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

2.1 Results on SynLethDB and SynLethDB-v2.0

In this work, for better comparison, in addition to AUC and AUPR, we also evaluate the performance of various methods using metric Recall@k. This metric is frequently used in other fields, such as recommendation systems (Wu et al., 2019). Table S2 shows the results of Recall@k (k=1000 and k=5000) on SynLethDB and SynLethDB-v2.0 under "1:1 setting", which keeps almost consistent with that of AUC and AUPR recorded in Table 2 in the manuscript. It should be noted that "1:1 setting" refers to the setting of using the same numbers of positive and negative samples for model training and testing. Negative SL pairs are randomly sampled from unknown pairs except for special instructions.

Representation learning methods (e.g., CMF and GRSMF) use all unknown pairs as negatives. For a fair comparison, we also conducted experiments to compare our proposed GCATSL model with four representation learning-based baseline methods under "All unknown setting". "All unknown setting" refers to the setting of using all unknown pairs as negatives. The results on SyLethDB and SyLethDB-v2.0 have been shown in Table S3. We can observe that our proposed model performs better than baseline methods on both datasets in terms of most of metrics.

In datasest SynLethDB and SynLethDB-v2.0, unknown SL pairs may include a gene that can be a SL partner or may have two genes that are not involved in any known SL pairs. To test both cases, we define negative SL pairs from DepMap (https://depmap.org/). In total, we extracted 275,557 gene pairs for 6375 genes in SynLethDB according to co-dependency coefficients between genes. Table S4 displays the performance of various methods under two different settings. Our model consistently outperforms five baseline methods. Meanwhile, we note that negative SL pairs extracted from DepMap can improve the performance of various methods including GRSMF, MetaSL and our GCATSL, demonstrating that DepMap can provide valuable genetic co-dependency information to define high-quality negative SL data.

Table S2. Comparison performance between our model and baseline methods on datasets SynLethDB and SynLethDB-v2.0 in terms of Recall@1000 and Recall@5000 under "1:1 setting".

Method	SynLethDB		SynLethDB-v2.0	
	R@1000	R@5000	R@1000	R@5000
CMF	0.2336	0.7997	0.1096	0.5121
SL^2MF	0.2512	0.8549	<u>0.1360</u>	0.6401
GRSMF	0.2508	<u>0.9229</u>	0.1361	<u>0.6745</u>
DDGCN	0.2499	0.8296	0.1161	0.5325
RF	0.2511	0.8627	0.1327	0.6087
DT	0.2247	0.8458	0.1264	0.5978
NB	0.2239	0.7697	0.1225	0.5120
SVM	0.2393	0.7927	0.1279	0.5324
KNN	0.2250	0.7896	0.1153	0.5115
Bagging	0.2498	0.8674	0.1314	0.6073
AdaBoost	0.2515	0.8194	0.1346	0.5480
GradientBoost	0.2530	0.8474	0.1351	0.5732
MNMC	0.2515	0.8560	0.1345	0.5731
MetaSL	<u>0.2528</u>	0.8736	0.1352	0.6067
GCATSL	**0.2568**	**0.9329**	**0.1422**	**0.6886**

Table S3. Comparison performance between our model and baseline methods on datasets SynLethDB and SynLethDB-v2.0 under "All unknown setting".

Method	SynLethDB				SynLethDB-v2.0			
	AUC	AUPR	R@1000	R@5000	AUC	AUPR	R@1000	R@5000
CMF	0.7240	0.0556	0.0605	0.1716	0.6921	0.0354	0.0216	0.1867
SL^2MF	0.8429	0.4369	0.2277	0.5011	0.7861	0.2970	0.0972	0.3339
GRSMF	**0.9243**	0.5351	0.2449	0.5580	0.9065	0.3260	**0.1327**	0.3469
DDGCN	0.8753	0.4883	0.1695	0.5693	0.8514	0.2776	0.0787	0.3176
GCATSL	0.9129	**0.5657**	**0.2517**	**0.5719**	**0.9136**	**0.3487**	0.1285	**0.3542**

Table S4. Comparison performance between our model and baseline methods on dataset SynLethDB with negative SL pairs defined by DepMap.

Method	1:1 setting				All unknown setting			
	AUC	AUPR	R@1000	R@5000	AUC	AUPR	R@1000	R@5000
CMF	0.8215	0.8441	0.2435	0.8691	0.9147	0.7125	0.2239	0.5854
SL^2MF	0.8432	0.8976	0.2540	0.8536	0.8448	0.7160	0.2512	0.6807
GRSMF	0.9284	0.9434	0.2536	0.9188	0.9302	0.5614	0.2510	0.5629
DDGCN	0.8782	0.9152	0.2358	0.8326	0.8775	0.7621	0.2444	0.5835
MetaSL	0.9092	0.9173	0.2529	0.9185	-	-	-	-
GCATSL	**0.9535**	**0.9556**	**0.2594**	**0.9576**	**0.9506**	**0.8000**	**0.2580**	**0.7948**

2.2 Results on Breast Cancer data

In this paper, to demonstrate the validity of our proposed model on specific cancer data, we perform GCATSL and four representation learning-based methods and the best feature-based method, i.e., MetaSL, on breast cancer data. Table S5 displays the comparison results of different methods under two different settings, from which we can find that our proposed model achieves better performance in most cases, demonstrating that GCATSL can be successfully applied for specific cancer type.

Table S5. Comparison performance between our model and five baseline methods on breast cancer-specific dataset under two different settings.

Method	1:1 setting				All unknown setting			
	AUC	AUPR	R@100	R@200	AUC	AUPR	R@1000	R@5000
CMF	0.7287	0.7284	0.2702	0.8474	0.6076	0.0242	0.1540	0.2398
SL^2MF	0.6203	0.6838	0.5135	0.7933	0.5670	0.0090	0.1113	0.3132
GRSMF	0.8702	0.9119	0.7504	0.9031	0.8600	**0.0549**	0.5060	0.7214
DDGCN	0.7975	0.8150	0.4586	0.8494	0.5745	0.0403	0.1026	0.1082
MetaSL	0.9103	0.9151	**0.7650**	0.9602	-	-	-	-
GCATSL	**0.9250**	**0.9226**	0.7593	**0.9730**	**0.9020**	0.0472	**0.5772**	**0.7351**

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 S6 displays the 36 SL pairs which are supported by previous literature.

Table S6. 36 confirmed SL pairs by SynLethDB-v2.0 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.

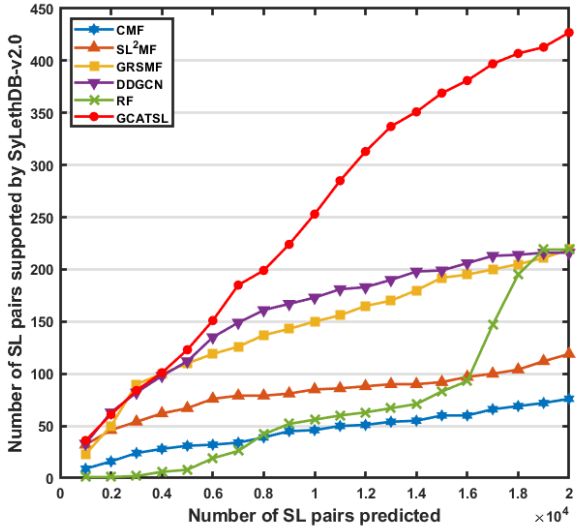

Fig. S1. Performance comparison between our model with 5 state-of-the-art methods in identifying potential SL pairs.

In addition, we conducted the second case study to further validate the effectiveness of our proposed model. More specifically, following Deng et al. (2019), we selected 10 genes as study objects, including BRCA1, BRCA2, TP53, PTEN, ATM, ATR, KRAS, HRAS and BRAF. We used all known SL pairs in SynLethDB to train our model. For each selected gene, we prioritized all its unknown pairs according to their prediction scores and calculated how many pairs among the top-100 and -500 predicted SL pairs can be validated by DepMap data and SynLethDBv2.0. As shown in Table S7 and S8, a total of 82 and 332 SL pairs could be successfully confirmed by DepMap and SynLethDB-v2.0 among the top-100 and -500 predicted SL pairs for these 10 genes. Note that the fourth column in Table S8 displays the number of SL pairs simultaneously validated by both DepMap and SynLethDB-v2.0.

Table S7. 78 confirmed SL pairs by database DepMap and SynLethDB-v2.0 among the top-100 predicted SL pairs for 10 selected genes.

No.	Gene1	Gene2	Source	No.	Gene1	Gene2	Source
1	ATM	SLC29A2	DepMap	40	EGFR	CCND1	SynLethDB-v2.0
2	ATM	MDM4	DepMap	41	HRAS	IRF7	DepMap
3	ATM	USP7	DepMap	42	KRAS	PLEK2	DepMap
4	ATM	MYBL2	DepMap	43	KRAS	TEX10	SynLethDB-v2.0
5	ATM	CD63	DepMap	44	KRAS	NFYB	SynLethDB-v2.0

6	ATR	TAF9	DepMap	45	KRAS	CEP57	SynLethDB-v2.0
7	ATR	PSMD12	DepMap	46	KRAS	ITGA3	SynLethDB-v2.0
8	ATR	RANBP3	DepMap	47	KRAS	VRK3	SynLethDB-v2.0
9	ATR	TOPBP1	DepMap	48	KRAS	ZNF83	SynLethDB-v2.0
10	ATR	LIG1	SynLethDB-v2.0	49	KRAS	PSMB3	SynLethDB-v2.0
11	ATR	SKP2	SynLethDB-v2.0	50	KRAS	BCAS2	SynLethDB-v2.0
12	BRAF	TP53	DepMap; SynLethDB-v2.0	51	PTEN	MAPK1	DepMap
13	BRAF	CYP3A4	DepMap	52	PTEN	DSCC1	DepMap
14	BRAF	LUC7L2	DepMap	53	PTEN	AKT1	DepMap
15	BRAF	MAPK1	DepMap; SynLethDB-v2.0	54	PTEN	UBE2H	DepMap; SynLethDB-v2.0
16	BRAF	EGFR	SynLethDB-v2.0	55	PTEN	THBS1	DepMap
17	BRAF	PIK3CA	SynLethDB-v2.0	56	PTEN	RNF146	DepMap
18	BRAF	CHEK1	SynLethDB-v2.0	57	PTEN	MRPL13	DepMap
19	BRAF	BRCA2	SynLethDB-v2.0	58	PTEN	SLC22A2	SynLethDB-v2.0
20	BRCA1	TOPBP1	DepMap	59	PTEN	RNF126	SynLethDB-v2.0
21	BRCA1	CCT2	DepMap	60	PTEN	HRAS	SynLethDB-v2.0
22	BRCA1	BRCA2	DepMap; SynLethDB-v2.0	61	PTEN	CHEK1	SynLethDB-v2.0
23	BRCA1	DSCC1	DepMap; SynLethDB-v2.0	62	PTEN	PSMD12	SynLethDB-v2.0
24	BRCA1	CD63	DepMap	63	PTEN	TACSTD2	SynLethDB-v2.0
25	BRCA1	BRAF	DepMap	64	PTEN	LIG1	SynLethDB-v2.0
26	BRCA1	PDGFRA	SynLethDB-v2.0	65	TP53	GPX8	DepMap
27	BRCA1	RIDA	SynLethDB-v2.0	66	TP53	ATAD5	DepMap
28	BRCA1	PIK3CA	SynLethDB-v2.0	67	TP53	MCM2	DepMap
29	BRCA1	MRPL13	SynLethDB-v2.0	68	TP53	PPM1D	DepMap
30	BRCA2	PTGS1	DepMap	69	TP53	RBM15	DepMap
31	BRCA2	CYP3A5	DepMap	70	TP53	NTRK1	SynLethDB-v2.0

32	BRCA2	DCK	SynLethDB-v2.0	71	TP53	ABL1	SynLethDB-v2.0
33	BRCA2	SKP2	SynLethDB-v2.0	72	TP53	PDGFRA	SynLethDB-v2.0
34	BRCA2	EZH2	SynLethDB-v2.0	73	TP53	PRNP	SynLethDB-v2.0
35	EGFR	TSPAN1	DepMap	74	TP53	ABCB1	SynLethDB-v2.0
36	EGFR	EPHA2	DepMap	75	TP53	RAD51	SynLethDB-v2.0
37	EGFR	S100A14	DepMap	76	TP53	PCNA	SynLethDB-v2.0
38	EGFR	LAD1	DepMap	77	TP53	PIK3CA	SynLethDB-v2.0
39	EGFR	PDGFRA	SynLethDB-v2.0	78	TP53	HMGB1	SynLethDB-v2.0

Table S8. The number of SL pairs confirmed by database DepMap and SynLethDB-v2.0 among the top-500 predicted SL pairs for 10 selected genes, respectively.

Genes	DepMap	SynLethDB-v2.0	DepMap & SynLethDB-v2.0
BRCA1	18	29	2
BRCA2	14	20	2
TP53	16	36	4
PTEN	15	28	1
ATM	23	4	2
ATR	21	6	0
KRAS	17	40	1
HRAS	6	1	0
BRAF	8	8	2
EGFR	32	4	0

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.

Wu, Q. et al. (2019) PD-GAN: Adversarial Learning for Personalized Diversity-Promoting Recommendation. In: *the 28th International Joint Conference on Artificial Intelligence*, 19, 3870-3876.

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.

Deng et al. (2019). SL-BioDP: Multi-Cancer Interactive Tool for Prediction of Synthetic Lethality andResponse to Cancer Treatment. Cancers. 11:1682.