19th European Conference on Computational Biology

Planetary Health and Biodiversity

31st August8th September 2020

Cancer Drug Response Prediction via a Hybrid Graph Convolutional Network

Qiao Liu

liuqiao@stanford.edu

Outline

- Background
- Methods
- Results
- Conclusions

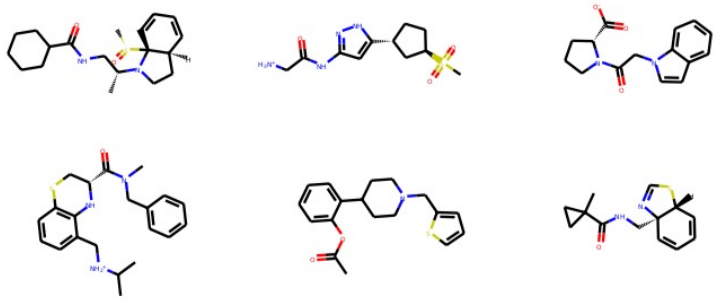
Cancer drug response

- Cancer drug response (CDR) describes how a cancer patient responses to a specific drug
  - Sensitive – a high degree of drug efficacy
  - Resistant – a low degree of drug efficacy
  - IC_{50} – half-maximal inhibitory concentration

- Identification of cancer drug response (CDR) is crucial for personalized cancer therapy

- Drug efficacy may vary due to the heterogeneity of cancer patients

- There are millions of compounds while only a minority can be considered as drugs



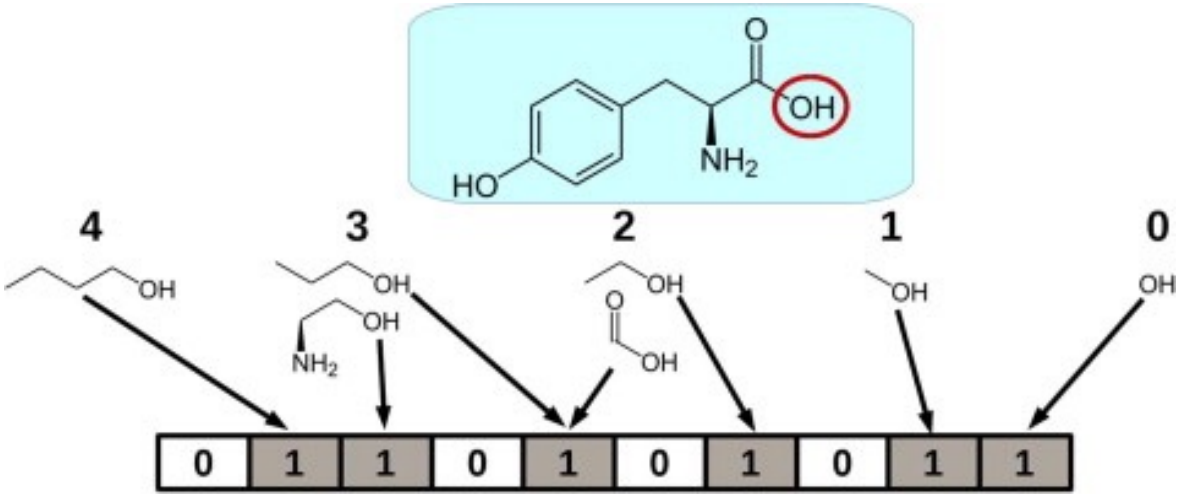
Two types of methods for predicting CDR

- Network-driven methods
  - The core idea is to construct a similarity-based model and assign the sensitivity profile of a known drug to a new drug if there are structurally similar.
  - Major limitation: low scalability and efficiency
- Machine learning methods
  - Taking large-scale profiles of both drugs and cancer cell lines as input
  - Typical approaches include LR (Geeleher et al., 2014), SVM (Dong et al., 2015), random forest (Daemen et al., 2013), neural networks (Chang et al., 2018)

Limitations of current methods

- Most previous works used engineered features of drugs such as molecular fingerprints

- Most previous works used single omics data of cancer cell lines, such as genomic mutation or gene expression



(Adrià et al., 2015)

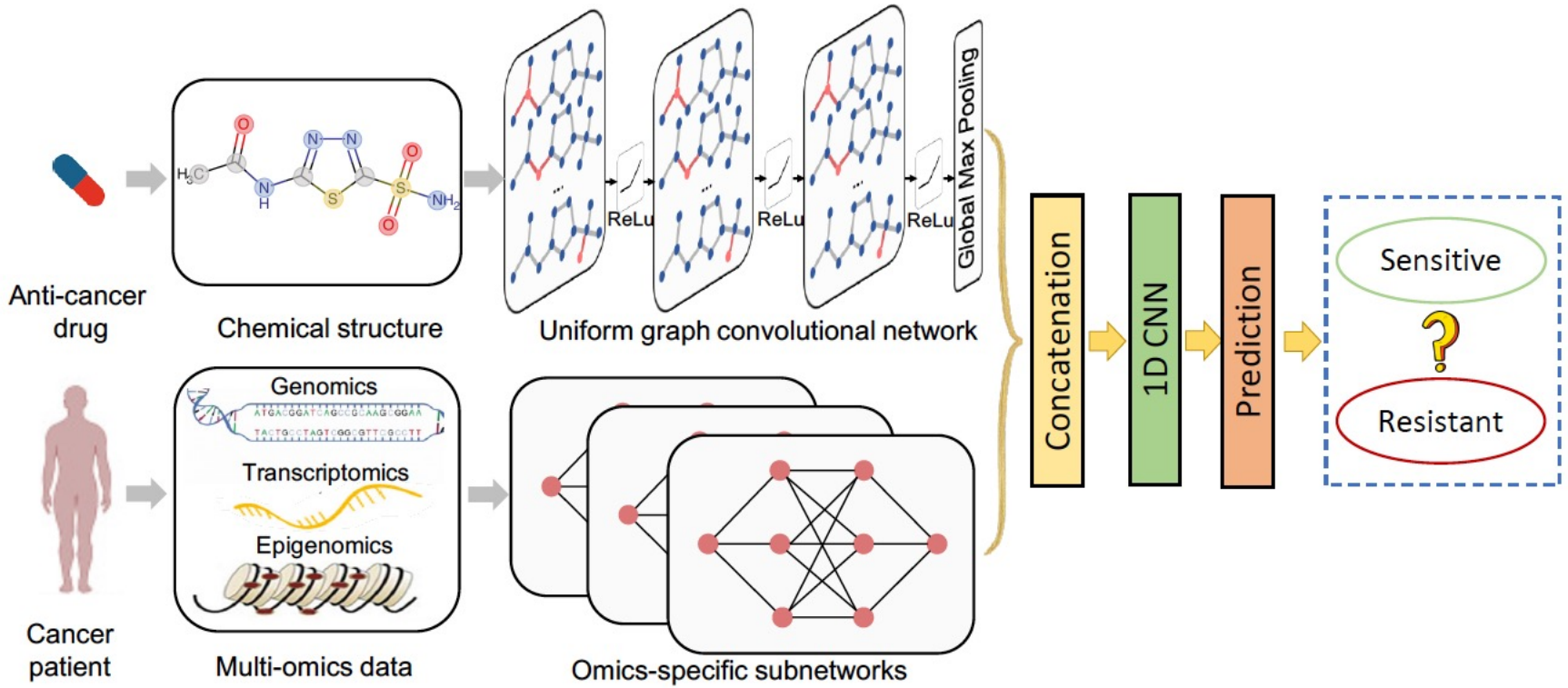
Motivation and intuition

- Drug can be naturally represented as a graph as it has unique topological structures, such as edges (chemical bond) and nodes (atoms)
- Multi-omics data of cancer cell line can reveal more biological information than single-omics data
- Identifying molecular signatures that determine drug response has not been fully explored

A hybrid graph convolutional network

- We proposed DeepCDR, as an end-to-end framework for predicting cancer drug response

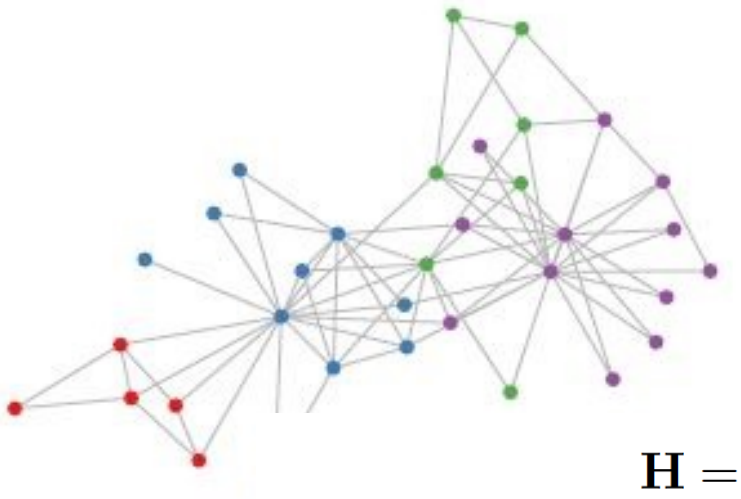
- DeepCDR takes chemical structures of drugs and multi-omics data of cancer cell line as input and outputs the sensitivity of drug



Convolutional graph networks

- Graph neural networks aim at graph-based representation of data by exploring topological relationships among nodes of a graph

- Convolutional graph networks (GCNs) introduced convolutions on graph largely based on spectral graph theory (Kipf et al., *ICLR,* 2017)



$$\mathbf{H} = \tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{X}\mathbf{\Theta} \quad (1)$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_N$ and $\tilde{\mathbf{D}}_{i,i} = \sum_j \tilde{\mathbf{A}}_{i,j}$. $\mathbf{X} \in \mathbb{R}^{N \times C}$ is the input signal with C channels and $\mathbf{\Theta} \in \mathbb{R}^{C \times F}$ denotes the parameters of F filters.

Uniform graph convolutional network(UGCN)

- Drug can be represented as $\{\mathcal{G}_i = (\mathbf{X}_i, \mathbf{A}_i)|_{i=1}^{M}\}$ where $\mathbf{X}_i \in \mathbb{R}^{N_i \times C}$ and $\mathbf{A}_i \in \mathbb{R}^{N_i \times N_i}$ is the feature matrix and adjacent matrix of i[th] drug

- Each atom in a drug was represented as a 75-dimensional feature vector, including chemical and topological properties (atom type, degree and hybridization)

- We extended the original GCN (Kipf et al., *ICLR,* 2017) for processing drugs with variable sizes and structures by introducing complementary drug

- Layer-wise operation of UGCN is derived as

$$\mathbf{H}_i^{(l+1,\alpha)} = \sigma(((\tilde{\mathbf{D}}_i + \mathbf{D}_i^B)^{-\frac{1}{2}}\tilde{\mathbf{A}}_i(\tilde{\mathbf{D}}_i + \mathbf{D}_i^B)^{-\frac{1}{2}}\mathbf{H}_i^{(l,\alpha)} + (\tilde{\mathbf{D}}_i + \mathbf{D}_i^B)^{-\frac{1}{2}}\mathbf{B}_i(\tilde{\mathbf{D}}_i^c + \mathbf{D}_i^{B^T})^{-\frac{1}{2}}\mathbf{H}_i^{(l,\beta)})\mathbf{\Theta}^{(l)})$$

$$\mathbf{H}_i^{(l+1,\beta)} = \sigma(((\tilde{\mathbf{D}}_i^c + \mathbf{D}_i^{B^T})^{-\frac{1}{2}}\mathbf{B}_i^T(\tilde{\mathbf{D}}_i + \mathbf{D}_i^B)^{-\frac{1}{2}}\mathbf{H}_i^{(l,\alpha)} + (\tilde{\mathbf{D}}_i^c + \mathbf{D}_i^{B^T})^{-\frac{1}{2}}\tilde{\mathbf{A}}_i^c(\tilde{\mathbf{D}}_i^c + \mathbf{D}_i^{B^T})^{-\frac{1}{2}}\mathbf{H}_i^{(l,\beta)})\mathbf{\Theta}^{(l)})$$

- Drugs with different size will be embedded into a fixed dimensional vector (default: 100)

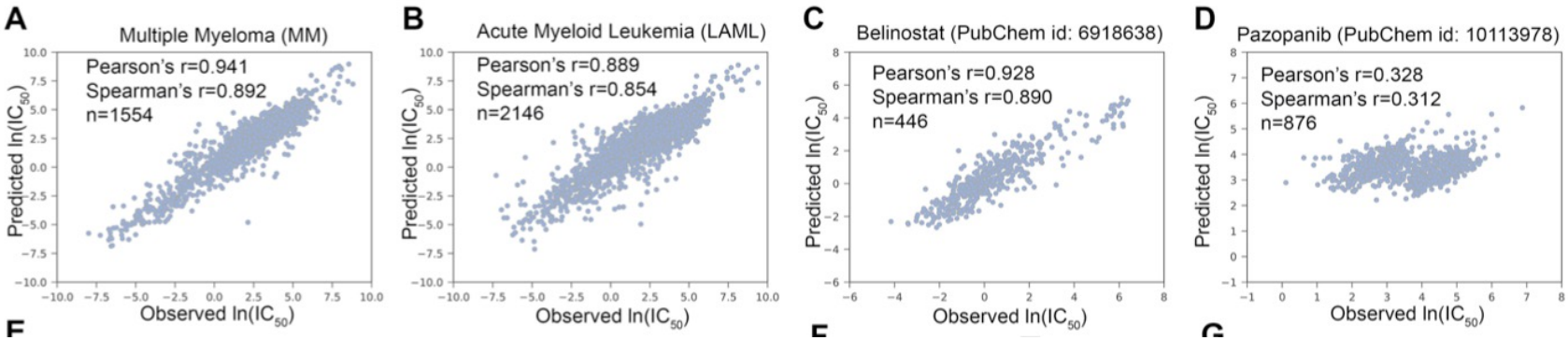
Dataset

- Drug structure data (.MOL) were from Pubchem database (Kim et al., 2018)
- Cancer cell lines data (genomic mutation, gene expression and DNA methylation) were downloaded from Cancer Cell Line Encyclopedia (CCLE) (Barretina et al. 2012)
- Drug sensitivity data were obtained from Genomics of Drug Sensitivity in Cancer (GDSC) (Iorio et al. 2016)
- We finally collected a dataset containing 107446 instances across 561 cancer cell lines and 223 drugs
- Each cell line corresponds to a TCGA cancer type.

Rediscovering cancer drug responses

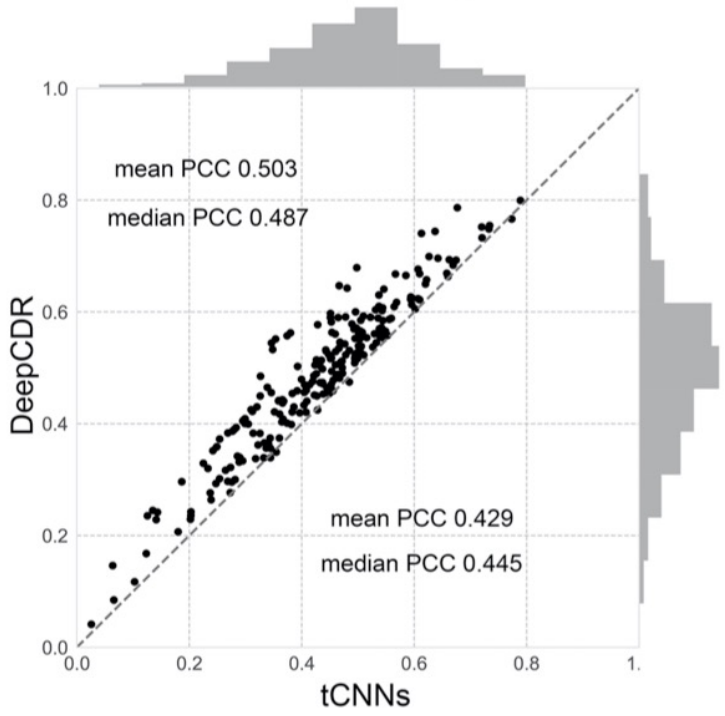
- Baseline methods: ridge regression, random forest and CDRscan
- DeepCDR outperforms other methods by a large margin

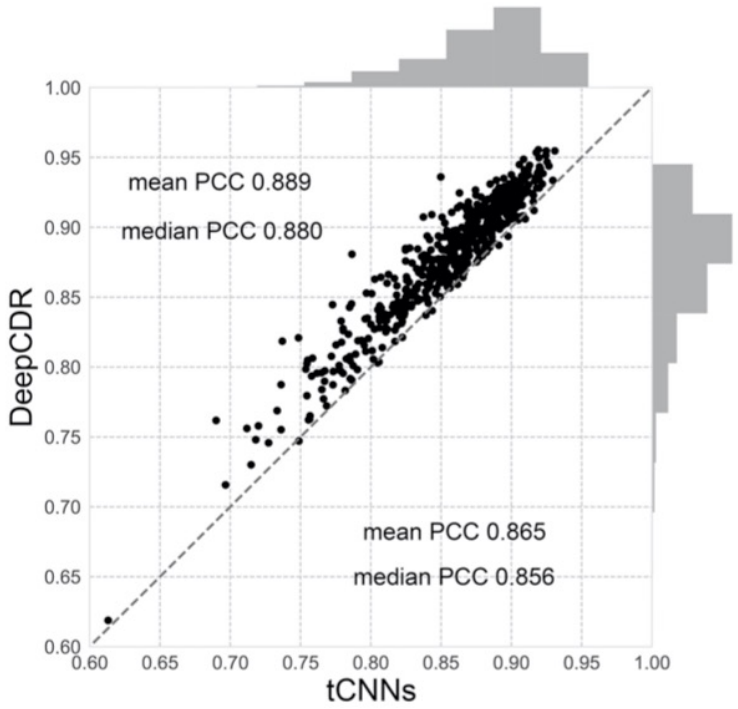
Methods	Pearson's correlation	Spearman's correlation	RMSE
Ridge Regression	0.780	0.731	2.368
Random Forest	0.809	0.767	2.270
MOLI	0.813±0.007	0.782±0.005	2.282±0.008
CDRscan	0.871±0.004	0.852±0.003	1.982±0.005
tCNNs	0.885±0.008	0.862±0.006	1.782±0.006
DeepCDR	**0.923±0.006**	**0.903±0.004**	**1.058±0.006**



Bind test for cancer cell line or drug

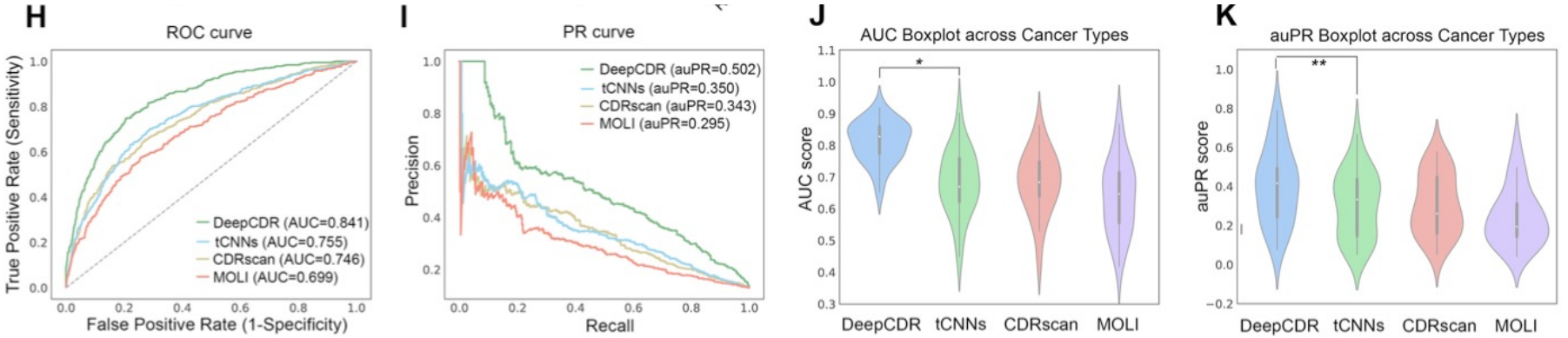
- DeepCDR still outperforms the best baseline method tCNNs
- Blind test for drug: 0.503 vs 0.429
- Blind test for cancer cell line: 0.889 vs 0.865





DeepCDR predicts the binary drug sensitivity status

- Using threshold to binarize IC_{50} values, each drug has a unique threshold IC_{50} value provided by Iorio et al. 2016
- DeepCDR again outperforms tCNNs, CDRscan and MOLI by achieving higher AUCs and auPRs



Ablation studies

- We evaluated the contribution of each type of omics data
- Besides, we verified the power of graph convolution by setting adjacent matrices to identity matrices
- We found that epigenomic data contributed most but was ignored by many previous studies

Table 2: Model ablation results with different experimental settings

Experimental setting	Pearson's correlation
Single genomics	0.846
Single transcriptomics	0.878
Single epigenomics	0.891
Multi-omics without adjacent info	0.886
Multi-omics with adjacent info	**0.921**

Model interpretation

- We calculated the absolute gradient of the predicted outcome from DeepCDR regression model with respect to the each gene's expression
- Top-5 prioritized genes were illustrated, many top ranked genes have been verified to be associated with cancers by existing literature

Table 3: Top-5 cancer-associated genes prioritized by DeepCDR

Drug	Cell line	TCGA type	Top-5 cancer-associated genes	$\ln(IC_{50})$	
				Observed	Predicted
Erlotinib	A3/KAW	DLBC	EGFR,ALK,BCL10,CREB3L1,STAG1	1.110	1.206
Lapatinib	BT-474	BRCA	ERBB2,MDS2,FOXL2,EGFR,MNX1	-1.028	-0.879
Bleomycin	A-375	SKCM	ACKR3,ASXL1,MTCP1,FOXL2,SALL4	-1.514	-1.428
Nilotinib	BHT-101	THCA	CBLC,ABI1,POU5F1,KLF4,ZNF198	-0.630	-0.714
Salubrinal	SUP-B15	ALL	JAK3,EIF1AX,NUMA1,PRDM1,IL21R	1.781	1.471

EGFR inhibitors

Conclusion

- DeepCDR is an end-to-end deep learning framework for predicting cancer drug response
- DeepCDR contains a novel UGCN architecture for representing topological information of drugs
- DeepCDR has potential translational value in guiding disease-specific drug design

Acknowledgement

Prof. Rui Jiang

sensetime

- I would like to thank Prof. Jiang and Jiang Group for their helpful discussion

- I would like to thank Dr. Mu Zhou and SenseTime AI medical group for their helpful discussion

Thank you！