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

Yifeng Tao[1,2,†] YIFENGT@CS.CMU.EDU
Shuangxia Ren[3,4,†] SHR81@PITT.EDU
Michael Q. Ding[3] DINGM@PITT.EDU
Russell Schwartz[1,5,*] RUSSELLS@ANDREW.CMU.EDU
Xinghua Lu[3,4,6,*] XINGHUA@PITT.EDU

[1] *Computational Biology Department, School of Computer Science, Carnegie Mellon University*
[2] *Joint Carnegie Mellon-University of Pittsburgh Ph.D. Program in Computational Biology*
[3] *Department of Biomedical Informatics, School of Medicine, University of Pittsburgh*
[4] *Intelligent Systems Program, School of Computing and Information, University of Pittsburgh*
[5] *Department of Biological Sciences, Carnegie Mellon University*
[6] *Department of Pharmaceutical Science, School of Medicine, University of Pittsburgh*
Pittsburgh, PA, USA

[†] *Both authors contributed equally to this work.*
[*] *To whom correspondence should be addressed.*

Editor: Editor's name

Abstract

Accurate anti-cancer drug recommendations and the 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. In this work, we present CADRE (Contextual Attention-based Drug REsponse), a model that accurately infers the response of cancer cell lines to a panel of candidate compounds based on the omics profiles, such as gene expressions, of cancer cells. 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. In addition, CADRE incorporates external knowledge of drug target pathways and co-expression patterns of genes to further improve feature representations and model performance. Comprehensive evaluations of CADRE and competing models on two large-scale pharmacogenomic datasets show its superiority in both prediction performance and interpretability. CADRE identifies as vital biomarkers genes related to intracellular vesicles and signaling receptor binding, shedding light on its translational potential in the clinical practice of cancer treatment.[1]

1. Code is available at https://github.com/yifengtao/CADRE.

1. Introduction

Precise prediction of drug sensitivities of tumors is one of the essential aspects of personalized treatment of cancers, which requires clinicians and researchers to assign the best potential anti-cancer drugs to individual patients (Prasad, 2016). With the rapid advancement of high-throughput sequencing technologies in the past decade, large amounts of tumor multi-omics data have become available at an acceptable cost (Reuter et al., 2015). However, inter- and intra-tumor heterogeneities (Schwartz and Schäffer, 2017) make tumor resistance to drugs a much more complex problem to resolve: Even cancer patients with the same cancer type may have distinct prognoses with the same clinical intervention (Priedigkeit et al., 2017). Furthermore, it is expensive and impractical to conduct systematic drug sensitivity assays directly on human beings. Although there is still a gap between cell lines and *in vivo* tumors, large-scale cancer-cell-line pharmacogenomic data (Yang et al., 2013; Barretina et al., 2012) provide a reasonable basis for understanding how cells and drugs interact and how inter-tumor heterogeneity can lead to distinct sensitivity profiles across tumors.

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. 1) Drug sensitivity data are **noisy** and often contain many missing entries (Liu et al., 2018). Therefore the model has to be robust, generalize well, and not otherwise overfit to the training data (Yuan et al., 2016). 2) The relationship between the molecular profiles of cell lines (such as gene expressions) and drug response is complex. Not every gene contributes equally to the response. In addition, a few genes may **interact** with each other to generate complex **contextual** effects (Zaitsev et al., 2019). 3) Cancer researchers and clinicians are especially concerned about the **interpretability** and clinical implications of the models, with an emphasis on how the critical biomarkers affect the final prediction results. Although deep learning models can achieve good to excellent performance in predicting sensitivities (Ding et al., 2018; Chiu et al., 2019), most of them behave like "black boxes" without achieving balanced performance and interpretability.

To address the three essential challenges mentioned above in cell line resistance inference, we proposed a model for accurate and interpretable drug sensitivity prediction, which we called CADRE (**C**ontextual **A**ttention-based **D**rug **RE**sponse). CADRE was built on the framework of a type of classical machine learning model called collaborative filtering (Schafer et al., 2007) to impute the sensitivities of untested cell lines to a panel of known drugs using their molecular profiles, such as gene expressions (section 3.1). Collaborative filtering captures shared features by jointly exploiting the similarities between drugs as well as similarities between cell lines, alleviating the significant **noise** in the sensitivity data (Wang et al., 2017). Furthermore, we developed and employed a contextual attention mechanism to identify the crucial inputs, capture the interactions between genes and drug targets, and thus encode the cell line features from their expression profiles effectively (See section 3.2-3.3 for implementation details). The attention mechanism is a family of deep learning modules/components which has been shown to be effective in encoding input features by assigning them different "attention weights" and thus further improving the model performance in many applications, such as computer vision (Xu et al., 2015), natural language processing (Yang et al., 2016), and computational biology (Tao et al., 2020a). Not

only did contextual attention increase prediction accuracy by capturing the **contextual interactions** of genes and drug targets, but it also improved the model **interpretability** by assigning higher weights to the genes that have a more significant effect on drug response. Although classical models such as random forest and linear models are well studied theoretically and have better interpretability compared with attention mechanism, attention mechanism still improves both model interpretability and empirical performance in deep learning. We refer readers to section 2.2 for details of how our approach is different from both classical models and conventional neural network models.

Generalizable Insights about Machine Learning in the Context of Healthcare

CADRE is a novel machine learning model for drug resistance prediction. It provides improved biological interpretability and potential translational biomarkers, compared with other deep learning models (section 5.4), and achieves better performance, compared with classical models such as collaborative filtering (section 5.2). The decisive improvements come from the contextual attention mechanism we used to encode the cell line features from gene expression levels (section 5.3). We utilized external knowledge to further boost the model by transferring gene embeddings pretrained in an unsupervised manner on a large-scale gene expression database (section 3.4,5.2). Our work could potentially be extended and applied to other clinical decisions, e.g., cancer subtype classification and cancer prognoses prediction (Chang et al., 2013), when the molecular information such as expression levels and genomic alterations are available, and when both the explainability and performance of the model are essential considerations in the task.

2. Related Work

2.1. Pharmacogenomic Datasets

An anti-cancer drug sensitivity dataset usually systematically measures the multi-omic molecular profiles of a collection of cancer cell lines, such as RNA expressions, DNA mutations, DNA copy number variations (CNVs), methylation levels, and protein abundances. It also includes the experimental dose vs. the response curves of cell growth inhibition from a panel of potential compounds. The NCI-DREAM challenge was one of the largest contests that tried to draw a community effort to solve the problem of accurately predicting the sensitivities of drugs in cancer cell lines (Costello et al., 2014). It contained response data of 53 breast cancer cell lines to 28 compounds. NCI-60 was a more massive dataset that contained the responses of 59 cell lines from various tissues to over 100k compounds (Shoemaker, 2006), providing an important resource for *in vitro* drug discovery. Cancer Genome Project (CGP; Garnett et al., 2012), Cancer Cell Line Encyclopedia (CCLE; Barretina et al., 2012), and Genomics of Drug Sensitivity in Cancer (GDSC; Yang et al., 2013) were more balanced datasets, each containing hundreds of pan-cancer cell lines with sensitivity data to tens or hundreds of compounds. The increased number of cell lines enabled further investigation of the genomic impact on drug resistance. While it is hard to examine the response of patients to different clinical options, The Cancer Genome Atlas (TCGA) collected follow-up records of around 10k pan-cancer patients in the form of survival/recurrence in response to given treatments (Chang et al., 2013).

2.2. Drug Response Prediction Models

A few classical machine learning models were proposed to solve the drug response prediction problem based on molecular profiles of cell lines, including ridge regression (Geeleher et al., 2014), elastic net (Yuan et al., 2016), support vector machine (SVM; Dong et al., 2015), random forest (Riddick et al., 2011), and Bayesian models (KBMTL; Gonen and Margolin, 2014). Researchers also developed network-based models that incorporated external knowledge of cell line similarities or drug similarities, e.g., CDCN (Wei et al., 2019) and dual-layer network (Zhang et al., 2015). NCFGER (Liu et al., 2018), SRMF (Wang et al., 2017), and KRL (He et al., 2018) were response prediction models based on collaborative filtering (Schafer et al., 2007), a class of models widely used in the area of recommender systems. Our model is different from NCFGER, which is neighbor-based, while CADRE is model-based and is more robust to noise and missing values. Compared with SRMF and KRL, CADRE considers the attention mechanism, which effectively encodes cell line features and increases the interpretability.

More deep learning models have been proposed recently. Ding et al. (2018) showed that simple neural networks such as multilayer perceptrons (MLPs) could be effective competitors with lasso or elastic net (Yuan et al., 2016). CDRscan (Chang et al., 2018) explored various architecture settings of MLPs to predict responses from cell line genomic data and drug fingerprints. Both DeepDR (Chiu et al., 2019) and MOLI (Sharifi-Noghabi et al., 2019) proposed MLP-based deep learning frameworks that used late integration to incorporate multi-omic cell line data for sensitivity prediction. PaccMann (Oskooei et al., 2018) applied attention mechanisms to the drug response prediction task using both gene expressions and drug structures. Our work is different from theirs since CADRE focuses on predicting the response of unknown cell lines to known drugs, instead of to unknown drugs as in PaccMann, which is a harder task with usually much lower accuracy in practice.

3. Methods

3.1. Overall Architecture: Collaborative Filtering

The overall architecture of CADRE is collaborative filtering (figure 1a,b; Schafer et al., 2007). Given a cell line c and a drug d, CADRE first maps the cell line and drug to two feature vectors, which we call cell line embedding $\boldsymbol{e}_c \in \mathbb{R}^s$ and drug embedding $\boldsymbol{e}_d \in \mathbb{R}^s$. CADRE then predicts the probability that the cell line c will be sensitive to drug d through the inner product and logistic function:

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

We define $\mathcal{W}$ as all the model parameters to be optimized/trained, such as gene embeddings, drug embeddings, target pathway embeddings, and neural network weights (table A1). At the training stage, we optimize the loss function:

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

where

$$\text{CrossEnt}(\hat{y}_{c,d}, y_{c,d}) = -\left[y_{c,d} \cdot \log \hat{y}_{c,d} + (1 - y_{c,d}) \cdot \log(1 - \hat{y}_{c,d})\right] \tag{3}$$

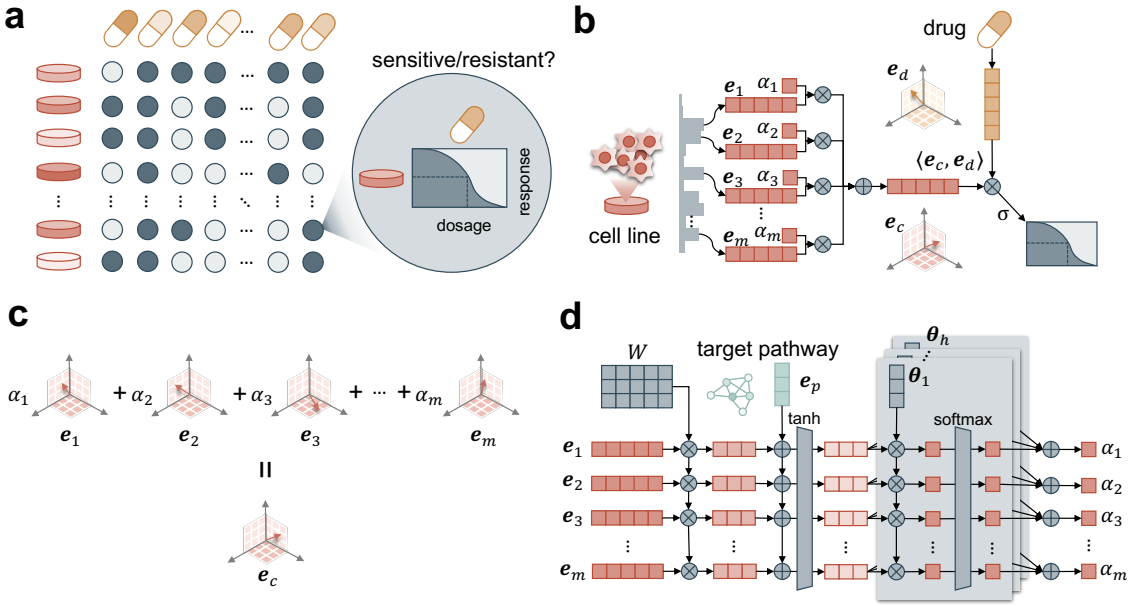

Figure 1: Diagram of the CADRE model. (**a**) The general goal of CADRE is to predict the responses of cancer cell lines to a panel of given anti-cancer drugs based on their gene expression profiles. (**b**) Given a pair of cell line c and drug d, CADRE first extracts the gene embeddings of m expressed genes in the cell line $\boldsymbol{e}_1$, $\boldsymbol{e}_2$, ..., $\boldsymbol{e}_m$ and the drug embedding $\boldsymbol{e}_d$. Then it generates the cell line embedding $\boldsymbol{e}_c$ as the weighted sum of gene embeddings. Finally, the predicted response will be $\sigma(\boldsymbol{e}_c^{\intercal}\boldsymbol{e}_d)$. (**c**) CADRE generates the cell line embedding $\boldsymbol{e}_c$ as a weighted sum of its consisting gene embeddings. (**d**) CADRE calculates the attention weights α_1, α_2, ..., α_m through the contextual attention mechanism, which was implemented as a sub neural network. It takes as input both the gene embeddings $\boldsymbol{e}_1$, $\boldsymbol{e}_2$, ..., $\boldsymbol{e}_m$ and drug target pathway embedding $\boldsymbol{e}_p$.

is the cross-entropy between predicted sensitivity $\hat{y}_{c,d}$ and ground truth sensitivity $y_{c,d}$, $\ell_2(\mathcal{W})$ is the ℓ_2-regularization term to prevent overfitting, λ_2 is the weight decay coefficient.

The mapping from drug d to its drug embedding $\boldsymbol{e}_d$ is direct, through a lookup table of drug embeddings $\mathcal{E}_{\mathcal{D}} = \{\boldsymbol{e}_d\}_{d\in\mathcal{D}}$, where $\mathcal{D}$ is the set of all the drugs in the dataset. We considered a total of 3,000 most varied genes, and for each cell line c, we had a set of m=1,500 genes that were highly expressed (section 4.3). Instead of a binary 3,000-dimensional vector, the input of the CADRE and collaborative filtering models is similar to the "bag of words": the indices of the m expressed genes $\{1, 2, ..., m\}$. We then mapped these gene indices into their corresponding gene embeddings $\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_m$ using a lookup table $\mathcal{E}_{\mathcal{G}} = \{\boldsymbol{e}_g\}_{g\in\mathcal{G}}$, where $\mathcal{G}$ is all the set of all the genes we considered ($|\mathcal{G}|$ = 3,000), and $\boldsymbol{e}_g \in \mathbb{R}^s$. We chose the parameter 3,000 and 1,500 following previous work (Ding et al., 2018; section 4.3). The major difference between CADRE and "vanilla" collaborative filtering is

in how to calculate the cell line embedding $\boldsymbol{e}_c$ from the gene embeddings $\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_m$. As we will introduce in section 3.2-3.3, SADRE (a CADRE variant) and CADRE incorporate attention mechanism to calculate $\boldsymbol{e}_c$. In vanilla collaborative filtering, however, we naively take the sum (or average) of all the gene embeddings $\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_m$:

$$\boldsymbol{e}_c = \sum_{i=1}^{m} 1 \cdot \boldsymbol{e}_i = 1 \cdot \boldsymbol{e}_1 + 1 \cdot \boldsymbol{e}_2 + ... + 1 \cdot \boldsymbol{e}_m. \tag{4}$$

We added a dropout layer (Srivastava et al., 2014) after the $\boldsymbol{e}_c$ to reduce the model complexity to prevent overfitting:

$$\boldsymbol{e}_c = \text{dropout}(\boldsymbol{e}_c; \rho), \tag{5}$$

where $\rho \in [0, 1]$ is the dropout rate. We use "vanilla" to refer to the standard collaborative filtering model, in contrast to more advanced models, such as CADRE in this work, which also build on the framework of collaborative filtering.

3.2. SADRE: Self-Attention-based Drug Response

Instead of summing up all the m gene embeddings with equivalent weights in the vanilla collaborative filtering (section 3.1; equation 4), we assumed that different genes should have different importance when we aggregate them into a single cell line embedding (figure 1b,c):

$$\boldsymbol{e}_c = \sum_{i=1}^{m} \alpha_i \cdot \boldsymbol{e}_i = \alpha_1 \cdot \mathbf{e}_1 + \alpha_2 \cdot \mathbf{e}_2 + ... + \alpha_m \cdot \mathbf{e}_m. \tag{6}$$

We could calculate the weights $\alpha_1, \alpha_2, ..., \alpha_m$ ($\alpha_i > 0$, $i = 1, 2, ..., m$) through various attention mechanisms (Yang et al., 2016; Oskooei et al., 2018), thus enabling a better feature representation of the cell line from its composing gene embeddings. We developed two attention-based collaborative filtering models in this paper: SADRE and CADRE. CADRE uses a slightly different attention mechanism from SADRE, and will be described in section 3.3. In SADRE (**S**elf-**A**ttention-based **D**rug **RE**sponse), the attention weights $\alpha_1, \alpha_2, ..., \alpha_m$ are the outputs of all the gene embeddings $\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_m$ using a Self-Attention function:

$$\alpha_1, \alpha_2, ..., \alpha_m = \text{Self-Attention}\left(\boldsymbol{e}_1, \boldsymbol{e}_1, ..., \boldsymbol{e}_m\right). \tag{7}$$

This self-attention function captures the contextual impact of other expressed genes when we calculate the weight of i-th gene α_i: it can not be solely calculated using $\boldsymbol{e}_i$. We implemented the Self-Attention function via a sub neural network (figure 1d; Yang et al., 2016), which first calculates the unnormalized attention weights:

$$\beta_{i,j} = \boldsymbol{\theta}_j^{\intercal} \tanh(W \boldsymbol{e}_i), \qquad i = 1, 2, ..., m, \;\; j = 1, 2, ..., h, \tag{8}$$

where $W \in \mathbb{R}^{q \times s}$, $\boldsymbol{\theta}_j \in \mathbb{R}^q$ are trainable model parameters. Then we normalize them:

$$\alpha_{1,j}, \alpha_{2,j}, ...\alpha_{m,j} = \text{softmax}(\beta_{1,j}, \beta_{2,j}, ...\beta_{m,j}), \qquad j = 1, 2, ..., h, \tag{9}$$

where h is the number of attention heads (discussed at the end of this paragraph). The softmax normalization is the key step of attention mechanisms to capture the interactions of all the input genes. The softmax function is defined as:

$$\alpha_{i,j} = \exp\left(\beta_{i,j}\right) \Big/ \sum_{i'=1}^{m} \exp\left(\beta_{i',j}\right), \qquad i = 1, 2, ..., m. \tag{10}$$

The final weights of the self-attention mechanism are:

$$\alpha_i = \sum_{j=1}^{h} \alpha_{i,j} = \alpha_{i,1} + \alpha_{i,2} + ... + \alpha_{i,h}, \qquad i = 1, 2, ..., m. \tag{11}$$

Note that in equation 8-11 we implemented the multi-head self-attention mechanism with h attention heads, instead of single-head self-attention. We used multi-head because single-head often pays most weight to a single gene embedding, while in practice, a few genes might be all-important, which could be selected through multiple independent heads, thus improving both the performance and interpretability of the model.

3.3. CADRE: Contextual Attention-based Drug Response

Although self-attention was already able to capture the contextual effects from other expressed genes to encode the cell line embedding, we hypothesized that by integrating the contextual information of drug targets, we could further improve the performance, leading to CADRE (**C**ontextual **A**ttention-based **D**rug **RE**sponse). Given the drug d and its target pathway p, instead of just using gene embeddings to calculate attention weights (equation 7), CADRE uses target pathway embedding $\boldsymbol{e}_p$ as well:

$$\alpha_1, \alpha_2, ..., \alpha_m = \text{Contextual-Attention}\left(\boldsymbol{e}_1, \boldsymbol{e}_1, ..., \boldsymbol{e}_m, \boldsymbol{e}_p\right). \tag{12}$$

All the steps to calculate attention weights in CADRE/Contextual-Attention are the same as SADRE/Self-Attention (equation 9-11), except equation 8, where CADRE calculates the unnormalized contextual-attention weights in the following way (figure 1d):

$$\beta_{i,j} = \boldsymbol{\theta}_j^{\intercal} \tanh(W\boldsymbol{e}_i + \boldsymbol{e}_p), \qquad i = 1, 2, ..., m, \;\; j = 1, 2, ..., h, \tag{13}$$

where $\boldsymbol{e}_p \in \mathbb{R}^q$ is the target pathway embedding of the pathway p. We mapped it from the lookup table of pathway embeddings $\mathcal{E}_{\mathcal{P}} = \{\boldsymbol{e}_p\}_{p \in \mathcal{P}}$, where $\mathcal{P}$ is the set of all possible pathways. Essentially, the drug target embedding $\boldsymbol{e}_p$ reflects the functional similarities of drugs, i.e., if two drugs share the same target, their target embeddings should be similar, leading to a similar function to calculate the attention weights from gene embeddings (equation 13). It is possible to directly use the drug embeddings here by replacing $\boldsymbol{e}_p$ with $V\boldsymbol{e}_d$ in equation 13, where $V \in \mathbb{R}^{q \times s}$ is a trainable model parameter. However, we did not find that this significantly improves performance.

3.4. Pretraining Gene Embeddings

Transfer learning research in the area of natural language processing and computational biology showed that word embeddings (Mikolov et al., 2013; Tao et al., 2019a) or gene embeddings (Tao et al., 2020a) pretrained on large-scale external unlabeled datasets could

improve related supervised learning tasks. To integrate the external knowledge of the co-expression pattern of genes, we utilized gene embeddings $\mathcal{E}_{\mathcal{G}} = \{\boldsymbol{e}_g\}_{g \in \mathcal{G}}$ pretrained on a large-scale Gene Expression Omnibus (GEO; Barrett et al., 2012) database. It is also possible to utilize other genetic association databases (Wang et al., 2019). The 200-dimensional gene embeddings were pretrained using the gene2vec algorithm (Du et al., 2019), which is a variant of the word2vec algorithm (Mikolov et al., 2013) in the scenario of gene expression. The co-expressed genes would also be close in the pretrained gene embedding space, and thus had a similar effect to the model. The full CADRE model directly used the fixed pretrained gene embeddings $\mathcal{E}_{\mathcal{G}}$ and did not optimize them at the time of training. If gene embeddings were randomly initialized and trained, the model reduced to CADREΔpretrain. To make a fair comparison, we also used fixed pretrained gene embeddings in the collaborative filtering and SADRE models.

3.5. Implementation and Training Procedure

We implemented the CADRE model using PyTorch (Paszke et al., 2019). We trained the model using the one cycle policy with mini-batch momentum gradient descent on an AWS GPU instance `g4dn.12xlarge` (Smith, 2018). To facilitate a balanced training process, we used a training batch size of $8 \times |\mathcal{D}|$, where $\mathcal{D}$ is the set of all the drugs. One cycle policy enabled fast convergence while preventing overfitting (superconvergence) by adjusting the learning rate and momentum. In the first 45% training steps (warm-up), we increased the learning rate linearly from $\eta/10$ to η, and decreased the momentum linearly from 0.95 to 0.85. In the following 45% training steps (cool-down), we decreased the learning rate linearly from η to $\eta/10$, and increased the momentum linearly from 0.85 to 0.95. In the last 10% training steps (annihilation), we decreased the learning rate linearly from $\eta/10$ to $\eta/100$, while keeping the momentum as 0.95. See section 5.1 for our tuning and evaluation protocols. The tuned hyperparameters of each dataset, such as maximum learning rate η, are available in table A2. We summarize the trainable/optimizable and fixed parameters of each model in table A1.

4. Datasets

4.1. Cohort Selection

We focused on two large-scale pharmacogenomic datasets: GDSC (Yang et al., 2013) and CCLE (Barretina et al., 2012). Both datasets included drug response data between hundreds of cell lines and tens/hundreds of anti-cancer drugs. We extracted the transcriptome data, i.e., the expression profiles of around 20k genes, as they were shown in previous multi-omics research to be the most dominant features compared with genomic and epigenomic data (Chiu et al., 2019; Sharifi-Noghabi et al., 2019). We also collected and summarized the targeted pathways of drugs of both datasets. The statistics of processed datasets are shown in table A3. Interested readers may refer to the original papers for more detailed characteristics of the two datasets (Yang et al., 2013; Barretina et al., 2012). We also list the data source in table A4.

4.2. Drug Sensitivity Discretization and Missing Value Imputation

GDSC and CCLE released both the half-maximal inhibitory concentration (IC50) and area under the curve (AUC)/activity area (AA) as the single continuous response value for each pair of cell line and drug. We discretized the AA into two categories of sensitive (one) vs. resistant (zero) using the waterfall algorithm for each drug (Barretina et al., 2012), following parameters in the previous work (Ding et al., 2018). Continuous IC50 (Chiu et al., 2019) and continuous AA (Barretina et al., 2012) were also widely used drug sensitivity measurements in the literature. However, AA is a more robust metric of sensitivity compared with IC50, which is crucial in our case since drug sensitivity data are usually noisy. In addition, binary sensitivity has a better clinical significance compared with the continuous one.

Although most of the response data of CCLE are available, 20% of response entries in the GDSC dataset are missing. At the time of training the model, if the sensitivity of a cell line to a drug was missing, we filled the missing value with the mode of the available sensitivities to this specific drug. At the time of evaluation, we skipped the missing values. There might exist alternative strategies, such as imputation with the mode of the k-nearest neighbors (Beretta and Santaniello, 2016). Another potential solution would be to modify the models by using a mask at the training phase to omit the unknown objective loss that resulted from them (Tao et al., 2019b). However, our preliminary experiments found that filling the values at the training stage could improve the performance on the validation set slightly.

4.3. Gene Expression Data Preprocessing

We downloaded the RNA expressions of cell lines in both GDSC and CCLE datasets. We calculated the variance of each gene using the quantile-normalized values in log-scale, and selected the 3,000 genes that had the largest variances. Within each cell line, we then annotated the top 1,500 highly expressed genes with ones, and the remaining genes with zeros. The parameters 3,000 and 1,500 were inherited from previous research (Ding et al., 2018).

4.4. Drug Target Pathway Extraction

We directly extracted the "target pathway" of each drug in the GDSC dataset. For the CCLE dataset, 15 out of its 24 drugs were shared with the GDSC dataset. Therefore, we used the same "target pathways" values for these 15 drugs in CCLE. For the remaining 9 drugs, we used their "class" as target pathways. One can find the mapping table of the drug target pathway of the CCLE dataset in table A5.

5. Results

5.1. Evaluation Approach

We trained and evaluated the models on the two datasets separately. For each dataset, we split the cell lines into three parts: training, validation, and test sets with a ratio of 60%, 20%, and 20%. All the models compared in this work shared the same split of datasets. We manually tuned all the models using the training and validation sets to optimize the

Table 1: Performance of different models and variants on the GDSC and CCLE datasets. We calculated the means and standard deviations from three repeated experiments. CADRE, SADRE, and collaborative filtering each utilized pretrained gene emebeddings. CADRE outperformed all the other competing models using four different metrics, validating its superior performance from the contextual attention mechanism and pretrained gene embeddings.

Dataset	Model	F1 Score	Accuracy	AUPR	AUROC
GDSC	DeepDR	62.1±0.55	78.4±0.74	67.2±0.91	81.8±0.92
	Collaborative filtering	60.9±0.18	76.9±0.37	67.0±1.36	81.2±0.13
	SADRE	62.9±0.20	78.2±0.25	68.7±1.43	82.9±0.29
	CADREΔpretrain	62.6±0.64	77.4±0.25	69.8±2.53	82.3±0.46
	CADRE	**64.3**±0.22	**78.6**±0.34	**70.6**±1.30	**83.4**±0.19
CCLE	DeepDR	51.5±2.95	77.8±0.39	53.5±2.80	78.5±1.09
	Collaborative filtering	48.6±1.26	77.6±0.42	48.9±1.10	78.6±0.52
	SADRE	50.1±1.61	77.5±0.31	56.4±2.02	78.7±0.80
	CADREΔpretrain	50.8±1.95	76.6±0.71	49.7±2.85	74.1±0.98
	CADRE	**54.4**±2.15	**79.1**±0.56	**60.0**±2.43	**80.9**±0.25

overall F1 score (See table A2 for tuned parameters). After we tuned the hyperparameters, we finally trained the models on the training and validation sets, and evaluated on the test set. Since the outputs of the models were slightly different across each run, we retrained and evaluated three times, and reported the mean and standard deviation. At the time of evaluation, we utilized multiple metrics in addition to the F1 score, including accuracy, area under the precision-recall curve (AUPR), and area under ROC (AUROC). AUPR and AUROC are more comprehensive evaluation metrics, which take as input the predicted probability instead of binary predictions, in contrast to the F1 score and accuracy.

5.2. CADRE Outperforms Competing Models

We compared the overall performance of CADRE (section 3.3) with other competing models and its ablated variants, including DeepDR (Chiu et al., 2019), vanilla collaborative filtering (section 3.1), SADRE (section 3.2), and CADREΔpretrain (CADRE without pretrained gene embeddings; section 3.4). DeepDR was previously shown to outperform both simple neural networks and classical models such as linear regression and SVM (Chiu et al., 2019). Note that CADRE, SADRE, and collaborative filtering each utilized pretrained gene embeddings in our experimental setting (table A1). DeepDR did not use pretrained gene embeddings, since we did not find it helpful in the preliminary experiments. As one can see from table 1, CADRE outperforms all these other models and variants on both GDSC and CCLE datasets. The attention mechanisms can significantly improve the performance of collaborative filtering. The contextual attention performs better than the self-attention, indicating the improved performance brought by the extra contextual information from the drug target pathway. The use of pretrained gene embeddings is also a crucial module for

the superior performance of CADRE, validating the transferred knowledge from external databases.

5.3. Effective Attention-Encoded Cell Line Representation Contributes to the Major Improvements of Performance

Using aggregated and flattened predictions and ground truth sensitivities, we showed the overall superior performance of CADRE over competing models (table 1). However, we were also concerned about the disentangled performance of individual cell lines to all drugs (per-cell-line performance), or individual drugs to all cell lines (per-drug performance). Since different cell lines or drugs could have distinct performances, we compared the distributions of their performances instead of single averaged values. As one can see from figure 2, all the methods had comparable and reasonably high AUPR per cell line. However, the attention-based models such as SADRE and CADRE significantly outperformed other models in per-drug ARPR. This indicated that the major improvements of the attention-based models might come from the useful cell line representations built by attention mechanisms, such that the cell lines with various expression profiles were appropriately distinguished.

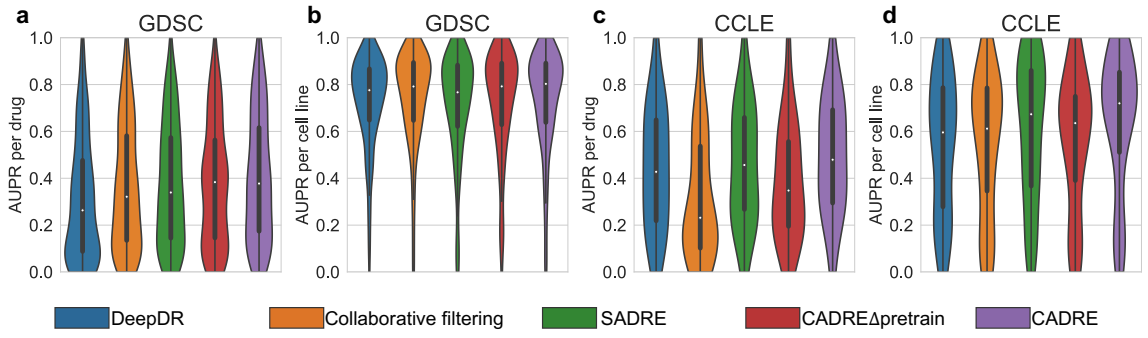

Figure 2: The distributions of dissected AUPR per cell line and AUPR per drug in different models on both GDSC and CCLE datasets. The per-cell-line performances of all models achieved reasonable high levels. Meanwhile, the major improvements of CADRE and SADRE came from the per-drug performance, indicating the well-designed attention mechanisms might encode the cell lines more effectively, so that different cell lines were more easily distinguished, leading to high per-drug AUPR.

To validate that attention-encoded cell line embeddings are more effective than non-attention-encoded vanilla cell line embeddings, we calculated the "correlation" of cell line embeddings and the origins of cell lines using "NN accuracy". NN accuracy is defined as the following expectation:

$$\text{NN accuracy} = \mathbb{E}_{c':\boldsymbol{e}_{c'}=\text{NN}(\boldsymbol{e}_c)}\left[\text{Tissue}(c') = \text{Tissue}(c)\right], \tag{14}$$

where $\text{Tissue}(c)$ outputs the tissue of cell line c and $\text{NN}(\boldsymbol{e}_c)$ returns the closest gene embedding to $\boldsymbol{e}_c$ using unnormalized cosine similarity. We approximated this expectation by

iterating c over all the cell lines in the dataset. The NN accuracy reflects how the distribution of cell line embeddings is consistent with their tissues. We focused on the GDSC dataset, since it had a larger sample size, and provided comprehensive annotations of the cell lines. As one can see from table A6, the attention-based cell line embeddings had a better correlation with the tissue type (NN accuracy=36.1%), compared with vanilla cell line embeddings (NN accuracy=20.2%) and random cases (shuffled and repeated for five times; NN accuracy=12.9±2.1%). The t-SNE plot (van der Maaten and Hinton, 2008) of attention-encoded cell line embeddings revealed distinct clusters of embeddings within the same tissue, and similar embeddings across different tissues (figure 3a). In contract, non-attention-encoded cell line embeddings only reflected similarities of tissues (figure 3b), and did not discover the similar subgroups observed in attention-encoded embeddings.

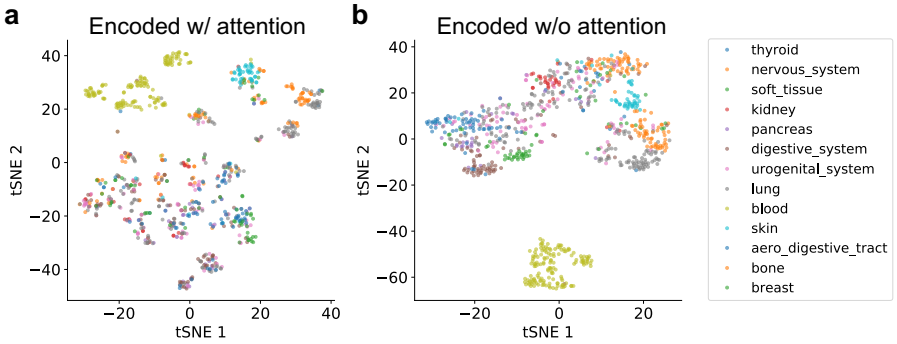

Figure 3: t-SNE visualization of cell line embeddings in the GDSC dataset. (**a**) CADRE-encoded cell line embeddings. The groups within the same tissue revealed the potential subtypes of the cancer that exhibited distinct drug response profiles. At the same time, even cell lines from different tissues can merge into the same clusters and thus share similar responses. (**b**) Cell line embeddings encoded without attention mechanism. These mainly reflected tissue-specific expression patterns rather than differences in response profiles, without finding subgroups of cell lines as in attention-encoded embeddings.

5.4. CADRE Identifies the Critical Biomarkers Related to Drug Resistance

CADRE assigned the heaviest weights to the genes that were most important to the drug response, providing a way to identify critical gene expression markers and biological processes. We extracted the attention weights from CADRE and counted expression frequency of the genes, and plotted the normalized attention weights vs. expression frequency. We shuffled the attention weights randomly 1,000 times to infer the significantly attended genes with a p-value threshold of 0.01 (Zaitsev et al., 2019). We mainly focused on the GDSC dataset due to its larger sample size. In general, the essential genes identified by CADRE are independent of the expression frequency (figure 4). The frequently expressed genes did not necessarily receive higher attention weights.

We then conducted Gene Ontology (GO) enrichment analysis on these significant genes (Mi et al., 2013). Two primary cell activities emerged (table A7). First, functions related to exporting the molecules from the cells were enriched, which is consistent with previous research that many cancer cells acquired drug resistance by expelling the compounds using microvesicles (Muralidharan-Chari et al., 2016). Secondly, functions related to signaling receptor binding were enriched, reflecting the fact that a lot of anti-cancer drugs are targeted to the receptors of specific signaling pathways such as EGFR and RTK (Yang et al., 2013). The clinically actionable genes of these CADRE-identified biomarkers could be potential targets of anti-cancer compounds for future exploration.

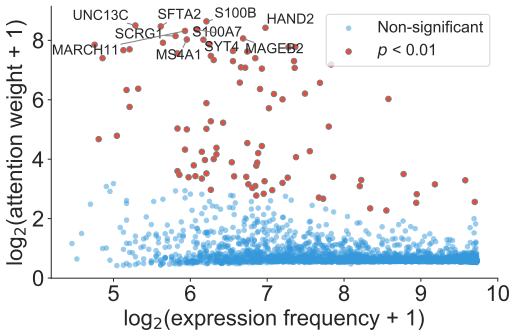

Figure 4: Landscape of the normalized contextual-attention weights of genes and their expression frequencies in the GDSC dataset. A higher expression frequency did not guarantee a higher attention weight.

6. Discussion

Personalized medicine in oncology requires that researchers and clinicians have tools to suggest effective anti-cancer drugs based on complex molecular profiles of the tumor and noisy pharmacogenomic assays, and to provide reasonable explanations for the recommendations. In pursuit of this goal, we created CADRE, an interpretable machine learning model that accurately predicted drug sensitivities of cancer cell lines from their expression levels. CADRE was built upon collaborative filtering, which is capable of dealing with noisy response assay data. The attention mechanism of CADRE improved both interpretability and performance of the model, by capturing the interactions and contextual effects of genes and drugs, and by encoding a better representation of cell lines from raw expression profiles. What is more, CADRE utilized gene representations transferred from an external database to boost its performance further. Through extensive evaluations and comparisons on the two primary pharmacogenomic datasets, we validated the superior performance of CADRE over competing models. Our results indicate that the genes assigned significant attention are involved in biological processes that can be expected to impact cellular responses to the presence of drugs. Thus these genes are potential novel biomarkers for designing more efficient test panels than whole-genome-scale sequencing.

Limitations and Future Directions Our model considered the simplified scenario of cancer cell lines, where each sample consists of only one cell population. However, a tumor tissue from a single cancer patient usually consists of multiple subpopulations, each exhibiting a distinct molecular profile (Schwartz and Schäffer, 2017; Tao et al., 2020b). A clinically applicable anti-cancer drug resistance model should not only consider the inter-tumor differences between cell lines or patients (CADRE considered the similarities/differences between cell lines by grouping them into subtypes in the embedding space; figure 3a), but also take into account and deconvolve the intra-tumor heterogeneity of the cell populations within the same tumor tissue. Network matching (Liu et al., 2020) or single-cell techniques (Lei et al., 2020) could be promising directions in bridging both sides of *in vitro* cancer cell lines and *in vivo* tumors. A few other promising directions also warrant pursuing in the future. We mainly validated the effectiveness of CADRE using RNA expression data of cell lines in this work. However, we expect that models similar to CADRE, with slight modifications, could apply to genomic or epigenomic data, or the combination of these omic data in the future. Although we incorporated the knowledge of drugs through their target pathways, other drug representations such as drug embeddings well-represented from their structures, such as inferred from fingerprints or SMILES, might further improve our model (Zhang et al., 2015). Finally, since CADRE integrates pathway information to capture the contextual information, it would be helpful to conduct a case study of the essential genes captured by attention mechanism in specific drug target pathways, in addition to the aggregated analysis (section 5.4; figure 4).

Acknowledgments

Y.T. was partially supported by Center for Machine Learning and Health Fellowship in Digital Health from Carnegie Mellon University. X.L. was partially supported by NIH award R01LM012011. M.Q.D. was supported by NLM grant 5T15LM007059-33. R.S. was partially supported by NIH awards R21CA216452 and R01HG010589 and the Mario Lemieux Foundation. This work was also partially supported by the AWS Machine Learning Research Awards granted to R.S. and X.L. The content does not necessarily represent the official views of the above funding agencies.

References

Jordi Barretina, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483:603, mar 2012.

Tanya Barrett, et al. NCBI GEO: archive for functional genomics data setsupdate. *Nucleic Acids Research*, 41(D1):D991–D995, 2012.

Lorenzo Beretta et al. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making*, 16 Suppl 3(Suppl 3):74, jul 2016.

Kyle Chang, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

Yoosup Chang, et al. Cancer Drug Response Profile scan (CDRscan): A deep learning model that predicts drug effectiveness from cancer genomic signature. *Scientific Reports*, 8(1):8857, 2018.

Yu-Chiao Chiu, et al. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Medical Genomics*, 12(1):18, 2019.

James C Costello, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, 32:1202, jun 2014.

Michael Q Ding, et al. Precision oncology beyond targeted therapy: Combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. *Molecular Cancer Research*, 16(2):269–278, feb 2018.

Zuoli Dong, et al. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. *BMC Cancer*, 15(1):489, 2015.

Jingcheng Du, et al. Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics*, 20(1):82, 2019.

Mathew J Garnett, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–575, 2012.

Paul Geeleher, et al. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biology*, 15(3):R47, mar 2014.

Mehmet Gonen et al. Drug susceptibility prediction against a panel of drugs using kernelized bayesian multitask learning. *Bioinformatics (Oxford, England)*, 30(17):i556–63, sep 2014.

Xiao He, et al. Kernelized rank learning for personalized drug recommendation. *Bioinformatics (Oxford, England)*, 34(16):2808–2816, aug 2018.

Haoyun Lei, et al. Tumor heterogeneity assessed by sequencing and fluorescence in situ hybridization (FISH) data. *bioRxiv*, 2020. doi: 10.1101/2020.02.29.970392.

Hui Liu, et al. Anti-cancer drug response prediction using neighbor-based collaborative filtering with global effect removal. *Molecular Therapy - Nucleic Acids*, 13:303–311, 2018.

Qingzhi Liu, et al. Network-based matching of patients and targeted therapies for precision oncology. *Pacific Symposium on Biocomputing*, 25:623–634, 2020.

Huaiyu Mi, et al. Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols*, 8:1551, jul 2013.

Tomas Mikolov, et al. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, pages 3111–3119, 2013.

Vandhana Muralidharan-Chari, et al. Microvesicle removal of anticancer drugs contributes to drug resistance in human pancreatic cancer cells. *Oncotarget*, 7(31):50365–50379, aug 2016.

Ali Oskooei, et al. PaccMann: Prediction of anticancer compound sensitivity with multi-modal attention-based neural networks, 2018.

Adam Paszke, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8026–8037. 2019.

Vinay Prasad. Perspective: The precision-oncology illusion. *Nature*, 537(7619):S63–S63, 2016.

Nolan Priedigkeit, et al. Intrinsic subtype switching and acquired ERBB2/HER2 amplifications and mutations in breast cancer brain metastases. *JAMA Oncology*, 3(5):666–671, may 2017.

Jason A Reuter, et al. High-throughput sequencing technologies. *Molecular Cell*, 58(4): 586–597, may 2015.

Gregory Riddick, et al. Predicting in vitro drug sensitivity using random forests. *Bioinformatics (Oxford, England)*, 27(2):220–224, jan 2011.

J Ben Schafer, et al. *Collaborative Filtering Recommender Systems*, pages 291–324. 2007.

Russell Schwartz et al. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics*, 18:213, feb 2017.

Hossein Sharifi-Noghabi, et al. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*, 35(14):i501–i509, jul 2019.

Robert H Shoemaker. The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer*, 6(10):813–823, 2006.

Leslie N. Smith. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay, 2018.

Nitish Srivastava, et al. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Yifeng Tao, et al. Effective feature representation for clinical text concept extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 1–14, Minneapolis, Minnesota, USA, jun 2019a.

Yifeng Tao, et al. Phylogenies derived from matched transcriptome reveal the evolution of cell populations and temporal order of perturbed pathways in breast cancer brain metastases. In *Mathematical and Computational Oncology*, pages 3–28, 2019b.

Yifeng Tao, et al. From genome to phenome: Predicting multiple cancer phenotypes based on somatic genomic alterations via the genomic impact transformer. In *Pacific Symposium on Biocomputing*, volume 25, pages 79–90, 2020a.

Yifeng Tao, et al. Robust and accurate deconvolution of tumor populations uncovers evolutionary mechanisms of breast cancer metastasis. *Bioinformatics*, 36(Supplement_1): i407–i416, jul 2020b.

Laurens van der Maaten et al. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Haohan Wang, et al. Automatic human-like mining and constructing reliable genetic association database with deep reinforcement learning. In *Pacific Symposium on Biocomputing*, volume 24, pages 112–123. World Scientific, 2019.

Lin Wang, et al. Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC Cancer*, 17(1):513, aug 2017.

Dong Wei, et al. Comprehensive anticancer drug response prediction based on a simple cell line-drug complex network model. *BMC Bioinformatics*, 20(1):44, 2019.

Kelvin Xu, et al. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

Wanjuan Yang, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(Database issue):D955–61, jan 2013.

Zichao Yang, et al. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, jun 2016.

Han Yuan, et al. Multitask learning improves prediction of cancer drug sensitivity. *Scientific Reports*, 6(1):31619, 2016.

Konstantin Zaitsev, et al. Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nature Communications*, 10(1):2209, 2019.

Naiqian Zhang, et al. Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. *PLoS Computational Biology*, 11(9):e1004498, 2015.

Appendix

Table A1: Trainable/optimizable and fixed parameters in the different models. If the gene embeddings $\mathcal{E}_\mathcal{G}$ were pretrained, it would be fixed during training. The pathway embeddings $\mathcal{E}_\mathcal{P}$ in both the CADREΔpretrain and CADRE models were randomly initialized by default in PyTorch and learned from the data. See section 3.1-3.3 for details and definitions of these models and parameters.

Model	Trainable parameters ($\mathcal{W}$)	Fixed parameters
Collaborative filtering	$\mathcal{E}_\mathcal{D}$	$\mathcal{E}_\mathcal{G}$
SADRE	$\mathcal{E}_\mathcal{D} \cup W \cup \{\boldsymbol{\theta}_j\}_{j=1}^h$	$\mathcal{E}_\mathcal{G}$
CADREΔpretrain	$\mathcal{E}_\mathcal{D} \cup \mathcal{E}_\mathcal{G} \cup \mathcal{E}_\mathcal{P} \cup W \cup \{\boldsymbol{\theta}_j\}_{j=1}^h$	$\emptyset$
CADRE	$\mathcal{E}_\mathcal{D} \cup \mathcal{E}_\mathcal{P} \cup W \cup \{\boldsymbol{\theta}_j\}_{j=1}^h$	$\mathcal{E}_\mathcal{G}$

Table A2: Tuned hyperparameters of CADRE model in GDSC and CCLE datasets. See section 3 for definitions and details of the hyperparameters.

	Dataset	
Hyperparameter	GDSC	CCLE
training batch size	8×260	8×24
training steps	48k	96k
maximum learning rate η	0.3	0.05
number of attention heads h	8	8
weight decay λ_2	3e-4	3e-4
dropout rate ρ	0.6	0.5
drug/gene/cell-line embedding dimension s	200	200
drug target pathway embedding dimension q	128	100

Table A3: Statistics of the two drug response datasets used in the work.

Dataset	# cell lines	# drugs ($\lvert\mathcal{D}\rvert$)	# pathways ($\lvert\mathcal{P}\rvert$)	% missing	% positive
GDSC	846	260	25	18.2%	32.2%
CCLE	409	24	11	3.5%	24.8%

Table A4: Downloading websites and file names of GDSC and CCLE data used in the work.

Dataset	Type	URL/Filename
GDSC	Download URL	`www.cancerrxgene.org/downloads/bulk_download`
	Drug sensitivity	`GDSC2_fitted_dose_response_25Feb20.xlsx`
	Gene expression	`Cell_line_RMA_proc_basalExp.txt`
	Target pathway	`Drug_listSun.csv`
CCLE	Download URL	`data.broadinstitute.org/ccle`
	Drug sensitivity	`CCLE_NP24.2009_Drug_data_2015.02.24.csv`
	Gene expression	`CCLE_RNAseq_rsem_genes_tpm_20180929.txt`
	Target pathway	`CCLE_NP24.2009_profiling_2012.02.20.csv`

Table A5: Cleaned mapping table from drugs to their target pathways in CCLE dataset.

Drug	Target pathway
Erlotinib	EGFR signaling
Lapatinib	EGFR signaling
PHA-665752	RTK signaling
PF-2341066	RTK signaling
TAE684	RTK signaling
Vandetanib	Other kinase inhibitor
Nilotinib	ABL signaling
AZD0530	RTK signaling
Sorafenib	RTK signaling
TKI258	Other kinase inhibitor
PD-0332991	Cell cycle
AEW541	Other kinase inhibitor
RAF265	Other kinase inhibitor
PLX4720	ERK MAPK signaling
PD-0325901	ERK MAPK signaling
AZD6244	ERK MAPK signaling
Nutlin-3	p53 pathway
LBW242	Other
17-AAG	Protein stability and degradation
L-685458	Other
Panobinostat	Other
Paclitaxel	Mitosis
Irinotecan	Other cytotoxic
Topotecan	Other cytotoxic

Table A6: NN accuracy of differently encoded cell line embeddings with respect to tissue type. CADRE-encoded cell line embeddings improved the NN accuracy by around 80% compared with the embeddings encoded without attention, indicating the attention mechanism enabled the cell line embeddings to achieve a higher correlation with their corresponding tissues.

Cell line embedding	NN accuracy (%)
Random	12.9±2.1
Encoded w/o attention	20.2
Encoded w/ attention	**36.1**

Table A7: Enriched biological functions of significantly weighted genes in GDSC dataset (red dots in figure 4). Genes related to intracelluar vesicles exporting compounds from cells, and signaling receptor bindings were heavily picked by CADRE.

GO domain	Enriched functions	FDR
biological process	export from cell	2.59e-3
biological process	secretion	2.84e-4
biological process	leukocyte activation	2.13e-2
molecular function	signaling receptor binding	6.24e-3
cellular component	intracelluar vesicle	9.64e-3
cellular component	vesicle lumen	4.79e-2
cellular component	extracellular region	1.05e-3