tested whether concatenating the multiomic features could further improve the performance of modelling cancer genes, but the results showed that no further improvements could be achieved (Fig. 1, row 10).

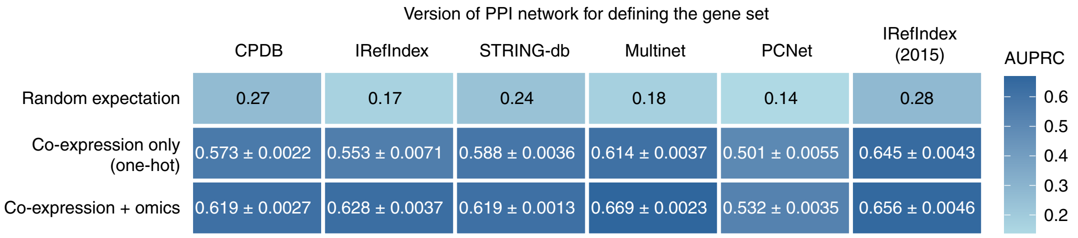
Due to the improved performance, we always use the fixed version of the EMOGI code for the remainder of this paper.

Predicting cancer genes using co-expression network

In order for a GNN to be useful, the interaction network should connect objects that are related to each other in the context of the target application, usually in the form of positive reinforcement. For the application of predicting cancer genes, a plausible reason for using PPIs to connect genes (based on the proteins that they code for) is that proteins that physically interact with each other may belong to certain common pathways, and that if one member in a pathway plays a crucial role in cancer, other members may also play crucial roles. Following this logic, other interaction networks that connect genes potentially belonging to the same pathways may also be useful for predicting cancer genes. One such candidate is the co-expression network, in which two genes are connected if they have strong co-expression across a large number of samples. The basic idea is that if two genes co-express, there could be a reason that their gene products should be available at the same time, which may suggest that they are functionally related.

To test if the co-expression network is useful for predicting cancer genes, we downloaded human gene co-expression data from COXPRESdb[8] (https://coxpresdb.jp/download/Hsa-u.c2-0/coex/). Using a mutual rank threshold, we produced a master co-expression network. In order to directly compare with the PPI results, we then produced six different versions of the co-expression network by keeping only genes in the master co-expression network that are also present in the six versions of the PPI network, respectively. We then modelled cancer genes with both the co-expression network and multiomic features together or with the co-expression network only. Hyper-parameters were tuned using a grid-search based on the validation-set performance of the co-expression network with genes from the CPDB PPI network.

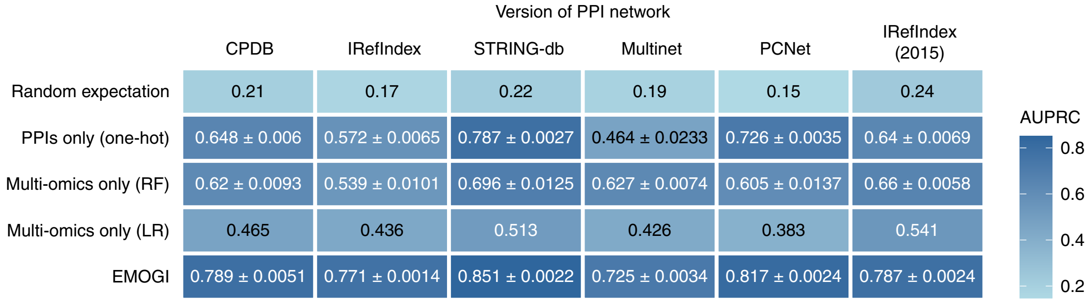
From the results (Fig. 2), when taking the co-expression network as input, the performance of EMOGI in predicting cancer genes was not as good as taking the PPI network as input (Fig. 1), but it was still much better than a random predictor (which would have an expected AUPRC value equal to the ratio of cancer genes to the total number of cancer genes and non-cancer genes). For example, when EMOGI was given both the co-expression network and multiomic features, the average AUPRC values were 0.532–0.669 with the co-expression network and 0.701–0.775 with the PPI network. These results are expected because two genes can co-express due to reasons other than belonging to the same pathway, such as being in two different pathways both downstream of a third pathway. In addition, our choice of a fairly large mutual rank threshold caused some genes with not very strong co-expression to be also connected. Should we have chosen a smaller threshold, the network edges would represent stronger co-expression gene pairs, but at the same time the number of edges in the network would reduce,



Version of PPI network for defining the gene set

	CPDB	IRefIndex	STRING-db	Multinet	PCNet	IRefIndex (2015)
Random expectation	0.27	0.17	0.24	0.18	0.14	0.28
Co-expression only (one-hot)	0.573 ± 0.0022	0.553 ± 0.0071	0.588 ± 0.0036	0.614 ± 0.0037	0.501 ± 0.0055	0.645 ± 0.0043
Co-expression + omics	0.619 ± 0.0027	0.628 ± 0.0037	0.619 ± 0.0013	0.669 ± 0.0023	0.532 ± 0.0035	0.656 ± 0.0046

Fig. 2 | Performance (AUPRC values) of predicting cancer genes using EMOGI but with its PPI network replaced by co-expression network. Each entry in the last two rows shows the average AUPRC and standard deviation across the five repeated runs with different random seeds.



Version of PPI network

	CPDB	IRefIndex	STRING-db	Multinet	PCNet	IRefIndex (2015)
Random expectation	0.21	0.17	0.22	0.19	0.15	0.24
PPIs only (one-hot)	0.648 ± 0.006	0.572 ± 0.0065	0.787 ± 0.0027	0.464 ± 0.0233	0.726 ± 0.0035	0.64 ± 0.0069
Multi-omics only (RF)	0.62 ± 0.0093	0.539 ± 0.0101	0.696 ± 0.0125	0.627 ± 0.0074	0.605 ± 0.0137	0.66 ± 0.0058
Multi-omics only (LR)	0.465	0.436	0.513	0.426	0.383	0.541
EMOGI	0.789 ± 0.0051	0.771 ± 0.0014	0.851 ± 0.0022	0.725 ± 0.0034	0.817 ± 0.0024	0.787 ± 0.0024

Fig. 3 | Performance (AUPRC values) of predicting essential genes using EMOGI and baseline methods. Each entry in the second, third and fifth rows shows the average AUPRC and standard deviation across the five repeated runs with different random seeds. Because the implementation of logistic regression that we used is deterministic without using any random seeds, we only ran it once. LR, logistic regression; RF, random forest.

which could also hamper prediction performance since the overall network would become more fragmented (Supplementary Table 2).

Nevertheless, this experiment does demonstrate the flexibility of the GCN framework, that the interaction network can be easily replaced by another one.

Predicting essential genes

Essential genes are genes whose loss-of-function perturbations have detrimental effects, which can happen at different levels, such as lowering production of critical metabolites and reducing cell survival. There are existing methods that aim at predicting gene essentiality using features such as those derived from sequences and protein interactions[9]. In the original paper, Schulte-Sasse et al. show that their newly predicted cancer genes (NPCGs) tend to be essential genes in cancer cell lines[4]. Due to this tendency, we hypothesized that information contained in the multi-omic features and PPI network captured by EMOGI could also predict essential genes directly.

To test it, we downloaded 16 human essential gene datasets from the DEG database[10], which contained 8,256 unique genes in total. Following a previous study[9], we took the genes contained in at least five datasets as the positive examples of essential genes and genes that were not included in any of the 16 datasets as the negative examples. We then used EMOGI to predict essential genes, with the genes not contained in the PPI network filtered in each case (Supplementary Table 3), based on a random division of the resulting genes into a left-out testing set (10%), a training set and a validation set (90% and 10% of the remaining genes, respectively). A grid search was performed to tune the hyper-parameters based on results of the validation set.

The results show that EMOGI achieved good performance in predicting essential genes (Fig. 3), with average AUPRC values ranging from 0.725 to 0.851. The prediction performance was consistently better when both the PPI network and multi-omic features were provided as compared to having only one of them (Fig. 3), again demonstrating the advantage of aggregating features over network neighbourhoods using GCN.

Graph attention networks

EMOGI was developed on top of GCNs. With the rapid development of graph learning methods, there are other GNN structures that have been shown to outperform GCNs in various applications (reviewed previously[11]), such as GraphSAGE (Graph SAmple and aggreGatE)[12], graph attention networks (GATs)[13] and graph isomorphism networks (GINs)[14]. Among them, GATs were shown to be consistently one of the best performers. The key idea behind GATs is the attention mechanism, which enables the assignment of different weights to different nodes in a neighbourhood.

To test if cancer genes can be predicted more accurately using GATs than GCNs, we replaced the GCN structure of EMOGI by a GAT structure. We used two different implementations of GAT, namely the TensorFlow implementation proposed in the original GAT paper[13] and the PyTorch implementation in Deep Graph Library (DGL)[15]. We repeated each experiment 10 times using different random seeds and report the whole distribution of performance values. Due to the long training time of the original TensorFlow implementation of GAT, in each run we could only train one model for each version of the PPI network, instead of using the original ensemble approach of EMOGI, which involved 10 models obtained by ten-fold cross-validation, and we only performed the tests using three versions of the PPI network.

The comparison results (Fig. 4) show that the performance of GATs based on the original TensorFlow implementation varied quite substantially across different runs. In comparison, the DGL implementation of GAT led to more accurate and stable predictions. In general, the performance of both implementations of GAT was lower than that of GCNs and the performance gap was smaller in terms of AUROC (area under the receiver–operator characteristics),

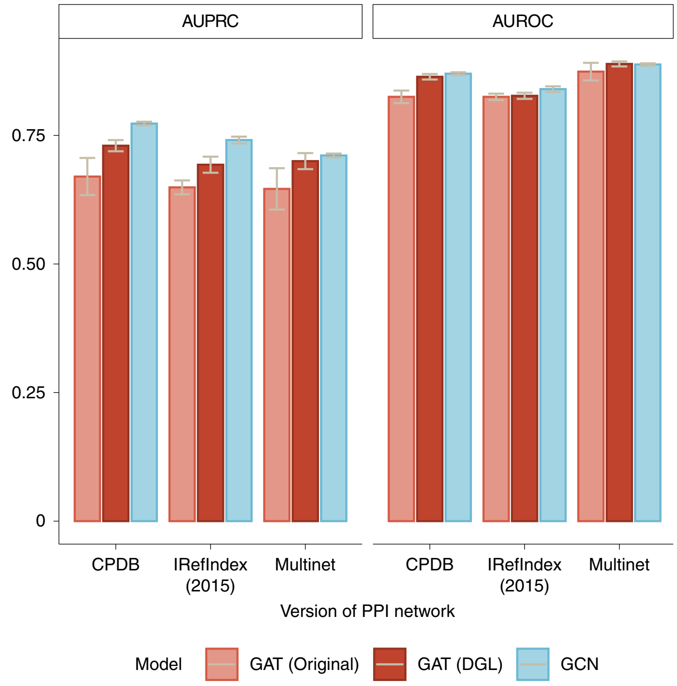


Fig. 4 | Comparison of GATs and GCNs in the prediction of cancer genes. Each error bar represents the standard deviation of the performance measure across all the runs based on that version of the PPI network. GAT (original): the TensorFlow implementation proposed in the original paper. GAT (DGL): the PyTorch implementation in DGL.

which is consistent with the results of some previous studies[16,17]. The different performance gaps of AUPRC and AUROC may be due to the characteristics of AUPRC, that it is more informative and sensitive than AUROC when the sizes of the positive and negative sets are imbalanced[18].

Discussion

In this study, we have shown that using the code and data provided by Schulte-Sasse et al., we were able to largely reproduce their main published results[4]. By fixing a programming issue and changing the encoding of network nodes, we were able to improve the reproduced results by 0.412–9.559% (rows 2–3 in Fig. 1) and 7.965–45.455% (rows 6–8), respectively. The latter result was particularly interesting, that by using information contained in the PPI network alone, cancer genes can be predicted with high accuracy even without considering multiomic features of the genes.

Conceptually, the one-hot encoding enables GNNs to memorize node IDs and thus distinguish between nodes with similar graph structural contexts[19]. It has led to good predictive power in social network applications without additional node features[20]. On the other hand, it also creates difficulties in applying the model when new nodes not seen during the training stage are introduced[19], which is not a major issue when the node set is (largely) fixed as in the case of protein-coding genes. Besides, when node features can also largely identify the nodes, the one-hot encoding would become redundant. This could explain why concatenating one-hot encoding and node features may not boost the performance in some tasks[21].

In addition to reproducing the results of Schulte-Sasse et al., we have demonstrated the flexibility of GCNs in terms of both replacing the interaction network (from PPIs to co-expression) and changing the prediction target (from cancer genes to essential genes). Since there are various biological networks available, such as chromatin interactions and transcription factor binding, and there are various interesting prediction targets, such as replication timing and gene function, it would be instructive to systematically study how individual networks and their integration, in the form of a heterogeneous network[22], perform in these predictions.

Our study also revealed practical difficulties in using GATs, at least the TensorFlow implementation of it that we used, in terms of its long training time and irreproducible results in our prediction tasks. The performance of GATs also seems to be fairly sensitive to values of its hyper-parameters. In general, GATs could be useful in some applications but there are apparently some specific prerequisites that are still not well characterized.

Data availability

The data used in our study are available at https://github.com/kevingroup/emogi-reusability[23]. All the data used in the original paper by Schulte-Sasse et al. for testing EMOGI are available at http://owww.molgen.mpg.de/~sasse/EMOGI/.

Code availability

The original EMOGI code is available at https://github.com/schulter/EMOGI. Our modified GAT-based version of it is available at https://github.com/kevingroup/emogi-reusability[23].

Received: 13 August 2021; Accepted: 1 February 2022;
Published online: 17 March 2022

References

1. Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M. & Monfardini, G. The graph neural network model. *IEEE Trans. Neural Networks* **20**, 61–80 (2009).
2. Micheli, A. Neural network for graphs: a contextual constructive approach. *IEEE Trans. Neural Networks* **20**, 498–511 (2009).
3. Yue, X. et al. Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics* **36**, 1241–1251 (2020).
4. Schulte-Sasse, R., Budach, S., Hnisz, D. & Marsico, A. Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. *Nat. Mach. Intell.* **3**, 513–526 (2021).
5. No node features. *GitHub* https://github.com/tkipf/gcn/issues/10 (3 March 2017).
6. Prechelt, L. In *Neural Networks: Tricks of the Trade* 55–69 (Springer, 1998).
7. Cui, H., Lu, Z., Li, P. & Yang, C. On positional and structural node features for graph neural networks on non-attributed graphs. In *Proc. Workshop of Deep Learning on Graphs: Methods and Applications, The 27th International ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2021).
8. Obayashi, T., Kagaya, Y., Aoki, Y., Tadaka, S. & Kinoshita, K. COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. *Nucleic Acids Res.* **47**, D55–D62 (2019).
9. Zhang, X., Xiao, W. & Xiao, W. DeepHE: accurately predicting human essential genes based on deep learning. *PLoS Comput. Biol.* **16**, e1008229 (2020).
10. Luo, H., Lin, Y., Gao, F., Zhang, C.-T. & Zhang, R. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Res.* **42**, D574–D580 (2014).
11. Wu, Z. et al. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Networks Learn. Syst.* **32**, 4–24 (2020).
12. Hamilton, W. L., Ying, R. & Leskovec, J. Inductive representation learning on large graphs. In *Proc. 31st Int. Conf. Neural Information Processing Systems* 1025–1035 (2017).
13. Veličković, P. et al. Graph attention networks. In *6th Int. Conf. Learning Representations* (2018).
14. Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How powerful are graph neural networks? In *7th Int. Conf. Learning Representations* (2019).
15. Wang, M. et al. Deep graph library: A graph-centric, highly-performant package for graph neural networks. Preprint at https://arxiv.org/abs/1909.01315 (2019).
16. Fanfani, V., Torne, R. V., Lio', P. & Stracquadanio, G. Discovering cancer driver genes and pathways using stochastic block model graph neural networks. Preprint at *bioRxiv* https://doi.org/10.1101/2021.06.29.450342 (2021).
17. Schulte-Sasse, R.,Budach, S., Hnisz, D. & Marsico, A. Graph convolutional networks improve the prediction of cancer driver genes. In *Int. Conf. Artificial Neural Networks* 658–668 (2019).
18. Davis, J. & Goadrich, M. The relationship between precision-recall and roc curves. In *Proc. 23rd Int. Conf. Machine Learning* 233–240 (2006).
19. You, J., Ying, R. & Leskovec, J. Position-aware graph neural networks. In *Int. Conf. Machine Learning* 7134–7143 (2019).

20. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. In *5th Int. Conf. Learning Representations* (2017).
21. About the data's feature. *GitHub* https://github.com/tkipf/gcn/issues/22 (5 February 2018).
22. Cao, Q. et al. A unified framework for integrative study of heterogeneous gene regulatory mechanisms. *Nat. Mach. Intell.* **2**, 447–456 (2020).
23. Hong, C., Cao, Q. & Zhang, Z. EMOGI-reusability v1.0 https://doi.org/10.5281/zenodo.5914506 (2022).

Acknowledgements

We thank Schulte-Sasse et al. for their help in our running of EMOGI and for sharing their processed data. Q.C. is supported by National Natural Science Foundation of China Youth Program 32100515. K.Y.Y. was supported by Hong Kong Research Grants Council Collaborative Research Funds C4015-20E, C4045-18W, C4057-18E and C7044-19G and General Research Funds 14107420 and 14203119, the Hong Kong Epigenomics Project (EpiHK), and the Chinese University of Hong Kong Young Researcher Award, Outstanding Fellowship and Project Impact Enhancement Fund.

Author contributions

Q.C. and K.Y.Y. conceived and supervised the project. C.H., Q.C., Z.Z., S.K.T. and K.Y.Y. designed the computational experiments and data analyses. C.H. and Z.Z. prepared the data. C.H. and Q.C. implemented the methods, conducted the experiments and performed data analyses. C.H., Q.C. and K.Y.Y. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-022-00454-y.

Correspondence and requests for materials should be addressed to Qin Cao or Kevin Y. Yip.

Peer review information *Nature Machine Intelligence* thanks Marinka Zitnik for her contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2022