Supplementary Materials for

"Graph Contextualized Attention Network for Predicting Synthetic Lethality in Human Cancers"

1. Feature extraction for genes

For genes, we extracted 8 pair-wise features from different genres of biological data and 10 node-wise network features from PPI network. Specifically, we first downloaded the ontology and annotation files from http://geneontology.org/. Then we calculated three semantic similarity matrices for genes based on the sub-ontologies "biological process (BP)", "molecular function" and "cellular component (CC)", using the method proposed by Wang et al. (2007). We further downloaded the PPI data from BioGrid to construct a PPI network. Note that we removed all the SL pairs curated in this PPI network constructed from BioGrid (Oughtred et al., 2019). Besides, we also constructed 4 features for each SL pair, derived from four sources: Pathway Co-membership, using the Canonical pathway database from Broad Institute's Molecular Signatures Database (MSigDB) (Subramanian et al., 2005); Protein Complex Co-membership, using the CORUM protein complex database (Giurgiu et al., 2018); Protein interaction scores, using human protein-protein interaction database (Hippie) (Gregorio et al., 2017); Protein top similarity, using human protein reference database (HPRD) (Prasad et al., 2009).

Node-wise network features were calculated based on the PPI network constructed from BioGrid. They included degree, closeness, betweenness, eigenvector centrality and clustering. Table S1 shows the name and description for each network feature.

Table S1. Names and descriptions of node-wise network features.

Name	**Type**	**Description**
BP	Pairwise	The number of biological process GO annotations shared between the source and target node.
MF	Pairwise	The number of molecular function GO annotations shared between the source and target node.
CC	Pairwise	The number of cellular component GO annotations shared between the source and target node.
Co-pathway	Pairwise	The number of protein pathways shared between the source and target node.
Co-complex	Pairwise	The number of protein complexes shared between the source and target node.
Protein score	Pairwise	A value to measure how well associated a given node is with the other node.
Protein top similarity	Pairwise	A value to measure the structure similarity between the source and target node.
PPI	Pairwise	A binary matrix recording whether a give node is confirmed to be associated with the other node.
Degree	Node-wise	The number of edges coming in to or out of the node.
Closeness	Node-wise	The number of steps required to reach all other nodes from a given node.
Betweenness	Node-wise	The number of shortest paths in the entire graph that pass through the node.
Eigenvector	Node-wise	A measure of how well connected a given node is to other well-connected nodes.
Clustering	Node-wise	The clustering coefficient of the node.

2. Comparison performance between our model with 14 state-of-the-art methods

In this work, for a fair comparison, we also conducted experiments to compare our proposed GCATSL model with four representation learning-based baseline methods with all unknown pairs as negative samples. The results on SyLethDB and SyLethDB-v2.0 have been shown in Table S2. We can observe that our proposed model consistently outperforms baseline methods in terms of AUC and AUPR on both datasets.

Table S2. Comparison performance between our model and four representation learning-based methods under the setting of using all unknown pairs as negative samples.

Method	SyLethDB		SyLethDB-v2.0	
	AUC	AUPR	AUC	AUPR
CMF	0.8918±0.0044	0.9066±0.0037	0.9016±0.0054	0.9278±0.0038
SL^2MF	0.8448±0.0052	0.8979±0.0042	0.7879±0.0031	0.8664±0.0016
GRSMF	0.8905±0.0050	0.9232±0.0027	0.9126±0.0016	0.9258±0.0011
DDGCN	0.8775±0.0049	0.9149±0.0032	0.8465±0.0067	0.8982±0.0041
GCATSL	**0.9286±0.0036**	**0.9385±0.0091**	**0.9266±0.0054**	**0.9378±0.0038**

3. Case study

In this work, we conducted case study to further validate the effectiveness of our model. In the experiment, we utilized all known SL pairs as positive samples to train our model, and prioritized all SL pairs according to their scores. We evaluate our model by checking how many unknown SL pairs among the top 1000 pairs are reported in SynLethDB-v2.0 and supported by biomedical literature. Table S3 displays the 36 SL pairs which are supported by previous literature.

Table S3. 36 confirmed SL pairs by previous literature among the top-1000 predicted SL pairs.

No.	Gene1	Gene2	Pubmed ID	Source
1	BCR	KRAS	27655641	in-silico prediction
2	DDR1	KRAS	24104479	shRNA screening
3	KRAS	RET	27655641	in-silico prediction
4	CMPK1	KRAS	24104479	shRNA screening
5	MYC	NTRK1	22623531	siRNA screening
6	BRCA1	KRAS	24104479	shRNA screening
7	KRAS	PIK3CA	26627737	CRISPR-Cas9
8	CHEK1	KRAS	27655641	in-silico prediction
9	KRAS	TBL1XR1	28700943	CRISPR screening
10	CYP1B1	KRAS	22613949	siRNA screening
11	KRAS	SSBP1	28700943	CRISPR
12	KRAS	MAPK1	26627737	CRISPR-Cas9
13	E2F1	KRAS	22613949	siRNA screening
14	EZH2	KRAS	25407795	RNAi screening
15	KRAS	WRAP53	28700943	CRISPR screening
16	KRAS	RPL13A	22613949	siRNA screening
17	CDC7	KRAS	27655641	in-silico prediction

18	ABL1	PDGFRB	26637171	siRNA screening
19	KRAS	POLR2A	22613949	siRNA screening
20	KIT	PDGFRB	26637171	siRNA screening
21	BID	KRAS	24104479	shRNA screening
22	KRAS	NHP2	28700943	CRISPR screening
23	KRAS	SSH3	24104479	shRNA screening
24	ABL1	KIT	26637171	siRNA screening
25	NTRK1	PDGFRB	26637171	siRNA screening
26	KIT	PDGFRA	31300006	in-silico prediction
27	KRAS	MSH2	27655641	in-silico prediction
28	KRAS	SRP9	28700943	CRISPR screening
29	KRAS	MCM2	24104479	shRNA screening
30	KRAS	SKP2	27655641	in-silico prediction
31	KRAS	LUC7L2	28700943	CRISPR screening
32	KRAS	TMED2	28700943	CRISPR screening
33	KRAS	RPS6KB1	27655641	in-silico prediction
34	KRAS	MAPRE1	24104479	shRNA screening
35	CDK1	KRAS	26881434	siRNA screening
36	ATP6V1C1	KRAS	24104479	shRNA screening

Besides, we compared our model with 5 state-of-the-art methods by observing the number of SL pairs supported by SynLethDB-v2.0 among top-r predicted SL pairs. We selected r from 1000 to 20000 with a step size of 1000. Fig. S1 shows our model performs better than baseline methods. In particular, our model outperforms significantly baseline methods from top 6000 to 20000. Therefore, we can conclusion that our model is an effective and promising tool in identifying potential SL pairs.

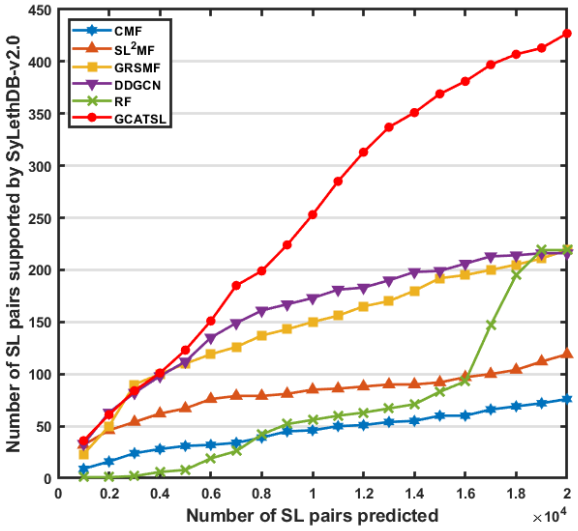

Fig. S1. Performance comparison between our model with 5 state-of-the-art methods in identifying potential SL pairs.

Reference

Wang, J. Z. *et al.* (2007). A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10), 1274–1281.

Prasad, T. S. K. *et al.* (2009). Human Protein Reference Database - 2009 Update. *Nucleic Acids Research*. 37, D767-D772.

Gregorio, A. L. *et al.* (2017). HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Research.* 45, D408-D414.

Giurgiu, M. *et al.* (2018). CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Research*. 47(D1), D559–D563.

Subramanian, A. *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. In: *Proceedings of the National Academy of Sciences*, 102(43), 15545–15550.

Oughtred, R. et al. (2019). The BioGRID interaction database: 2019 update. *Nucleic acids research* 47(D1), D529-D541.