Predicting Drug Sensitivity of Cancer Cell Lines via Collaborative Filtering with Contextual Attention

Yifeng Tao[1,†], Shuangxia Ren[2,†], Michael Q. Ding[2], Russell Schwartz[1,*], Xinghua Lu[2,*]

[1]Computational Biology Department, School of Computer Science, Carnegie Mellon University

[2]Department of Biomedical Informatics, School of Medicine, University of Pittsburgh

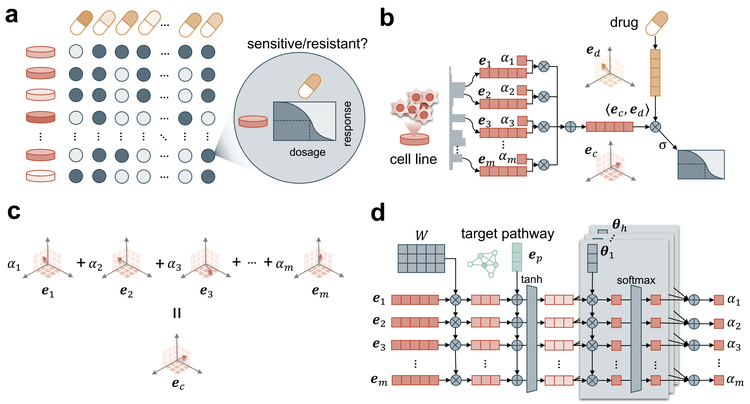
INTRODUCTION

Accurate anti-cancer drug recommendations and identification of essential biomarkers for this task are crucial to precision oncology. Large-scale drug response assays on cancer cell lines provide a potential way to understand the interplay of drugs and cancer cells. Predicting the sensitivities of cell lines to a panel of potential molecules based on omics data of the cell lines is challenging in three primary aspects.

- **Robust model**: Drug sensitivity data are noisy and often contain many missing entries.
- **Contextual effects**: The relationship between the molecular profiles of cell lines and drug response is complex.
- **Interpretability**: Cancer researchers and clinicians are concerned about the clinical implications of the models, with an emphasis on how the critical biomarkers affect the final prediction results.

METHODS

We present CADRE (Contextual Attention-based Drug REsponse). CADRE builds on the framework of collaborative filtering, which provides **robustness** to the noise of biological data by leveraging similarities within drugs and cell lines. It utilizes the **contextual** attention mechanism to identify informative biomarkers of these cell lines, which boosts prediction accuracy and affords **interpretability** of results.



Overall architecture: collaborative filtering

$$\hat{y}_{c,d} = \sigma\left(\langle \boldsymbol{e}_c, \boldsymbol{e}_d \rangle\right) = \frac{1}{1+\exp(-\boldsymbol{e}_c^{\mathsf{T}}\boldsymbol{e}_d)}$$

$$\ell(\hat{y}_{c,d}, y_{c,d}; \mathcal{W}) = \text{CrossEnt}(\hat{y}_{c,d}, y_{c,d}) + \frac{\lambda_2}{2} \cdot \ell_2(\mathcal{W})$$

$$\boldsymbol{e}_c = \sum_{i=1}^{m} 1 \cdot \boldsymbol{e}_i = 1 \cdot \boldsymbol{e}_1 + 1 \cdot \boldsymbol{e}_2 + ... + 1 \cdot \boldsymbol{e}_m$$

SADRE: Self-Attention-based Drug REsponse

Cell embedding is the weighted sum of gene embeddings:

$$\boldsymbol{e}_c = \sum_{i=1}^{m} \alpha_i \cdot \boldsymbol{e}_i = \alpha_1 \cdot \mathbf{e}_1 + \alpha_2 \cdot \mathbf{e}_2 + ... + \alpha_m \cdot \mathbf{e}_m$$

$$\alpha_1, \alpha_2, ..., \alpha_m = \text{Self-Attention}\left(\boldsymbol{e}_1, \boldsymbol{e}_1, ..., \boldsymbol{e}_m\right)$$

Self-attention implemented as a sub-neural network:

$$\beta_{i,j} = \boldsymbol{\theta}_j^{\mathsf{T}} \tanh(W\boldsymbol{e}_i), \qquad i = 1, 2, ..., m, \;\; j = 1, 2, ..., h$$

$$\alpha_{1,j}, \alpha_{2,j}, ...\alpha_{m,j} = \text{softmax}(\beta_{1,j}, \beta_{2,j}, ...\beta_{m,j}), \qquad j = 1, 2, ..., h$$

$$\alpha_{i,j} = \exp\left(\beta_{i,j}\right) \Big/ \sum_{i'=1}^{m} \exp\left(\beta_{i',j}\right), \qquad i = 1, 2, ..., m$$

$$\alpha_i = \sum_{j=1}^{h} \alpha_{i,j} = \alpha_{i,1} + \alpha_{i,2} + ... + \alpha_{i,h}, \qquad i = 1, 2, ..., m$$

CADRE: Contextual Attention-based Drug REsponse

Drug pathway knowledge is integrated:

$$\alpha_1, \alpha_2, ..., \alpha_m = \text{Contextual-Attention}\left(\boldsymbol{e}_1, \boldsymbol{e}_1, ..., \boldsymbol{e}_m, \boldsymbol{e}_p\right)$$

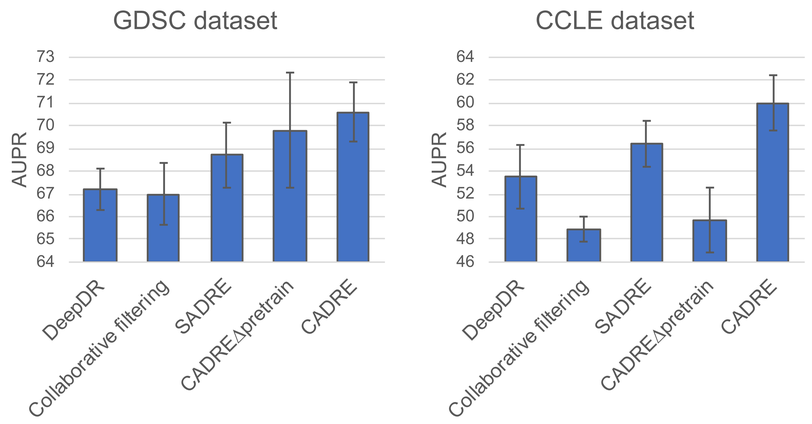
$$\beta_{i,j} = \boldsymbol{\theta}_j^{\mathsf{T}} \tanh(W\boldsymbol{e}_i + \boldsymbol{e}_p), \qquad i = 1, 2, ..., m, \;\; j = 1, 2, ..., h$$

Pretraining gene embeddings

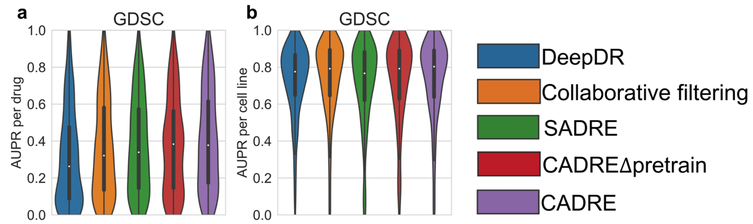
Gene embedding pretrained using gene2vec, a variant of word2vec variant, on the GEO database.

RESULTS

CADRE outperforms competing models

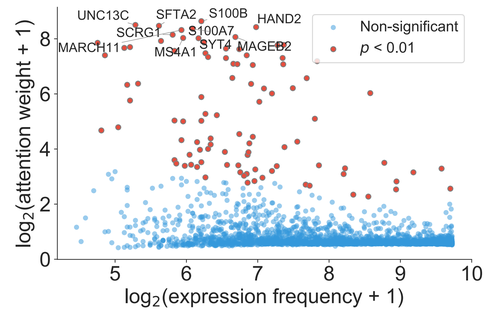


Effective attention-encoded cell line representation contributes to the major improvements of performance



CADRE identifies the critical biomarkers related to drug resistance

Two enriched pathways: export from cell & signaling receptor binding.



CONCLUSIONS

CADRE is an interpretable machine learning model that accurately predicts drug sensitivities of cancer cell lines from their expression levels.

- CADRE built upon collaborative filtering, which is capable of coping with noisy data.
- The attention mechanism of CADRE improves both model interpretability and performance.
- Gene representations transferred from external database boost CADRE performance further.

REFERENCES

1. Wanjuan Yang, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells, 2013.
2. Zichao Yang, et al. Hierarchical attention networks for document classification, 2016.
3. Ali Oskooei, et al. PaccMann: Prediction of anticancer compound sensitivity with multi-modal attention-based neural networks, 2018.
4. Jingcheng Du, et al. Gene2vec: distributed representation of genes based on co-expression, 2019.