Prediction of ligand-receptor interactions based on CatBoost and deep forest and its application in cell-cell communication analysis

Wei Wu, Zhao Wang, Longlong Liu, Junfeng Huang, Haifan Qiu, Lihong Peng*, and Libo Nie*

Figure **S1-S2** discuss the performance of CellCDmT under different feature dimensions.

Table **S1** provides the parameter settings.

Table **S2** provides the performance of CellCDmT and other 7 baselines on four LRI datasets.

Table **S3-S6** provide molecular docking results of the top 10 predicted interacting LRPs on each dataset.

Figure **S3-S5** provides the UpSetR maps within CRC, HNSCC, melanoma.

Table **S7-S9** provide CCC prediction results of CellCDmT and five baselines within CRC, HNSCC, and melanoma.

1. *Effect of feature dimensions*

Considering that the AUC and AUPR values can reflect the prediction performance of models more comprehensively, we compared the effect of several feature dimensions on the performance of CellCDmT upon 5-fold cross-validation on the four datasets. We used XGBoost for feature selection, as shown in Figures S1, when the feature dimension was set to 450, we obtained the highest AUC values on the four datasets. As shown in Figures S1, although we obtained the second best AUPR value on dataset 1, the highest AUPR values were obtained on the other three datasets. Therefore, we set the feature selection dimension to 450.

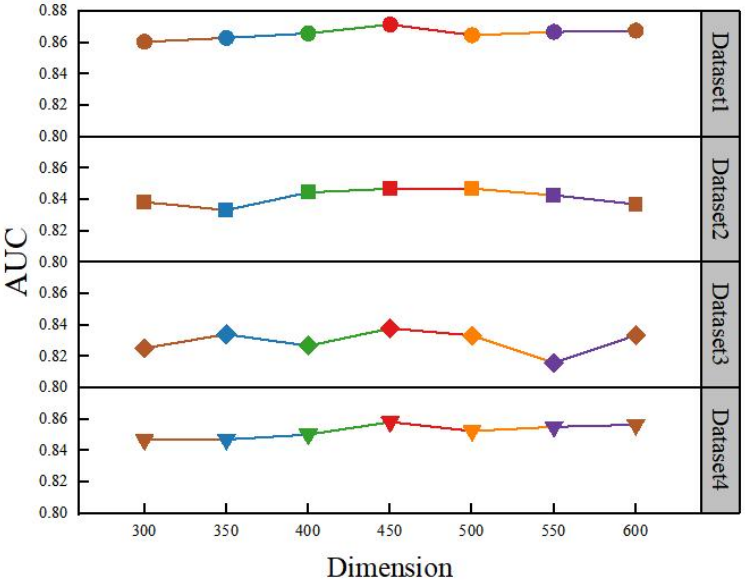


Figure S1: AUC of CellCDmT under different feature dimensions.

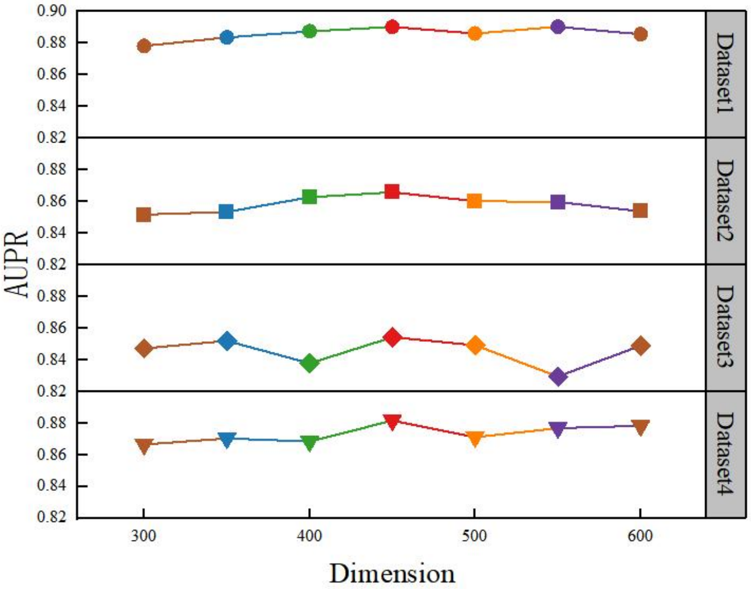


Figure S2: AUPR of CellCDmT under different feature dimensions.

2. *Parameter settings*

Table S1: Parameter settings

Method	Parameter setting
PIPR	n__estimators = 10, depth = 5, split = 5, neighbors = 3

OR-RCNN	learning__rate = 0.01, n__estimators = 20,
	max__depth = 3, criterion = ‘friedman__mse’, loss = ‘deviance’, min__samples__split = 2

DNNXGB	n__estimators = 500, max___depth = 15,
	min__child__weight = 1, max___delta__step = 0, learning__ rate = 0.2

CellEnBoost	**AdaBoostCNN**: n__estimators = 10, learning__ rate = 1, epoch=1 **LightGBM**: n___estimators = 500, learning__rate = 0.1, min__split__gain = 0, min__child__ weight =0.001

CellComNet	**HNBM:** object = ’logloss’, num__round = 4000, min__max__ depth = 1, max__max___depth = 24;
	DNN: layers = 6, β = 0.6,
	epoch = 200, activation = ’elu’, learning___ rating = 0.001

CellGiQ	**interBM:** learning___rate = 0.01,
	interactions = 600, max__bins = 256,
	early__stopping__tolerance = 1e-2,
	min__samples___leaf = 3, max___leaves = 3, max__ rounds = 900, subsample = 2
	GBNN: total___nn = 300, num__nn___step = 8, eta = 0.75, solver = lbfgs,
	max__iter = 1200, random___state = None, tol = 0.0, activation = logistic

CellDialog	n__estimators = 400, max___depth = 9,
	learning__rate = 0.1, n__components = 100, base__learner = "combined", kernel = "rbf", update___step = ’hybrid’

CellCDmT	**CatBoost:** boosting__type=’Ordered’, max__depth = 8, n__estimators = 3000;
	Deep forest: n___bins=255, bin___subsample = 2e5, bin__type = "percentile", predictor = "forest",
	n___estimators = 3, n___trees = 160

3. *The performance comparison of CellCDmT and 7 LRI baselines*

Table S2: The performance comparison of CellCDmT and 7 LRI baselines

Metric	Dataset	PIPR	OR-RCNN	DNNXGB	CellEnBoost	CellComNet	CellGiQ	CellDialog	CellCDmT
Precision	Dataset 1	0.7203±0.0078	0.7227±0.0096	0.7920±0.0132	0.7917±0.0135	0.8361±0.0033	0.7834±0.0028	0.8088±0.0060	**0.8386±0.0024**
	Dataset 2	0.6856±0.0090	0.7037±0.0103	0.7696±0.0297	0.7528±0.0132	0.7854±0.0059	0.7471±0.0047	0.7810±0.0080	**0.8116±0.0091**
	Dataset 3	0.6893±0.0101	0.6891±0.0078	0.7495±0.0112	0.7404±0.0158	0.7835±0.0044	0.7491±0.0058	0.7754±0.0089	**0.8170±0.0045**
	Dataset 4	0.7399±0.0072	0.7417±0.0060	0.8026±0.0109	0.7829±0.0087	0.8229±0.0037	0.7841±0.0024	0.8146±0.0041	**0.8326±0.0076**
Recall	Dataset 1	0.7403±0.0140	0.7217±0.0151	0.7625±0.0029	0.7567±0.0166	0.6745±0.0045	0.7607±0.0033	**0.7662±0.0069**	0.7382±0.0106
	Dataset 2	0.7068±0.0188	0.7223±0.0179	**0.7612±0.0153**	0.7235±0.0175	0.6285±0.0072	0.7355±0.0057	0.7443±0.0089	0.7166±0.0122
	Dataset 3	0.7068±0.0172	0.7240±0.0125	**0.7458±0.0320**	0.7128±0.0228	0.6399±0.0091	0.7404±0.0052	0.7322±0.0081	0.7034±0.0066
	Dataset 4	0.7666±0.0111	0.7665±0.0061	0.7349±0.0128	**0.7829±0.0087**	0.6533±0.0046	0.7477±0.0012	0.7606±0.0038	0.7308±0.0078
Accuracy	Dataset 1	0.7256±0.0055	0.7213±0.0053	0.7810±0.0081	0.7787±0.0104	0.7711±0.0025	0.7751±0.0023	0.7924±0.0054	**0.7980±0.0036**
	Dataset 2	0.6907±0.0061	0.7083±0.0071	0.7661±0.0229	0.7428±0.0114	0.7282±0.0047	0.7431±0.0041	0.7676±0.0073	**0.7749±0.0052**
	Dataset 3	0.6934±0.0070	0.7052±0.0082	0.7483±0.0153	0.7311±0.0141	0.7313±0.0045	0.7460±0.0044	0.7598±0.0074	**0.7728±0.0031**
	Dataset 4	0.7479±0.0036	0.7496±0.0033	**0.8038±0.0071**	0.7668±0.0064	0.7563±0.0027	0.7709±0.0015	0.7937±0.0029	0.7918±0.0056
F1-score	Dataset 1	0.7292±0.0061	0.7209±0.0063	0.7770±0.0092	0.7737±0.0111	0.7465±0.0032	0.7718±0.0024	**0.7867±0.0057**	0.7849±0.0053
	Dataset 2	0.6949±0.0079	0.7120±0.0083	0.7654±0.0181	0.7377±0.0123	0.6980±0.0057	0.7411±0.0041	0.7620±0.0076	**0.7673±0.0112**
	Dataset 3	0.6969±0.0079	0.7052±0.0082	0.7477±0.0148	0.7362±0.0156	0.7040±0.0063	0.7460±0.0044	0.7524±0.0074	**0.7558±0.0038**
	Dataset 4	0.7524±0.0038	0.7537±0.0026	0.7675±0.0061	0.7600±0.0074	0.7282±0.0033	0.7654±0.0013	**0.7866±0.0029**	0.7783±0.0061
AUC	Dataset 1	0.7910±0.0049	0.7888±0.0055	0.8260±0.0110	0.8424±0.0104	0.8430±0.0025	0.8456±0.0020	0.8609±0.0041	**0.8717±0.0041**
	Dataset 2	0.7605±0.0063	0.7802±0.0065	0.8028±0.0089	0.8070±0.0114	0.8018±0.0044	0.8157±0.0026	0.8413±0.0066	**0.8470±0.0050**
	Dataset 3	0.7580±0.0060	0.7647±0.0079	0.7981±0.0113	0.8002±0.0141	0.7951±0.0036	0.8164±0.0035	0.8279±0.0063	**0.8379±0.0040**
	Dataset 4	0.8222±0.0036	0.8182±0.0029	0.8184±0.0071	0.8328±0.0064	0.8306±0.0017	0.8433±0.0011	**0.8616±0.0026**	0.8582±0.0046
AUPR	Dataset 1	0.7796±0.0059	0.7878±0.0078	0.7740±0.0294	0.8547±0.0103	0.8537±0.0028	0.8580±0.0117	0.8816±0.0038	**0.8901±0.0033**
	Dataset 2	0.7515±0.0069	0.7830±0.0069	0.7718±0.0115	0.8213±0.0144	0.8112±0.0055	0.8236±0.0027	0.8597±0.0057	**0.8658±0.0046**
	Dataset 3	0.7494±0.0061	0.7639±0.0086	0.7622±0.0303	0.8029±0.0170	0.8052±0.0037	0.8276±0.0038	0.8492±0.0063	**0.8543±0.0028**
	Dataset 4	0.8176±0.0045	0.8148±0.0049	0.7696±0.0113	0.8536±0.0066	0.8399±0.0022	0.8607±0.0009	**0.8828±0.0026**	0.8817±0.0041

4. *The CCC inference result validation*

Table S3: Molecular docking results of the top 10 predicted interacting LRPs on Dataset 1

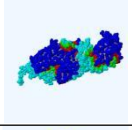
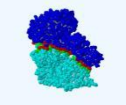
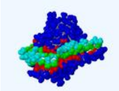
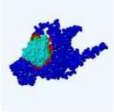
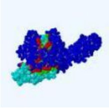
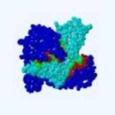
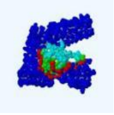
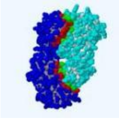
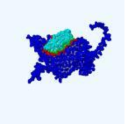
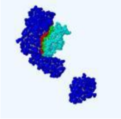
LRI(Gene Names)	Molecular docking	Binding Enegy (kcal/mol)	Hydrogen bonds	Interface area
MIF-CD74		-41.6	3.09	3862.5
B2M-TFRC		-38.7	3.65	2743.5
WNT7B-FZD1		-32.0	2.99	1783.4
LYZ-CNR1		-35.8	2.90	2329.1
NDP-ADRA2A		-33.8	2.88	2628.4
PRL-ADRA2A		-32.1	2.88	2731.3
CCL1-ADRA2A		-36.2	2.86	1741.8
TFPI-CNR1		-19.4	2.65	1741.9
PDCD1LG2-CX3CR1		-25.3	2.72	1805.1
WNT3-ADRA2A		-22.0	2.60	1167.3

Table S4: Molecular docking results of the top 10 predicted interacting LRPs on Dataset 2

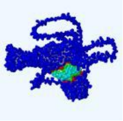
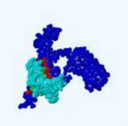
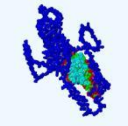
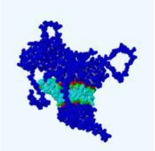
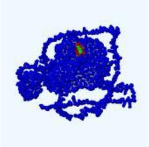
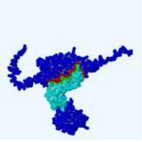
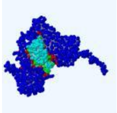
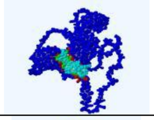
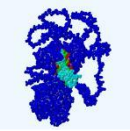
LRI(Gene Names)	Molecular docking	Binding Enegy (kcal/mol)	Hydrogen bonds	Interface area
Agrn-Fzd6		-37.5	3.52	2274.2
Bmp8b-Bmpr1b		-18.2	3.70	1074.5
Csf3-Acvr2b		-18.6	3.62	1294.4
Il5-Fzd6		-61.8	3.31	4442.4
Il33-Fzd6		-46.7	2.95	2965.5
Ptgs2-Fzd6		-17.4	2.73	547.1
Rspo2-Ackr4		-17.1	3.57	1123.3
Rspo2-Ccr1		-27.5	2.76	2446.6
Rspo2-Fzd6		-26.6	2.58	2306.7
Rspo2-Ptch1		-26.7	2.58	2496.0

Table S5: Molecular docking results of the top 10 predicted interacting LRPs on Dataset 3

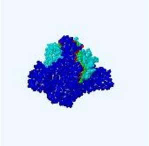
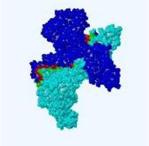
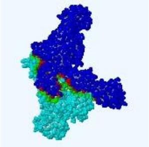
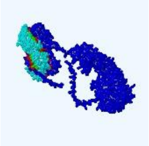
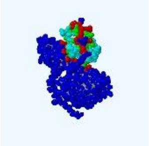
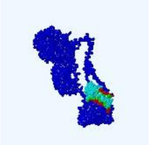
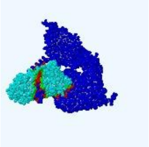
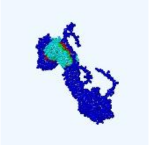
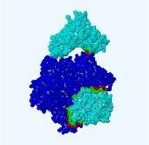
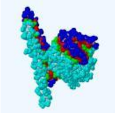
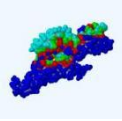
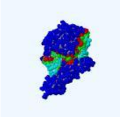
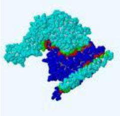
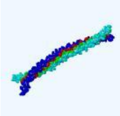
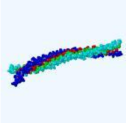
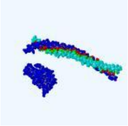
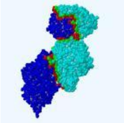
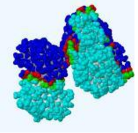
LRI(Gene Names)	Molecular docking	Binding Enegy (kcal/mol)	Hydrogen bonds	Interface area
Bmp10-Itgb1		-55.7	3.73	3742.3
Cdh1-Itgb1		-49.4	3.75	3546.0
Cntn2-Itgb1		-46.9	2.82	3723.2
Col2a1-Itga5		-39.2	3.04	2675.4
Fgf20-Fgfr2		-21.7	3.12	1404.8
Gnai2-Bmpr1b		-8.2	2.94	705.6
Hspg2-Itga5		-26.9	2.74	1678.6
Lama1-Itga5		-16.4	2.99	1941.1
Lama1-Itgav		-17.2	2.99	1732.7
Selplg-Itgb1		-47.4	3.76	3720.3

Table S6: Molecular docking results of the top 10 predicted interacting LRPs on Dataset 4

LRI(Gene Names)	Molecular docking	Binding Enegy (kcal/mol)	Hydrogen bonds	Interface area
APOC-CTSD		-47.3	2.84	4251.6
FN1-BSG		-29.0	3.15	2169.1
FN1-PRDX4		-24.9	2.92	3067.2
ITGA5-COL4A5		-31.4	2.16	1991.7
ITGA5-CTSD		-42.2	3.7	2555.7
KRT1-AP2M1		-55.2	2.69	2467.0
KRT1-COL4A5		-55.2	2.69	2464.8
KRT1-PRDX4		-55	2.69	2454.4
TFRC-PRDX4		-56.7	3.36	5016.1
LGALS1-PRDX4		-26.1	2.92	3075.0

5. CCC prediction results within CRC, HNSCC, and melanoma tissues.

Table S7: Comparison of CellCDmT with six CCC prediction tools in melanoma

Method	Ranking 1	Ranking 2	Ranking 3	Ranking 4	Ranking 5	Ranking 6
CellChat	CAFs	Endothelial cells	Macrophages	T cells	NK cells	B cells
iTALK	CAFs	Macrophages	Endothelial cells	NK cells	T cells	B cells
CellPhoneDB	Macrophages	Endothelial cells	CAFs	T cells	B cells	NK cells
NATMI	CAFs	Endothelial cells	T cells	Macrophages	B cells	NK cells
CellComNet	CAFs	Macrophages	Endothelial cells	NK cells	T cells	B cells
CellEnBoostp	CAFs	Macrophages	Endothelial cells	NK cells	T cells	B cells
CellEnBoosts	CAFs	Macrophages	Endothelial cells	T cells	B cells	NK cells
CellEnBoost c	CAFs	Macrophages	Endothelial cells	T cells	NK cells	B cells
CellGiQ	CAFs	Macrophages	Endothelial cells	T cells	B cells	NK cells
CellDialog	CAFs	Macrophages	Endothelial cells	T cells	NK cells	B cells
CellCDmT	CAFs	Macrophages	Endothelial cells	NK cells	T cells	B cells

Table S8: Comparison of CellCDmT with six CCC prediction tools in HNSCC tissue.

Method	Ranking 1	Ranking 2	Ranking 3	Ranking 4	Ranking 5	Ranking 6	Ranking 7	Ranking 8
CellChat	Fibroblasts	Endothelial cells	Macrophages	T cells	Mast cells	Dendritic cells	B cells	Myocytes
iTALK	Fibroblasts	Endothelial cells	Macrophages	Myocytes	Dendritic cells	Mast cells	B cells	T cells
CellPhoneDB	Macrophages	Endothelial cells	Fibroblasts	T cells	Dendritic cells	Mast cells	B cells	Myocytes
NATMI	Macrophages	Endothelial cells	Myocytes	Fibroblasts	Dendritic cells	Mast cells	B cells	T cells
CellComNet	Endothelial cells	Macrophages	Fibroblasts	Dendritic cells	Mast cells	T cells	Myocytes	B cells
CellEnBoostp	Macrophages	Endothelial cells	Fibroblasts	Dendritic cells	Mast cells	T cells	Myocytes	B cells
CellEnBoosts	Fibroblasts	Endothelial cells	Macrophages	T cells	Myocytes	Dendritic cells	Mast cells	B cells
CellEnBoostc	Macrophages	Endothelial cells	Fibroblasts	Dendritic cells	Mast cells	T cells	Myocytes	B cells
CellGiQ	Macrophages	Fibroblasts	Endothelial cells	Myocytes	Dendritic cells	B cells	Mast cells	T cells
CellDialog	Fibroblasts	Endothelial cells	T cells	Dendritic cells	Macrophages	Mast cells	Myocytes	B cells
CellCDmT	Fibroblasts	T cells	Endothelial cells	Macrophages	Mast cells	Dendritic cells	B cells	Myocytes

Table S9: Comparison of CellCDmT with six CCC prediction tools in CRC

Method	Ranking 1	Ranking 2	Ranking 3	Ranking 4	Ranking 5	Ranking 6	Ranking 7
CellChat	Macrophages	Endothelial cells	B cells	Mast cells	Fibroblasts	T cells	Epithelial cells
iTALK	Fibroblasts	Epithelial cells	B cells	Endothelial cells	T cells	Macrophages	Mast cells
CellPhoneDB	Fibroblasts	T cells	Epithelial cells	Macrophages	B cells	Endothelial cells	Mast cells
NATMI	Endothelial cells	Macrophages	Fibroblasts	Mast cells	T cells	B cells	Epithelial cells
CellComNet	Endothelial cells	Macrophages	T cells	Fibroblasts	B cells	Epithelial cells	Mast cells
CellEnBoostp	Endothelial cells	Macrophages	T cells	Fibroblasts	B cells	Epithelial cells	Mast cells
CellEnBoosts	Epithelial cells	Fibroblasts	B cells	T cells	Macrophages	Endothelial cells	Mast cells
CellEnBoostc	Fibroblasts	Endothelial cells	Epithelial cells	T cells	Mast cells	B cells	Macrophages
CellGiQ	Fibroblasts	Macrophages	Endothelial cells	Epithelial cells	T cells	B cells	Mast cells
CellDialog	T cells	Fibroblasts	Epithelial cells	Mast cells	Endothelial cells	Macrophages	B cell
CellCDmT	Epithelial cells	Fibroblasts	B cells	T cells	Endothelial cell	Mast cells	Macrophages

6. *The CCC inference result validation*

Figures S3-S5 present UpSetR map, analyzing LRI detection concordance across five CCC inference tools in melanoma, HNSCC, and CRC tissues respectively. The left panels of Figures S3-S5 illustrate comparative LRI$_{sensitivity}$ metrics across the five analytical tools through vertical bar charts. Corresponding right panels employ intersecting dot-plot matrices to visualize pairwise LRI overlaps between tools, with overlying histograms quantifying the cardinality of shared LRIs among tool combinations. Comparative analysis revealed that CellCDmT achieved superior performance in both evaluation metrics, demonstrating: (1) the highest LRI$_{sensitivity}$ values among all examined tools, (2) the greatest number of overlapping LRIs when cross-validated against the other four computational methods. This dual superiority suggests CellCDmT maintains optimal balance between detection sensitivity and consensus validation in multi-tool CCC analyses.

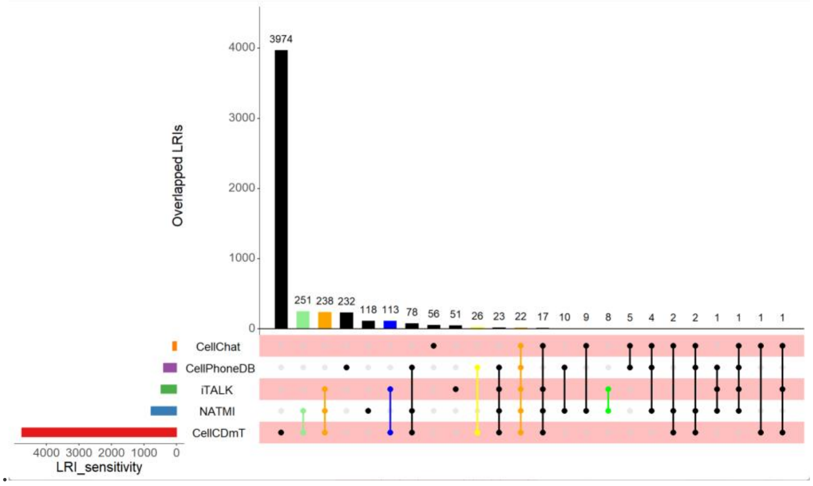


Figure S3: UpsetR map of Melanoma

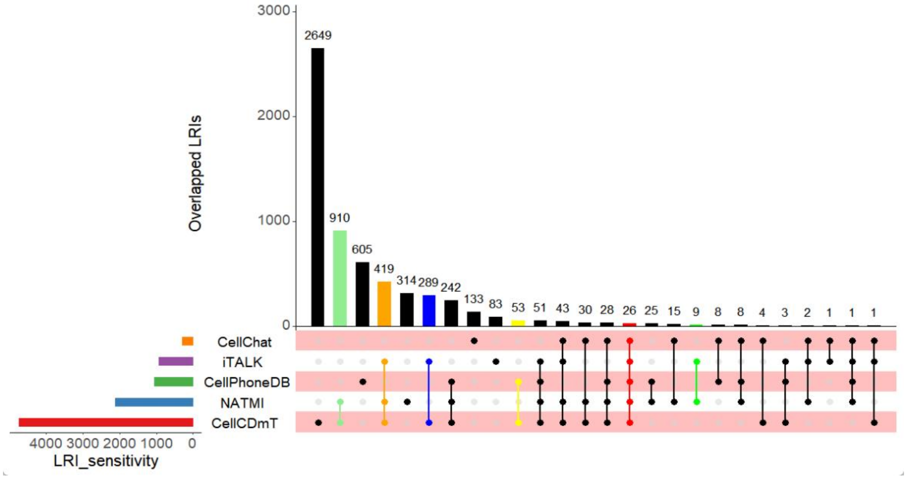


Figure S4: UpsetR map of HNSCC

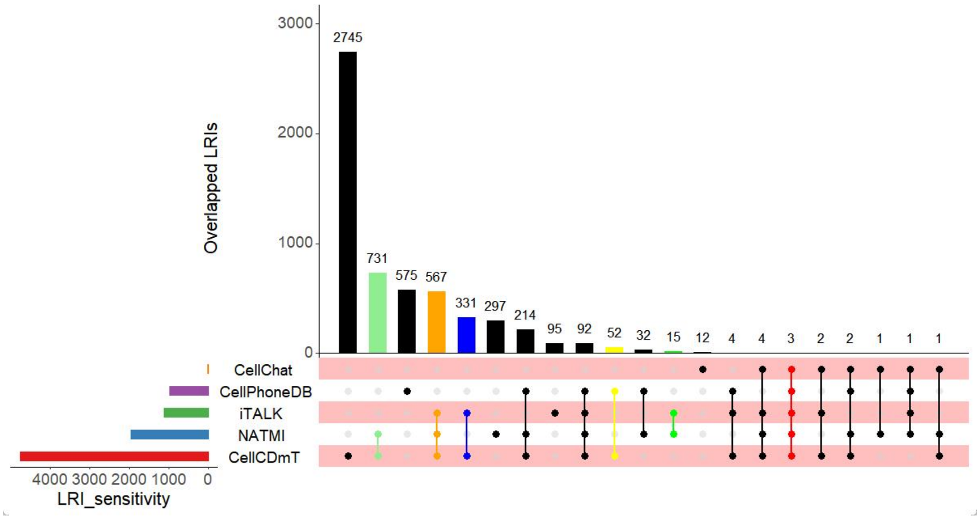


Figure S5: UpsetR map of CRC