This is ©Bioinformatics Journal, 2024, pp. 1–12

doi: DOI HERE
MRes Bioengineering Thesis
Paper

PAPER

Self-Supervised Disentangled Omics Representation learning for Robust Drug Response Prediction

Yuanjie Zou[1] and Pedro Ballester[1, *]

[1]Department of Bioengineering, Imperial College London, UK
[*]Corresponding author. p.ballester@imperial.ac.uk

Abstract

The rapid development of high-throughput technologies has revolutionized the field of drug response prediction, with omics data emerging as a new paradigm. However, discovering biomarkers for all cancer types and drugs remains challenging due to the heterogeneity of cancer and the scarcity of labeled pharmacogenomic datasets. Additionally, omics data is highly noisy due to technical and biological confounders that can entangle with the desired signal. To address these challenges, we propose a pre-training disentangled omics representation model (DOR) for drug response prediction. DOR employs specialized encoders to divide samples into salient (informative) and mutual (confounding) embeddings, and a carefully designed reconstruction loss function to make the salient embedding capture the most critical information. This approach can disentangle confounders and align data from different domains in the latent space. We evaluated the performance of DOR on the TCGA database, consisting of 10 drugs and 5 types of omics data. DOR outperformed competing tree-based, linear or non-linear, and pre-training models, achieving the top rank of prediction. DOR also demonstrated superior domain adaptation ability when trained and tested on gender-specific subsets. Finally, the low-dimensional salient embeddings of DOR enabled effective multi-omics integration for improved predictions. These results highlight DOR's potential for expression-based precision oncology.

Key words: Drug Response Prediction, Disentangled Representation Learning, Domain Adaptation, Multi-omics Integration, Precision Oncology

Introduction

The advent of high-throughput technologies has started a new phase in the field of drug response prediction, with omics data emerging as the dominant paradigm [1, 2, 3, 4, 5, 6, 7]. This technological revolution has enabled the mining of valuable information from various types of omics data, which can provide crucial knowledge that determines the anti-cancer drug response to patients. A notable example of this is the methylation status of the MGMT promoter, a well-established biomarker of response to Temozolomide that has been widely adopted in clinical decision-making procedures [8]. However, it is impossible to discover efficient biomarkers for all cancer types and drugs due to the inherent heterogeneity of cancer [9]. Such heterogeneity leads to different gene activity patterns across various locations and times, even for the same patient, making it difficult to identify commonly applicable biomarkers [10]. As a result, only a small subset of drugs and cancers currently have known genetic biomarkers, which limits the application of precision oncology approaches in clinical practice [11].

To address this issue, machine learning approaches have been employed to analyze whole gene expression profiles, aiming to construct a more complex, comprehensive, and diverse representation of a patient's status for classification purposes [12, 13, 14, 15]. Nevertheless, compared to the extremely high dimension of omics data ($\geq$20k features), the current pharmacogenomic datasets with clinical drug response labels ($\leq$300 samples) are scarcity, leading to over-fitting and the curse of dimension problem. To solve this critical problem, a modern strategy is to leverage label-sufficient ex-vivo data, such as cell-line and patient-derived xenograft (PDX), to train models and subsequently transfer the learned knowledge to the patient domain [16]. Recently, foundation models using large-scale unlabeled samples for pre-training have demonstrated promise in generating superior embeddings that may benefit downstream drug response prediction tasks [17]. These advancements in machine learning techniques offer potential solutions to the challenges posed by the high dimension of omics features and the scarcity of labeled pharmacogenomic data problem, showing the way for more accurate and robust drug response prediction models in precision oncology.

However, label scarcity is not the only challenge that omics data have. The high dimension of omics features also introduces the problem of noisy data [18]. It has been proved that omics data contains a significant level of noise, which can be attributed to both technical confounders (e.g. batch effects), and biological confounders (e.g. sex) [19]. Ideally, models should be able to predict response robustly under

different confounding factors. However, these confounders often entangled with the signal we desired, causing models to learn information that is not directly related to the target [20]. Typical databases collect data from diverse sources, containing different regions, sexes, and gene sequencer types. Consequently, training a machine learning model using data from multiple confounders becomes a challenging task [21]. Both technical and biological confounders can be treated as the domain adaptation problem [22]. SAVER-X [23] is an early work that addresses the issue of noisy data in omics datasets by employing a Bayesian model to decompose the structure of the data. However, SAVER-X requires external gene-gene correlation data, which may not always be available. AttentionMOI [24] proposed a bootstrap-based feature selection method to select highly informative features. Nevertheless, such bootstrap algorithms are computationally expensive and not suitable adapted to large-scale datasets. A more recent approach, Velodrome [25], proposed aligning embeddings from different domains using contrastive loss. Although Velodrome is proven to remove systematic confounders such as those present in in-vitro and ex-vivo data, it cannot effectively deal with the inherent noise such as sex factor in the data itself. ADAE [26] utilized an adversarial model to align domains from different datasets. However, it requires explicit confounders labels such as dataset for adversarial training, which may not be used for many unknown potential confounders. VAEN [27] was developed to facilitate pre-training learning task with unlabeled data, but it does not demonstrate the ability to remove confounding.

Recent advancements in embedding learning have developed many efficient methods for information extraction and representation. Notably, CODE-AE [28] introduced a domain separation network (DSN) model [29], which is designed to distinguish between specific and common information across two different domains. This innovative approach has inspired us to explore the use of Deep Subspace Network (DSN) architecture to extract essential information from a single dataset. Specifically, we aim to identify the most representative features after removing confounding variables through a disentanglement process. We hypothesize that, when presented with two identical datasets, the DSN should still be able to output specific and common view of data for this single dataset. This concept of information disentanglement within a single domain has already been studied by previous research. For instance, the contrastive learning approach employed by cVAE [30, 31] points out that a sample can be disentangled into two distinct views: a salient view containing the sample's specific information, and a mutual view containing the common background information shared across datasets. Following this framework, recent studies such as SwitchTab [32] and Swap AutoEncoder [33] have proposed a novel training procedure aimed at disentangling salient and mutual embeddings from pre-training data. These approaches have demonstrated the potential to generate high-quality embeddings that can be effectively utilized in various downstream tasks.

Inspired by DSN and SwitchTab, we propose a novel pre-training disentangled omics representation (DOR) model for drug response prediction. Our approach begins with self-supervised pre-training on both labeled and unlabeled data. During this phase, two specialized projector map each sample into salient and mutual embeddings. A reconstruction loss function then guides the projector on how to disentangle data. Subsequently, we fine-tune the DOR model using the salient embedding on the labeled dataset. Unlike conventional pre-training models that attempt to reconstruct all details from the latent embedding, DOR requires accurate reconstruction using only combined salient and mutual embeddings. In this approach, the salient embedding retains only critical information. In summary, This strategy potentially reduces omics data dimension and disentangles meaningful signals from noise, leveraging both labeled and unlabeled data. Through disentangled learning, the salient embedding demonstrates enhanced generalizability across diverse domains and improved predictive accuracy on unseen data, aligning with the objectives of precision oncology.

To evaluate the efficacy of our model, we conducted comprehensive comparative analyses of the drug response prediction task for DOR against several competing models using the TCGA (The Cancer Genome Atlas) database, covering 10 drugs and 5 types of omics data. DOR's superior performance derives from its ability to learn disentangled representations, outperforming other models in comparative tests, including tree-based, linear, and pre-training-based approaches. Furthermore, we demonstrate that DOR performs better in domain adaptation tasks, particularly when the model is trained on gender-specific data and tested on data from the opposite gender. A visualized t-SNE embeddings will be provided for a better understanding of DOR's behavior. We also evaluate DOR's performance in multi-omics integration. The model's ability to effectively reduce high-dimensional omics data enabled us to integrate multiple omics data types and discover their interactions. In conclusion, our findings suggest that DOR holds significant potential for addressing drug response prediction tasks in the context of expression-based precision oncology.

Methodology

Overview

This study aims to develop a disentangled representation of gene expression features for enhanced drug response prediction through Disentangled Omics Representation (DOR) technique. The proposed architecture of the DOR model contains three essential components: an encoder that maps omics data into a latent space, two projectors that transform the latent features into salient (informative) and mutual (confounding) embeddings, and a decoder that reconstructs the original input from the combined salient and mutual embeddings, as depicted in Figure 1. Notably, the model employs an asymmetric design with a larger encoder compared to the decoder. The aim of this design is to enhance the encoder's ability to capture subtle patterns in the databased on existing research [34]. This innovative disentanglement approach aims to enhance both the predictive accuracy and domain adaptation ability of drug response models by effectively disentangling informative features from confounding factors for omics data.

Disentangled Representation Learning

In this section, we introduce details of disentangled representation learning model. Given a batch B of b samples with d dimensions, denoted as $B \in \mathbb{R}^{b \times d}$, we first split the batch into two equal sub-batches, B_1 and B_2, each of size $\frac{b}{2} \times d$. These two sub-batches could be treated as separate input domains within the original Domain Separation Network (DSN) framework. A shared encoder, f_e, is then used to encode

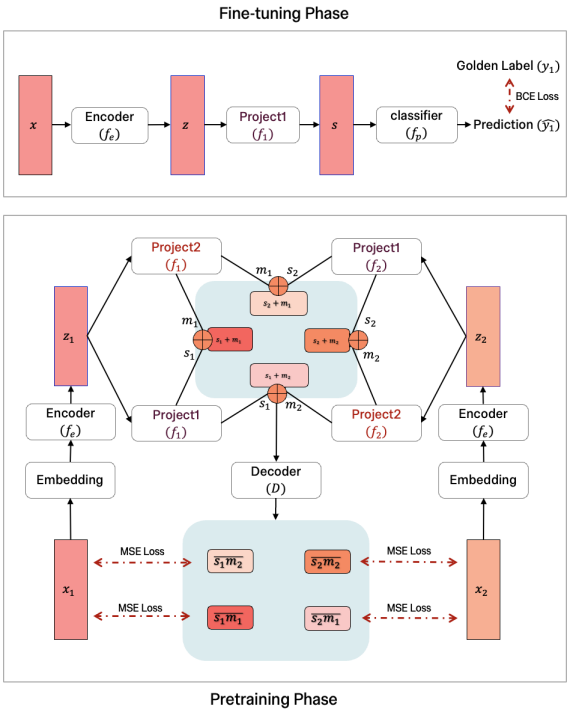


Fig. 1. Overview of the DOR model architecture. In the Pre-training Phase, the model processes input embeddings x1 and x2 through encoders and projects them into a shared latent space. Then those embedding will be combined each other and fed into the Decoder to reconstructs input. The Fine-tuning Phase have a classifier to predict the response label.

the two sub-batches into latent feature (encoder embedding) representations z_1 and z_2:

$$f_e : \mathbb{R}^{b\times d} \to \mathbb{R}^{b\times l}$$

$$z_1 = f_e(x_1) \in \mathbb{R}^{b\times l}, \quad z_2 = f_e(x_2) \in \mathbb{R}^{b\times l}$$

where l is the dimension of latent space. The encoder backbone can be flexibly chosen, including ResNet or a simple Multilayer Perceptron (MLP). For this study, we use MLP due to its simplicity and effectiveness. Next, the latent features are disentangled using two distinct projectors, f_1 and f_2, which generate four different embeddings:

$$f_1 : \mathbb{R}^{b\times l} \to \mathbb{R}^{b\times l}, \quad f_2 : \mathbb{R}^{b\times l} \to \mathbb{R}^{b\times l}$$

$$s_1 = f_1(z_1) \in \mathbb{R}^{b\times l}, \quad m_1 = f_2(x_1) \in \mathbb{R}^{b\times l}$$

$$s_2 = f_1(z_2) \in \mathbb{R}^{b\times l}, \quad m_2 = f_2(x_2) \in \mathbb{R}^{b\times l}$$

Here, s_1 and s_2 is the salient embeddings, and m_1 and m_2 is the mutual embeddings. Salient embeddings are expected to capture the most informative features, whereas mutual embeddings contain confounding and noisy information. To guide this disentanglement, the four embeddings are combined using an addition operation and then fed into a decoder D to reconstruct the input data:

$$D : \mathbb{R}^{b\times l} \to \mathbb{R}^{b\times d}$$

$$s_1\bar{m}_1 = D(s_1 + m_1), \quad s_2\bar{m}_1 = D(s_2 + m_1)$$

$$s_1\bar{m}_2 = D(s_1 + m_2), \quad s_2\bar{m}_2 = D(s_2 + m_2)$$

where $s_i\bar{m}_i$ is the reconstructed data. While either addition or concatenation can be used for combining embeddings, we choose addition here due to its lower parameter requirements. The reconstruction loss function for the pre-training phase is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{recon}} = &||\overline{s_1 m_1} - x_1||_2 + ||\overline{s_1 m_2} - x_1||_2 \\ &+ ||\overline{s_2 m_1} - x_2||_2 + ||\overline{s_2 m_2} - x_2||_2\end{aligned}$$

This loss function ensures that the reconstruction is primarily driven by the salient embeddings, guiding the model to capture the most representative information. This implies that both $s_1\hat{m}_1$ (original) and $s_1\bar{m}_2$ (switched) are expected to demonstrate comparable information of representation for x_1. Consequently, the salient feature s_1 should contain the most representative information of x_1 to reconstruct the input. Conversely, the mutual embedding is intended to be shared and interchangeable between the two sub-batches. The mutual embedding m_2 should not contain any information that could uniquely represent x_2. If it did, the decoder would be unable to determine whether to reconstruct x_1 or x_2 since the restriction of loss term $||s_1\bar{m}_2 - x_1||_2$. Similar expectations apply to the other sub-batch x_2. Through this mechanism, $\mathcal{L}_{\text{recon}}$ guides the model to disentangle salient and mutual information.

This representation learning approach does not require labeled data, enabling training on large-scale unlabeled datasets. However, unlike common pre-training models, indiscriminate addition of more pre-training data is not recommended for this disentangled representation learning method. As the volume of data grows, the potential for both additional confounders and increased mutual information rises accordingly. When incorporating a dataset unrelated to the desired downstream task, there is a risk that the true signal we seek might become a confounder in this unrelated dataset. Consequently, potentially useful information may be captured in the mutual embedding rather than the salient one.

Orthogonal Embedding

To encourage the projectors to disentangle omics data effectively, we introduce an additional loss term that promotes orthogonality between the salient and mutual embeddings:

$$\mathcal{L}_{\text{frobenius}} = \|s_1 s_2\|_F^2 + \|m_1 m_2\|_F^2$$

where $\|M\|_F$ is the Frobenius norm of a matrix M:

$$\|M\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}^2}$$

The total pre-training loss function thus becomes:

$$\mathcal{L}_{\text{pre-training}} = \mathcal{L}_{\text{recon}} + \lambda \times \mathcal{L}_{\text{frobenius}}$$

where λ is a hyperparameter that balances the contribution of the Frobenius loss.

Algorithm 1 DOR Pre-training

Require: N: batch size, λ: Frobenius loss coefficient

Initialize encoder, decoder, salient and mutual layers with given hyperparameters
for each sample x in dataset **do**
  $encoded \leftarrow \text{encoder}(x)$
  $x_{salient} \leftarrow \text{salient\_projector}(encoded)$
  $x_{mutual} \leftarrow \text{mutual\_projector}(encoded)$
  $x_{recon\_enc} \leftarrow \text{decoder}(encoded)$
  $x_{recon\_s1m1} \leftarrow \text{decoder}(x_{salient} + x_{mutual})$
  # Split x into two halves x_1, x_2
  $x_{salient1}, x_{salient2} \leftarrow \text{split}(x_{salient}, 2)$
  $x_{mutual1}, x_{mutual2} \leftarrow \text{split}(x_{mutual}, 2)$
  $x_{recon\_s1m1} \leftarrow \text{decoder}(x_{salient1} + x_{mutual1})$
  $x_{recon\_s1m2} \leftarrow \text{decoder}(x_{salient1} + x_{mutual2})$
  $x_{recon\_s2m1} \leftarrow \text{decoder}(x_{salient2} + x_{mutual1})$
  $x_{recon\_s2m2} \leftarrow \text{decoder}(x_{salient2} + x_{mutual2})$
  Calculate reconstruction losses: $\mathcal{L}_{recon}$
  if Frobenius loss is enabled **then**
    Compute Frobenius loss $\mathcal{L}_{frobenius}$ from salient and mutual representations
  end if
  Total loss ← sum of recon losses and Frobenius loss
  Backpropagate total loss
end for

Fine-Tuning

After pre-training, the DOR model is fine-tuned for drug response prediction using labeled data. The pre-trained salient embeddings f_p are then input into a binary classifier f_p:

$$f_p : \mathbb{R}^{b\times l} \to \mathbb{R}^{b\times 1}$$
$$\hat{y} = f_p(s)$$

The supervised cross-entropy loss $\mathcal{L}_{\text{sup}}$ is used to optimize the model for predicting the drug response $y \in [0, 1]$:

$$\mathcal{L}_{\text{sup}} = -\sum_{j=1}^{b} y_j \log\left(\hat{y}_j\right)$$

Multi-Omics Integration

In light of DOR could extract salient features in a low-dimensional space, we investigate the potential of integrating diverse omics data types to construct a multimodal model. The integration process can be implemented at various stages, categorized as early, middle, or late, as defined in [14]. Our study applies a middle-stage integration approach, where latent embeddings from distinct omics datasets are combined into a unified representation using a cross-attention mechanism. This methodology enables the model to leverage the complementary information inherent in different omics modalities, potentially enhancing its predictive ability and providing a more comprehensive understanding of complex biological systems.

For each omics dataset, the embeddings E_m are first transformed into query (Q), key (K), and value (V) matrices:

$$Q_m = W_Q E_m, \quad K_m = W_K E_m, \quad V_m = W_V E_m$$

where $W_Q, W_K, W_V \in \mathbb{R}^{l\times l}$ are learnable weight matrices. The attention score between omics i and omics j is computed as:

$$\text{Attention}(E_i, E_j) = \text{softmax}\left(\frac{Q_i K_j^\top}{\sqrt{l}}\right) V_j$$

The integrated representation for the i-th omics embedding, $\hat{E}_i$, is then obtained by summing the attended embeddings from all other omics datasets:

$$\hat{E}_i = \sum_{j=1}^{M} \text{Attention}(E_i, E_j)$$

Finally, to produce a unified representation $\hat{E}$ that captures information from all modalities, we aggregate the integrated embeddings $\hat{E}_m$:

$$\hat{E} = \frac{1}{M}\sum_{m=1}^{M} \hat{E}_m$$

This integrated embedding $\hat{E}$ is then used for downstream classification tasks.

Experiment Setting

Dataset and Pre-processing

In this study, we leveraged omics data and corresponding clinical drug response records obtained from The Cancer Genome Atlas (TCGA) database [35]. Our investigation focused on ten drugs: Carboplatin, Cisplatin, Cyclophosphamide, Docetaxel, Doxorubicin, Etoposide, Fluorouracil, Gemcitabine, Leucovorin, and Paclitaxel. We select those drugs since they have sufficient labels (#label $\geq$ 100) compared to other drugs in database. Drug response labels were extracted from files named as 'nationwidechildrens.org_clinical_drug_drug.txt'. The original drug response records in TCGA are categorized according to the Response Evaluation Criteria in Solid Tumors (RECIST): Complete Response (CR), Partial Response (PR), Stable Disease (SD), and Progressive Disease (PD). Following established methodologies in previous studies, we binarized the response data, categorizing CR and PR as positive responses, and SD and PD as negative responses citebomane2019paclitaxel. To standardize drug names and correct for aliases, we employ the reference table developed by [36]. We excluded patients who received drug treatments before tumor resections to avoid potential bias from drug-resistant tumors. Additionally, we removed patients who underwent multiple treatments in a single period to prevent multiple drug response records for individual patients.

We collected and analyzed five distinct types of omics data: microRNA (miRNA), isoform microRNA (isomiR), messenger RNA (mRNA), copy number variation (CNV), and DNA methylation β-values. The experimental strategy, workflow, and additional information on omics data collection are detailed in Supplementary Table 6. For the DNA methylation analysis data, we only select genes with differentially methylated region (DMR) labels. These DMR labels were obtained from humanmethylation450_15017482_v1-2.csv file, available through Illumina's official website. To ensure data quality, we implemented a filtering process wherein genes with more than 40% missing expression values were excluded from the analysis. The remaining missing values were imputed using the mean expression of each gene. All data are downloaded through GDC portal (https://portal.gdc.cancer.gov/) and the command to filter data is discredited in supplementary information table 8.

Table 1. Average model performance rank comparison in all datasets.

Type	Model	MCC ↑	ROC-AUC ↑	PR-AUC ↑
Pre-training	**DOR**	**2.58 ± 1.11**	**2.22 ± 1.54**	**2.02 ± 1.15**
Tree-based	RF	4.26 ± 2.16	4.70 ± 2.22	5.62 ± 2.28
Linear	LR	4.52 ± 2.76	4.84 ± 2.51	5.88 ± 2.58
Linear	EN	4.62 ± 2.49	7.04 ± 2.33	3.12 ± 2.57
Pre-training	VAE	5.04 ± 2.59	4.56 ± 2.14	5.46 ± 2.22
Non-linear	MLP	5.16 ± 2.69	4.46 ± 2.56	5.34 ± 2.54
Pre-training	AE	5.34 ± 2.65	4.52 ± 2.31	5.84 ± 2.16
Tree-based	XGB	7.18 ± 2.32	6.02 ± 2.49	6.78 ± 2.45
Linear	SVM	7.84 ± 2.01	7.90 ± 1.93	8.18 ± 2.00
Tree-based	CART	8.46 ± 2.06	8.74 ± 2.12	6.76 ± 3.05

Note: The rank are presented as mean ± standard deviation.

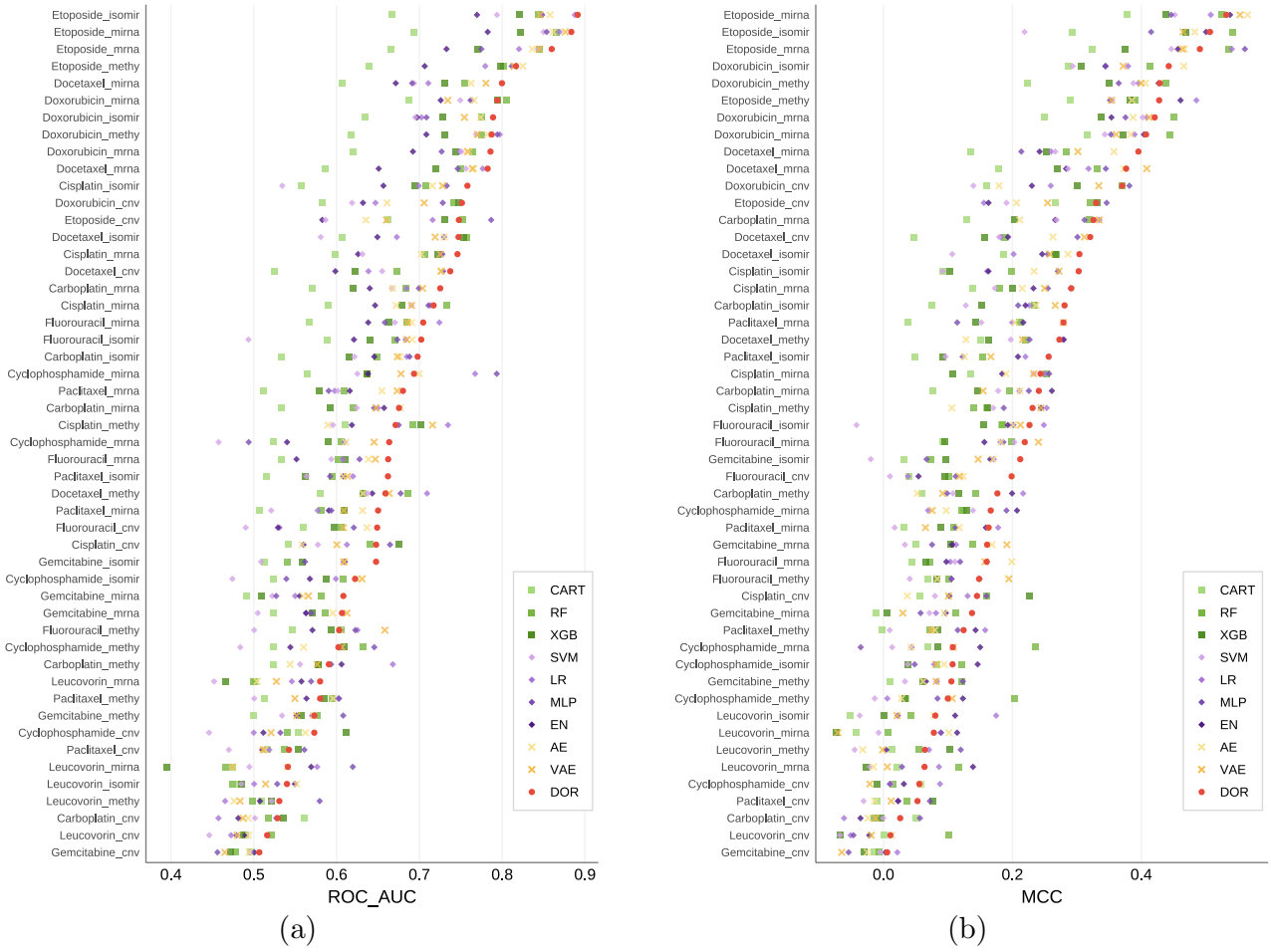


Fig. 2. Model performance comparison. Performance comparison of TCGA drug response prediction as measured by ROC-AUC 2(a); MCC 2(b)

Workflow

In the pre-training phase of DOR model, we collect all labeled and unlabeled samples associated with each drug as pre-training dataset. The pre-training dataset was subsequently partitioned into 90% training data and 10% testing data for the early-stopping proposal. Quantile standardization, implemented via the Scikit-learn library [37], was employed to fit the training data and transform both the training and validation datasets. The DOR model was trained with an early stopping criterion with 5 patience epochs. After pre-training, model parameters and Sklearn normalization object will be saved for the following fine-tuning phase.

For downstream tasks, a classifier was appended to the pre-trained model. During the fine-tuning phase, we conducted a 5-fold nested cross-validation test. Following the previous drug response prediction training pipeline [38], the model was trained on the inner fold's training dataset and validated

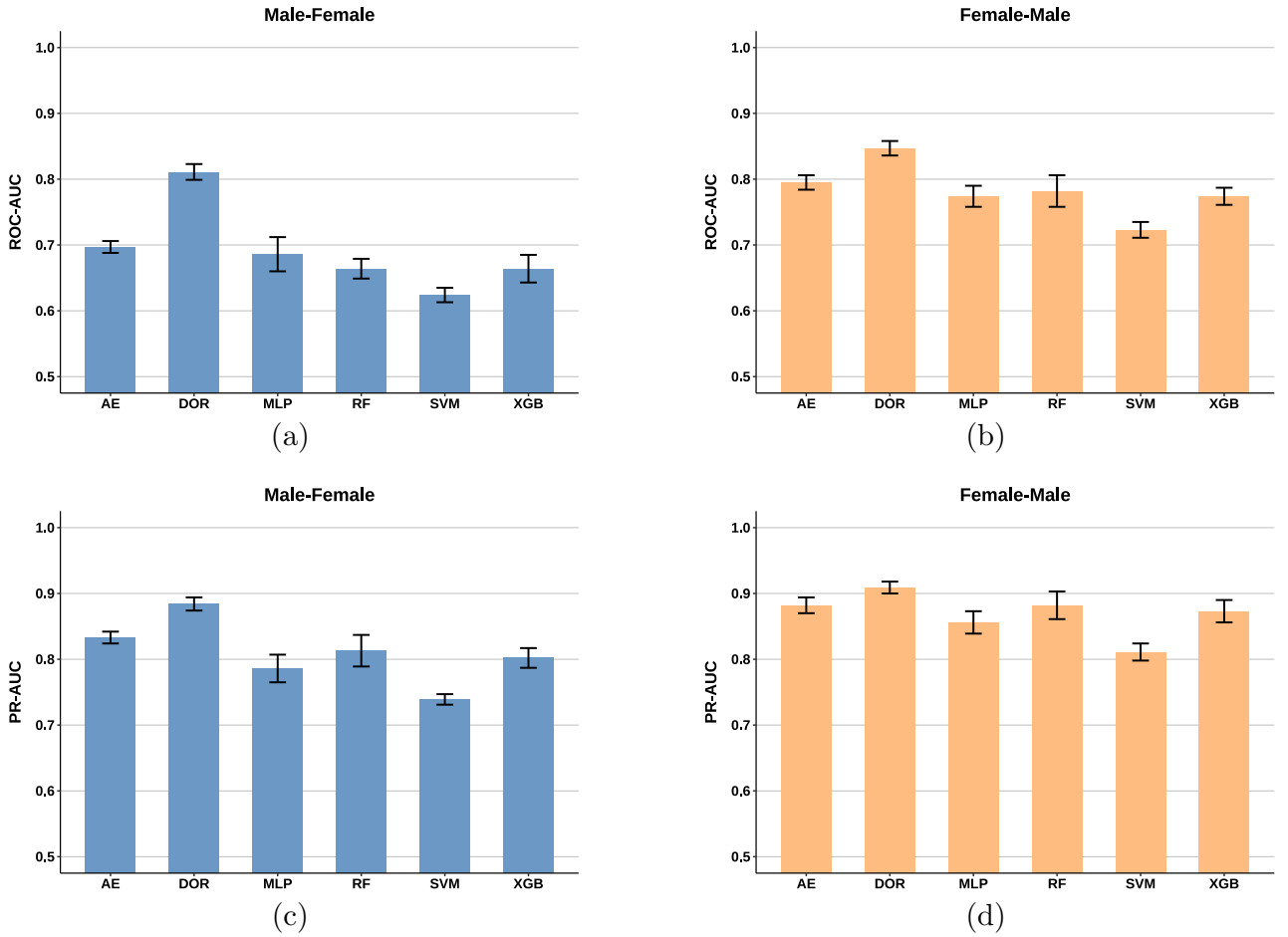


Fig. 3. Performance comparison of various models for cancer subtype prediction, with sex as a confounding factor. The plots illustrate the model performance when trained on male samples and evaluated on female samples (Male-Female) in terms of ROC-AUC (a) and PR-AUC (c). Similarly, model performance when trained on female samples and evaluated on male samples (Female-Male) is shown in terms of ROC-AUC (b) and PR-AUC (d). Error bars represent the standard deviation from cross-validation.

on the corresponding validation dataset to determine the optimal stopping point. The model was then evaluated on the independent test dataset from the outer fold, generating predictions. After predictions from the five inner models were recorded, the final prediction for the outer test data was computed as the mean of these five inner predictions. The ultimate performance metric for nested cross-validation was calculated as the average performance across the five outer folds.

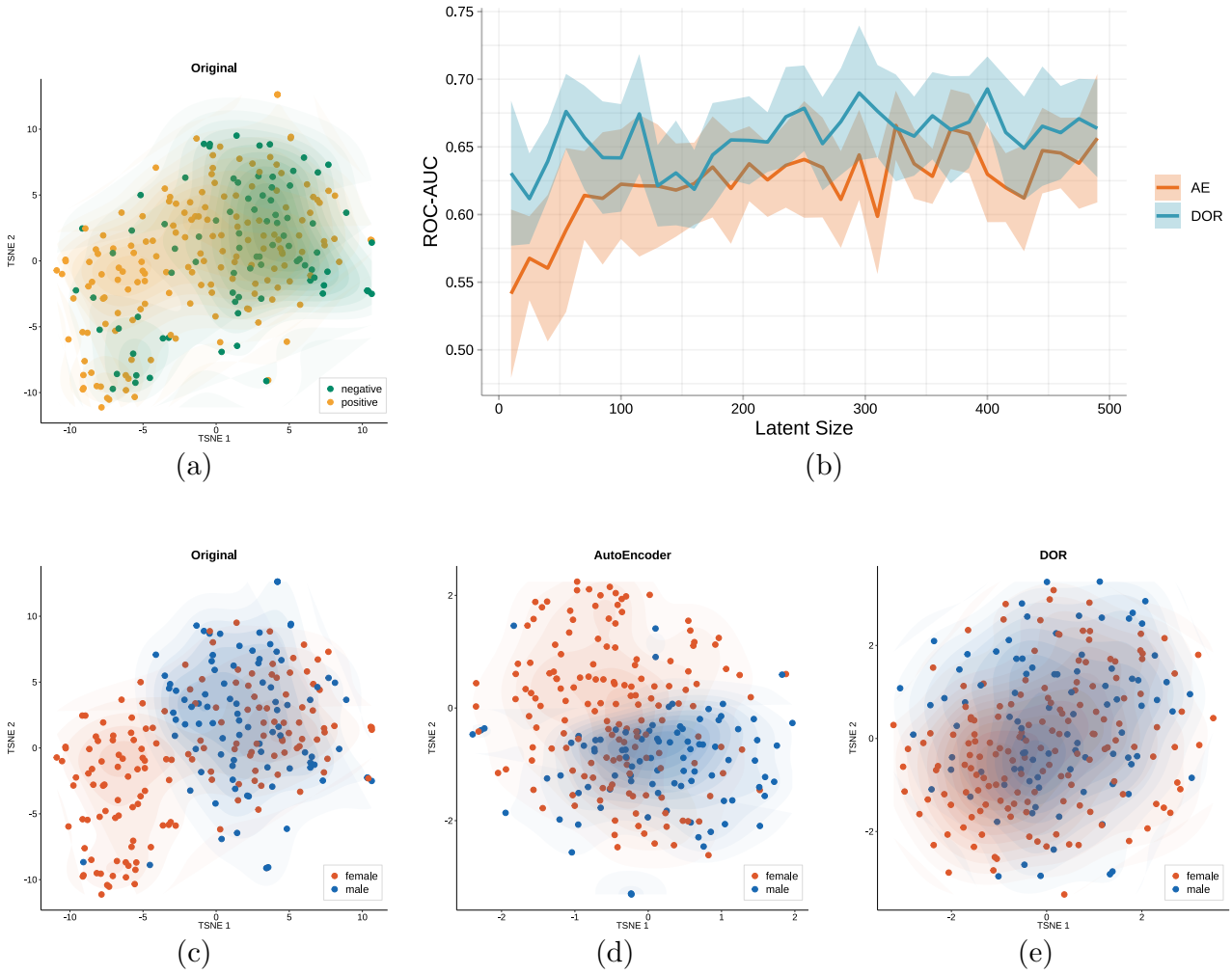
Given the inherent class imbalance in drug response data, we employed a comprehensive set of classification metrics. These included the Area Under the Receiver Operating Characteristic Curve (ROC-AUC), Matthews Correlation Coefficient (MCC), and Precision-Recall AUC (PR-AUC). ROC-AUC provides an assessment of overall model performance, while MCC evaluates the model's capability to handle imbalanced data. Additionally, we utilized the PR-AUC, which, similar to MCC, is particularly suited for imbalanced data scenarios. It is important to note that the classification threshold was optimized using the validation dataset to ensure an unbiased evaluation of the test set.

For hyperparameter optimization, we employed a Bayesian search technique. For each omics data type, we search optimized neural network hyper-parameters configuration on platform Weights & Biases (WandB), including latent size, number of layers, and learning rate. For tree-based models, we utilized the 'BayesSearchCV' class in Scikit-learn to optimize the model parameters during cross-validation. The search space of tuning can be found in supplementary information table 3. To ensure reproducibility, all experiments were replicated with ten different random seeds: [1013, 1481, 3054, 7967, 3658, 3783, 2843, 6384, 666, 8920].

Baseline

We conducted a comparative analysis of the Drug-Omics Response (DOR) model against three distinct categories of baseline models: machine learning-based, tree-based, and pre-training-based approaches. The machine learning category contains Multilayer Perceptron (MLP), Support Vector Machine (SVM), Logistic Regression (LR), and Elastic Net (EN) models. Tree-based models included Random Forest (RF), XGBoost (XGB), and Classification and Regression Trees (CART). Pre-training-based models contain AutoEncoder and Variational AutoEncoder (VAE). It is noteworthy that certain contemporary denoising models for drug response prediction, specifically CODE-AE and ADAE, were not included in this

Fig. 4. Decounfounding result of DOR. 5(a) show the tSNE visualization with response label. 5(c)-5(e) show the tSNE visualization with sex label of Original data 5(c), AutoEncoder embedding 5(d), and DOR embedding 5(e). 5(b) show the variation of ROC-AUC associated with latent size.



study. This exclusion was due to their dependency on cell-line data for training, which falls outside the scope of our research focus on patient-derived data.

To ensure fair comparisons, all neural networks in this study are set with identical configuration for encoders and decoders. The detailed model architecture specifications can be found in Supplementary Information table 4. Each neural network employs a single-layer MLP classifier. We have selected the Sigmoid Linear Unit (SiLU) [39] as the activation function due to its similar behavior to the ReLU function, but with improved stability near the zero point. Furthermore, the weight λ of the Frobenius loss term is set to 1. The learning rate is fixed at 1×10^{-4}, and the batch size is maintained at 32 across all experiments.

Result and Discussion

DOR Enhances Drug Response Predictions in TCGA Cancer Datasets

Our preliminary investigation assessed the effectiveness of the DOR model in predicting drug responses. We compared its performance to other competitive models using clinical data from The Cancer Genome Atlas (TCGA) database. Table 1 illustrates that the DOR model consistently outperformed other approaches across all evaluated metrics, achieving the highest average performance rank. This observation suggests that DOR's ability to disentangle omics features significantly enhances its predictive power. Notably, tree-based models, particularly Random Forest, also demonstrated robust performance, aligning with findings from previous studies [40]. The efficacy of these models can be attributed to their inherent ability to mitigate the influence of non-informative features [41]. Linear models, specifically Elastic Net (EN), displayed remarkable PR-AUC performance in a part of dataset. The relatively simple structure of these models, which involves fewer parameters, may lower the risk of overfitting, making them survive in highly noisy omics dataset. Pre-training models, while generally effective, exhibited vulnerability to overfitting when applied to omics datasets. This limitation likely stems from their large parameter space and the inherent complexity of high-dimensional omics data. Notably, DOR, although using pre-training technique for produce embedding, overcomes the aforementioned limitation. It achieves this by learning disentangled representations that effectively separate informative features from noise, thus enhancing its ability to

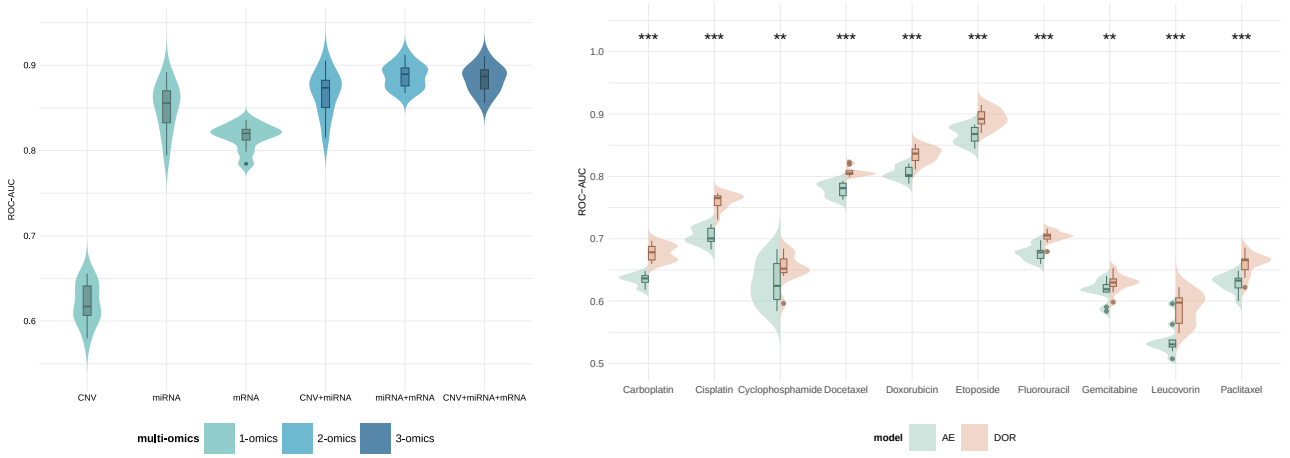


Fig. 5. Multi-omics performance. 5(f) show the performance when train model with different number of omics types. 5(g) show the performance comparison of AutoEncoder and DOR with multi-omics training in 10 dataset

generalize across diverse datasets containing confounders. This unique approach distinguishes DOR from other pre-training models in the omics field.

Figure 2 provides a comprehensive comparison of model performances across 50 datasets. The analysis demonstrates that certain models exhibit superior performance exclusively within specific contextual parameters. For instance, Elastic Net achieved the highest MCC on the Leucovorin-mRNA and Cyclophosphamide-isomiR datasets, but its performance was suboptimal on others. In contrast, DOR demonstrated remarkable consistency across diverse datasets. Its performance was particularly noteworthy for the chemotherapeutic agent Etoposide, where it achieved an ROC-AUC exceeding 0.8. The detailed results of all experiments, including the median values of ROC-AUC and MCC, are provided in the Supplementary Information. Those result can be found in Supplementary Information table 9 - 18.

This comparative analysis highlights the importance of model selection in drug response prediction and underscores the potential advantages of disentangled representation learning in handling complex, high-dimensional omics data. However, it is crucial to acknowledge that model performance may vary depending on the specific characteristics of the dataset and the drug under investigation.

Deconfounding Biological Variables Using DOR

We further examined the capacity of DOR model to deconfound biological variables. A robust representation should exhibit generalization ability across diverse domains. To evaluate this, we followed the experimental protocols established by ADAE and CODE-AE. Our approach involved pre-training models on a comprehensive dataset, using gender as a biological confounder. We then fine-tuned and tested these models on gender-specific subsets.

For this study, we selected the Etoposide-miRNA dataset due to its sufficient sample size and significant domain shift, making it particularly well-suited for evaluating domain generalization capabilities. Figure 3 illustrates the outcomes of the domain adaptation task in both Male-to-Female and Female-to-Male directions. We compared six models: pre-training models (AutoEncoder), tree-based models (Random Forest and XGB), linear models (MLP and SVM), and our disentangled learning approach, DOR.

DOR demonstrated superior performance in both directions, as quantified by ROC-AUC and PR-AUC metrics. The dataset exhibits a notable gender imbalance, with 154 female and 91 male samples. This disparity introduces a significant challenge for domain adaptation, particularly when transferring knowledge from the male subset to the female cohort. While the ROC-AUC for DOR decreased by 0.073 in the Male-to-Female direction compared to Femal-to-Male, other models showed statistically significant declines of at least 0.1 ($p \leq 0.01$). This outcome suggests that DOR effectively learns crucial cancer-related information from omics data, largely unaffected by gender bias. The AutoEncoder model demonstrated enhanced effectiveness in this transfer task when compared to conventional prediction approaches. This improvement can likely be attributed to the model's ability to derive embeddings from the complete dataset. However, it still performed worse than DOR, possibly because it tends to capture gender-specific biases during pre-training.

To further analyze these results, we employed t-distributed stochastic neighbor embedding (t-SNE) to visualize the embedding distributions of the Etoposide-miRNA dataset, as shown in Figure 4. When comparing the original data distribution labeled by response and gender, we observed a notable cluster of female samples in the lower-left quadrant of the gender plot. However, in the response label plot, this cluster includes both positive and negative samples, indicating a shift in labels between gender and response. While the AutoEncoder method was unsuccessful in aligning the female and male domains (as shown in Figure 5(d)), the DOR approach demonstrated superior performance. Figure 5(e) shows that the DOR method successfully accomplished this alignment, as demonstrated by the well-integrated distribution of female and male samples.

We hypothesize that the observed differences in performance between the AutoEncoder and DOR models can be attributed to their distinct strategies for information retention. While the AutoEncoder attempts to capture all aspects of the

Table 2. Performance metrics for various model configurations.

Model Configuration	ROC-AUC	MCC	PR-AUC
DOR	**0.800 ± 0.05**	**0.395 ± 0.07**	**0.901 ± 0.02**
w/o frobenius loss	0.761 ± 0.03	0.349 ± 0.05	0.874 ± 0.02
encoder embedding	0.794 ± 0.04	0.371 ± 0.06	0.898 ± 0.03
concatenation	0.789 ± 0.06	0.385 ± 0.08	0.892 ± 0.03
whole data pre-train	0.783 ± 0.08	0.362 ± 0.05	0.889 ± 0.02

Note: The performance metrics are presented as mean ± standard deviation.

original data, including gender-related information, DOR models appear to selectively preserve the most informative features. To test this hypothesis, we conducted an experiment using the complete Etoposide-miRNA dataset, training both AutoEncoder and DOR models across a range of embedding latent sizes and evaluating their performance using ROC-AUC scores. Figure 5(b) illustrates the ROC-AUC performance as a function of latent size. When the latent size exceeds 100, indicating sufficient capacity to adequately represent the data, DOR models marginally outperform AutoEncoders. However, as the latent size decreases below 100, the performance of the AutoEncoder declines more rapidly, while DOR's performance diminishes at a slower rate. These findings suggest that AutoEncoders, in their attempt to reconstruct all input features, may allocate limited latent space to noise or less relevant information. In comparison, DOR models demonstrate a greater capability for distinguishing between informative features and irrelevant data, effectively prioritizing the retention of essential information under the constraints of limited latent space. This selective preservation of relevant features may contributes to the superior performance of DOR models, especially in contexts where latent dimensions are constrained.

Performance of DOR with Multi-omics data

Given that DOR demonstrated great performance in disentangling omics data to obtain high-quality embeddings, we aimed to investigate the potential of combining these embeddings to develop a multi-omics model. Our primary objective was to determine whether this multi-omics approach could improve drug response prediction using data from The Cancer Genome Atlas (TCGA). For this study, we specifically selected three types of omics data: copy number variation (CNV), messenger RNA (mRNA), and microRNA (miRNA).

As illustrated in Figure 5(f), when employing single-omics models in isolation, miRNA yielded the most promising results, while CNV demonstrated comparatively lower efficacy. However, the integration of multiple omics types significantly improved ROC-AUC. That strongly supports the advantages of multi-omics integration in this context.

Subsequently, we conducted a comparative analysis of DOR's performance against that of AutoEncoder in the integration of multi-omics data (specifically, miRNA, isomiR, mRNA, CNV, and methylation) across a diverse set of 10 drug datasets. The results, as depicted in Figure 5(g), reveal that DOR significantly outperformed AutoEncoder in nearly all datasets. This finding shows the DOR's superior ability in multi-omics integration.

Ablation Study

To optimize the Disentangled Omics Representation (DOR) model, we conducted a comprehensive ablation study utilizing the Doxorubicin-miRNA dataset to assess various configurations. As table **??** shows, the implementation of Frobenius loss substantially improved the model's performance by enhancing its capacity to extract diverse features from the omics data. Other modifications, such as the utilization of encoder embeddings and concatenation operations when combining salient and mutual embeddings, led to slight decreases in performance. These findings align with findings from previous studies [32], which we attribute to the encoder's inherent capacity to achieve some degree of disentanglement during the pre-training phase.

Notably, training DOR on the entire database ($\geq$ 10,000 samples) resulted in a surprising decline in performance. Upon analysis of the embeddings, we observed that the salient embeddings exhibited behavior similar to those generated by an AutoEncoder, trying to reconstruct input data preciously. We hypothesize that this phenomenon may be attributed to the heterogeneous signals within the expansive dataset, causing DOR to capture only a limited amount of relevant information and subsequently leading to overfitting. This observation highlights the importance of carefully considering dataset size and composition in model training.

Discussion of Future Work

In recent years, large-scale models have demonstrated exceptional performance across various domains, particularly in natural language processing [42] and computer vision [17]. Motivated by these advancements, we would like to extend the DOR model by including transformer architectures for omics studies. This approach treats genes as tokens and utilizes gene expression prediction as a pre-training objective, which is similar to token prediction in natural language processing. Previous research has illustrated the potential opportunity of applying BERT and GPT structures to biological tasks [43]. These studies highlight the potential of leveraging knowledge transfer from large-scale, unlabeled datasets to address specific tasks, which is particularly important in advancing precision oncology.

A major challenge in this approach is the process of converting scalar expression data into vector tokens that are compatible with transformer models. While various vectorization methods have been proposed [44, 45, 46], we chose to apply the modern approach introduced by [47], which employs periodic functions for robust embedding. Moreover, due to the typically large feature sizes found in many omics datasets, choosing an appropriate transformer backbone is essential. We intend to use Performer model [48], which has demonstrated efficacy in handling long sequences.

Although training the DOR model on large-scale datasets is not recommended for obtaining high-quality salient embeddings, it is possible to train the encoder component independently on such datasets. We hypothesize that the incorporation of a transformer architecture and large-scale pre-training will enhance the encoder's knowledge representation capabilities. We evaluated our proposed approach on multiple datasets. However, the performance was sometimes very low or even no better than random chance (ROC-AUC=0.5). We hypothesize that this issue is related to hyper-parameter settings, prompting us to conduct an extensive search across a broad range of hyper-parameters. By doing so, we achieved performance comparable to the standard version of DOR. Despite utilizing multi-GPU acceleration, the hyper-parameter search process was exceedingly time-consuming due to the large size of both the transformer model and the pre-training dataset. Consequently, it was impossible to apply this search process to each type of omics data. Future work should focus on developing more efficient transformers and conducting comprehensive experiments with the transformer-based version of DOR.

Conclusions

In this study, we introduced DOR, a disentangled omics representation model for drug response prediction. By disentangling salient and mutual information in omics data during pre-training, DOR learns robust embeddings that capture the most informative features for predicting drug response. Our experiments on the TCGA database demonstrated that DOR consistently outperforms other machine learning models across various metrics. Furthermore, DOR's ability to deconfound biological variables like gender enables it to generalize well across different patient subgroups. Visualizations of the learned embeddings confirmed DOR's effectiveness in aligning different domains. We also show the potential application of DOR in multi-omics integration, where combining the disentangled embeddings from different omics types led to improved predictive performance. An ablation study identified the optimal configuration for DOR, highlighting the importance of the Frobenius loss in promoting diversity in the learned features. Overall, DOR represents a novel approach to learning robust representations from omics data for drug response prediction. By disentangling the informative signal from confounding noise, DOR may bring us one step closer to the goal of expression-based precision oncology.

References

1. Pedro J Ballester and Javier Carmona. Artificial intelligence for the next generation of precision oncology. *NPJ Precision Oncology*, 5(1):79, 2021.
2. Delora Baptista, Pedro G Ferreira, and Miguel Rocha. Deep learning for drug response prediction in cancer. *Briefings in bioinformatics*, 22(1):360–379, 2021.
3. Francisco Azuaje. Computational models for predicting drug responses in cancer research. *Briefings in bioinformatics*, 18(5):820–829, 2017.
4. Farzaneh Firoozbakht, Behnam Yousefi, and Benno Schwikowski. An overview of machine learning methods for monotherapy drug response prediction. *Briefings in bioinformatics*, 23(1):bbab408, 2022.
5. George Adam, Ladislav Rampášek, Zhaleh Safikhani, Petr Smirnov, Benjamin Haibe-Kains, and Anna Goldenberg. Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ precision oncology*, 4(1):19, 2020.
6. James C Costello, Laura M Heiser, Elisabeth Georgii, Mehmet Gönen, Michael P Menden, Nicholas J Wang, Mukesh Bansal, Muhammad Ammad-Ud-Din, Petteri Hintsanen, Suleiman A Khan, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology*, 32(12):1202–1212, 2014.
7. Michael P Menden, Dennis Wang, Mike J Mason, Bence Szalai, Krishna C Bulusu, Yuanfang Guan, Thomas Yu, Jaewoo Kang, Minji Jeon, Russ Wolfinger, et al. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nature communications*, 10(1):2674, 2019.
8. Madison Butler, Lorinc Pongor, Yu-Ting Su, Liqiang Xi, Mark Raffeld, Martha Quezado, Jane Trepel, Kenneth Aldape, Yves Pommier, and Jing Wu. Mgmt status as a clinical biomarker in glioblastoma. *Trends in cancer*, 6(5):380–391, 2020.
9. Smriti Chawla, Anja Rockstroh, Melanie Lehman, Ellca Ratther, Atishay Jain, Anuneet Anand, Apoorva Gupta, Namrata Bhattacharya, Sarita Poonia, Priyadarshini Rai, et al. Gene expression based inference of cancer drug sensitivity. *Nature communications*, 13(1):5680, 2022.
10. Ibiayi Dagogo-Jack and Alice T Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nature reviews Clinical oncology*, 15(2):81–94, 2018.
11. Antonio Passaro, Maise Al Bakir, Emily G Hamilton, Maximilian Diehn, Fabrice André, Sinchita Roy-Chowdhuri, Giannis Mountzios, Ignacio I Wistuba, Charles Swanton, and Solange Peters. Cancer biomarkers: Emerging trends and clinical implications for personalized treatment. *Cell*, 187(7):1617–1635, 2024.
12. Adeolu Ogunleye, Chayanit Piyawajanusorn, Ghita Ghislat, and Pedro J Ballester. Large-scale machine learning analysis reveals dna methylation and gene expression response signatures for gemcitabine-treated pancreatic cancer. *Health Data Science*, 4:0108, 2024.
13. Hossein Sharifi-Noghabi, Soheil Jahangiri-Tazehkand, Petr Smirnov, Casey Hon, Anthony Mammoliti, Sisira Kadambat Nair, Arvind Singh Mer, Martin Ester, and Benjamin Haibe-Kains. Drug sensitivity prediction from cell line-based pharmacogenomics data: guidelines for developing machine learning models. *Briefings in Bioinformatics*, 22(6):bbab294, 2021.
14. Parminder S Reel, Smarti Reel, Ewan Pearson, Emanuele Trucco, and Emily Jefferson. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology advances*, 49:107739, 2021.
15. Stefan Naulaerts, Cuong C Dang, and Pedro J Ballester. Precision and recall oncology: combining multiple gene mutations for improved identification of drug-sensitive tumours. *Oncotarget*, 8(57):97025, 2017.
16. Jianzhu Ma, Samson H Fong, Yunan Luo, Christopher J Bakkenist, John Paul Shen, Soufiane Mourragui, Lodewyk FA Wessels, Marc Hafner, Roded Sharan, Jian Peng, et al. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nature Cancer*, 2(2):233–244, 2021.

17. Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, pages 1–11, 2024.
18. Xiaoyu Zhang, Jingqing Zhang, Kai Sun, Xian Yang, Chengliang Dai, and Yike Guo. Integrated multi-omics analysis using variational autoencoders: application to pan-cancer classification. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 765–769. IEEE, 2019.
19. Joël Simoneau, Simon Dumontier, Ryan Gosselin, and Michelle S Scott. Current rna-seq methodology reporting limits reproducibility. *Briefings in bioinformatics*, 22(1):140–145, 2021.
20. Jonathan M Raser and Erin K O'shea. Noise in gene expression: origins, consequences, and control. *Science*, 309(5743):2010–2013, 2005.
21. Luca Parca, Gerardo Pepe, Marco Pietrosanto, Giulio Galvan, Leonardo Galli, Antonio Palmeri, Marco Sciandrone, Fabrizio Ferrè, Gabriele Ausiello, and Manuela Helmer-Citterich. Modeling cancer drug response through drug-specific informative genes. *Scientific reports*, 9(1):15222, 2019.
22. Shuoran Jiang, Qingcai Chen, Yang Xiang, Youcheng Pan, Xiangping Wu, and Yukang Lin. Confounder balancing in adversarial domain adaptation for pre-trained large models fine-tuning. *Neural Networks*, 173:106173, 2024.
23. Jingshu Wang, Divyansh Agarwal, Mo Huang, Gang Hu, Zilu Zhou, Chengzhong Ye, and Nancy R Zhang. Data denoising with transfer learning in single-cell transcriptomics. *Nature methods*, 16(9):875–878, 2019.
24. Jiali Pang, Bilin Liang, Ruifeng Ding, Qiujuan Yan, Ruiyao Chen, and Jie Xu. A denoised multi-omics integration framework for cancer subtype classification and survival prediction. *Briefings in Bioinformatics*, 24(5):bbad304, 2023.
25. Hossein Sharifi-Noghabi, Parsa Alamzadeh Harjandi, Olga Zolotareva, Colin C Collins, and Martin Ester. Out-of-distribution generalization from labelled and unlabelled gene expression data for drug response prediction. *Nature Machine Intelligence*, 3(11):962–972, 2021.
26. Ayse B Dincer, Joseph D Janizek, and Su-In Lee. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics*, 36(Supplement_2):i573–i582, 2020.
27. Peilin Jia, Ruifeng Hu, Guangsheng Pei, Yulin Dai, Yin-Ying Wang, and Zhongming Zhao. Deep generative neural network for accurate drug response imputation. *Nature communications*, 12(1):1740, 2021.
28. Di He, Qiao Liu, You Wu, and Lei Xie. A context-aware deconfounding autoencoder for robust prediction of personalized clinical drug response from cell-line compound screening. *Nature Machine Intelligence*, 4(10):879–892, 2022.
29. Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *Advances in neural information processing systems*, 29, 2016.
30. Abubakar Abid, Martin J Zhang, Vivek K Bagaria, and James Zou. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature communications*, 9(1):2134, 2018.
31. Abubakar Abid and James Zou. Contrastive variational autoencoder enhances salient features. *arXiv preprint arXiv:1902.04601*, 2019.
32. Jing Wu, Suiyao Chen, Qi Zhao, Renat Sergazinov, Chen Li, Shengjie Liu, Chongchao Zhao, Tianpei Xie, Hanqing Guo, Cheng Ji, et al. Switchtab: Switched autoencoders are effective tabular learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15924–15933, 2024.
33. Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems*, 33:7198–7211, 2020.
34. Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
35. John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.
36. John Christian Givhan Spainhour and Peng Qiu. Identification of gene-drug interactions that impact patient survival in tcga. *BMC bioinformatics*, 17:1–8, 2016.
37. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
38. Alexandra Bomane, Anthony Gonçalves, and Pedro J Ballester. Paclitaxel response can be predicted with interpretable multi-variate classifiers exploiting dna-methylation and mirna data. *Frontiers in genetics*, 10:1041, 2019.
39. Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018.
40. Adeolu Z Ogunleye, Chayanit Piyawajanusorn, Anthony Gonçalves, Ghita Ghislat, and Pedro J Ballester. Interpretable machine learning models to predict the resistance of breast cancer patients to doxorubicin from their microrna profiles. *Advanced Science*, 9(24):2201501, 2022.
41. Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520, 2022.
42. Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022.
43. Yuting He, Fuxiang Huang, Xinrui Jiang, Yuxiang Nie, Minghao Wang, Jiguang Wang, and Hao Chen. Foundation model for advancing healthcare: Challenges, opportunities, and future directions. *arXiv preprint arXiv:2404.03264*, 2024.
44. Talip Ucar, Ehsan Hajiramezanali, and Lindsay Edwards. Subtab: Subsetting features of tabular data for self-supervised representation learning. *Advances in Neural Information Processing Systems*, 34:18853–18865, 2021.

45. Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6679–6687, 2021.
46. Ehsan Hajiramezanali, Nathaniel Lee Diamant, Gabriele Scalia, and Max W Shen. Stab: Self-supervised learning for tabular data. In *NeurIPS 2022 First Table Representation Workshop*, 2022.
47. Yury Gorishniy, Ivan Rubachev, and Artem Babenko. On embeddings for numerical features in tabular deep learning. *Advances in Neural Information Processing Systems*, 35:24991–25004, 2022.
48. Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Pedro Ballester, for his invaluable guidance, unwavering support, and insightful feedback throughout my MRes journey. His expertise and mentorship have been instrumental in shaping my research and helping me navigate the challenges of this endeavour. I am truly grateful for the opportunity to have worked under his supervision.

I would also like to extend my sincere thanks to PhD students in Dr. Pedro's group, Hanqin Du, Krinos Li, Qianrong Guo and Chayanit Piyawajanusorn, for their constant support and collaboration. Their willingness to share ideas, provide constructive feedback, and lend a helping hand in daily work has been a source of motivation and inspiration. I am fortunate to have had such wonderful colleagues who have made this journey more enriching and enjoyable.

I would like to thank the Department of Bioengineering, Imperial College for providing me with the resources and opportunities necessary to pursue my research. This work would not have been possible without the support of this institution.

Finally, we acknowledge computational resources and support provided by the Imperial College Research Computing Service (http://doi.org/10.14469/hpc/2232)

Supplementary information

Omics data summary

Table 6 show the number and label ratio for dataset. 8 show how to filter data in GDC portal (https://portal.gdc.cancer.gov/) to download omics data. Tabel 3 show the hyper-parameter searching space of tree-based model. Table 4 show the hyper-parameter setting of AutoEncoder and Variational AutoEncoder. Table 5 show the hyper-parameter setting of DOR. Table 9 - Table 18 shows the value of ROC-AUC and MCC for each model on 10 dataset.

Table 3. Summary of Hyperparameters for Tree Models

Model	Hyperparameter	Range/Values
CART	criterion	gini, entropy, log_loss
	max_depth	2 to round(nfeature$^{0.25}$)
	min_samples_split	2 to 10
RF	max_depth	2 to round(log(nfeature))
	min_samples_leaf	1 to 5
	criterion	gini, entropy, log_loss
	max_features	0.01 to 1.0
XGB	max_depth	2 to round(nfeature$^{0.25}$)
	max_leaves	3 to round(nfeature$^{0.25}$ × 2)
	grow_policy	depthwise, lossguide
	base_score	0.1 to 0.9
	reg_alpha	0.0 to 3.0
	reg_lambda	0.0 to 3.0
LGBM	num_leaves	3 to round(nfeature$^{0.25}$ × 2)
	max_depth	2 to round(nfeature$^{0.25}$)
	reg_lambda	0.0 to 3.0
	min_child_samples	5 to 40
	min_child_weight	0.0001 to 0.01
	reg_alpha	0.0 to 3.0

Table 4. Summary of Hyperparameters for AutoEncoder and Variational AutoEncoder

Parameter	miRNA	isomiR	mRNA	Methylation	CNV
latent_size	500	1200	500	5000	1000
encoder_hidden_dim	500	1200	500	5000	1000
encoder_num_layers	1	1	2	2	4
predictor_hidden_dim	500	1200	500	5000	1000
predictor_num_layers	1	1	1	1	1

Table 5. Summary of Hyperparameters Used by Model DOR

Parameter	miRNA	isomiR	mRNA	Methylation	CNV
latent_size	500	1200	500	5000	1000
encoder_hidden_dim	500	1200	500	5000	1000
encoder_num_layers	1	1	2	2	4
predictor_hidden_dim	500	1200	500	5000	1000
predictor_num_layers	1	1	1	1	1
salient_hidden_dim	500	1200	500	N/A	N/A
salient_num_layers	1	1	1	N/A	N/A
mutual_hidden_dim	500	1200	500	N/A	N/A
mutual_num_layers	1	1	1	N/A	N/A
decoder_hidden_dim	500	N/A	N/A	N/A	N/A
decoder_num_layers	0	N/A	N/A	N/A	N/A

Table 6. Sample distribution across different molecular data types and drugs.

Drug	miRNA	mRNA	isomiR	CNV	Methylation
Carboplatin	245 (160/85)	245 (160/85)	245 (160/85)	237 (157/80)	239 (154/85)
Cisplatin	421 (325/96)	416 (321/95)	421 (325/96)	414 (321/93)	413 (318/95)
Cyclophosphamide	166 (153/13)	170 (157/13)	166 (153/13)	168 (156/12)	131 (119/12)
Docetaxel	156 (105/51)	160 (108/52)	156 (105/51)	153 (105/48)	139 (88/51)
Doxorubicin	162 (118/44)	163 (119/44)	162 (118/44)	160 (118/42)	146 (103/43)
Etoposide	107 (87/20)	106 (86/20)	107 (87/20)	106 (87/19)	106 (86/20)
Fluorouracil	249 (180/69)	255 (186/69)	249 (180/69)	254 (186/68)	193 (133/60)
Gemcitabine	219 (102/117)	220 (102/118)	219 (102/117)	213 (99/114)	224 (101/123)
Leucovorin	107 (77/30)	109 (79/30)	107 (77/30)	109 (79/30)	69 (47/22)
Paclitaxel	239 (167/72)	237 (166/71)	239 (167/72)	232 (162/70)	230 (158/72)

Table 7. Dimension of omics data

	miRNA	mRNA	isomiR	CNV	Methylation
Dimension	1770	19564	1636	34708	32840

Table 8. Dataset Filters

Dataset	**Filter**
clinical record	cases.project.program.name in [”TCGA”] and cases.project.project_id in [”TCGA-LGG”] and files.data_format in [”bcr biotab”]
mRNA	cases.project.project_id in [”TCGA-LGG”] and files.analysis.workflow_type in [”STAR - Counts”] and files.data_type in [”Gene Expression Quantification”]
miRNA	cases.project.project_id in [”TCGA-LGG”] and files.data_type in [”miRNA Expression Quantification”] and files.experimental_strategy in [”miRNA-Seq”]
isomiR	cases.project.project_id in [”TCGA-LGG”] and files.data_type in [”Isoform Expression Quantification”] and files.experimental_strategy in [”miRNA-Seq”]
methylation β value	cases.project.project_id in [”TCGA-LGG”] and files.experimental_strategy in [”Methylation Array”] and files.data_type in [”Methylation Beta Value”]
CNV	cases.project.project_id in [”TCGA-LGG”] and files.analysis.workflow_type in [”ASCAT2”] and files.data_type in [”Gene Level Copy Number”]

Table 9. MCC of mRNA omics for different drugs and methods

Drug	**CART**	**RF**	**XGB**	**LR**	**EN**	**MLP**	**AE**	**DOR**	**SVM**
Carboplatin	0.1259	0.3196	0.2062	0.3320	0.2665	0.2987	0.2143	0.3008	0.2636
Cisplatin	0.1280	0.1847	0.1844	0.2852	0.2327	0.2309	0.2127	0.2664	0.1647
Cyclophosphamide	0.0850	0.2467	0.1518	0.1047	0.1500	-0.0825	0.0500	0.0826	0.0130
Docetaxel	0.1700	0.3035	0.2584	0.3143	0.2833	0.2950	0.3643	0.3515	0.2671
Doxorubicin	0.2537	0.4413	0.3491	0.4049	0.3537	0.3874	0.4196	0.3958	0.4059
Etoposide	0.3229	0.5329	0.4234	0.5316	0.4568	0.5036	0.4532	0.4486	0.4676
Fluorouracil	0.0502	0.0712	0.0628	0.1037	0.0975	0.1262	0.1745	0.1352	0.1029
Gemcitabine	0.0471	0.1387	0.1031	0.1130	0.1064	0.1039	0.1378	0.1358	0.0404
Leucovorin	-0.0001	0.0819	-0.0283	0.1053	0.1388	-0.0079	-0.0281	0.0275	-0.0423
Paclitaxel	0.0303	0.2001	0.1304	0.1933	0.2169	0.1062	0.2171	0.2542	0.1494

Table 10. MCC of miRNA omics for different drugs and methods

Drug	**CART**	**RF**	**XGB**	**LR**	**EN**	**MLP**	**AE**	**DOR**	**SVM**
Carboplatin	0.0731	0.1834	0.1313	0.2111	0.2613	0.1776	0.2239	0.2160	0.1723
Cisplatin	0.1241	0.2241	0.1096	0.2445	0.1841	0.2542	0.1805	0.2187	0.2305
Cyclophosphamide	0.1455	0.2066	0.2009	0.1019	0.2072	0.2427	0.0956	0.0641	0.1285
Docetaxel	0.1204	0.2840	0.2476	0.2646	0.2422	0.2349	0.3105	0.3705	0.2840
Doxorubicin	0.3159	0.4451	0.3707	0.3959	0.4041	0.3549	0.3788	0.3825	0.3634
Etoposide	0.3718	0.4988	0.4269	0.4982	0.5377	0.4407	0.5615	0.5071	0.4516
Fluorouracil	0.0799	0.1795	0.0981	0.1842	0.1566	0.1743	0.2163	0.1942	0.1638
Gemcitabine	-0.0102	0.1143	0.0102	0.0733	0.0967	0.0377	0.0529	0.1124	0.0594
Leucovorin	-0.0320	0.0002	-0.0462	0.0971	0.1143	0.1049	0.0362	0.0531	0.0200
Paclitaxel	0.0334	0.1159	0.0790	0.1701	0.1588	0.1300	0.1393	0.1382	-0.0020

Table 11. MCC of methylation omics for different drugs and methods

Drug	CART	RF	XGB	LR	EN	MLP	AE	DOR	SVM
Carboplatin	0.0473	0.1308	0.1123	0.2146	0.1995	0.0809	0.0468	-0.0085	0.0887
Cisplatin	0.1577	0.1307	0.1665	0.2506	0.2444	0.1997	0.1245	–	0.1487
Cyclophosphamide	0.0500	0.2357	0.0445	0.0071	0.1233	0.1106	0.0325	0.0282	-0.0455
Docetaxel	0.1469	0.2298	0.1609	0.2713	0.2798	0.2150	0.1325	0.0472	0.1828
Doxorubicin	0.2170	0.4519	0.3485	0.3949	0.4269	0.3790	0.3775	0.4037	0.3810
Etoposide	0.2825	0.4024	0.4111	0.5174	0.4609	0.3295	0.3704	0.2973	0.3077
Fluorouracil	0.0736	0.0917	0.0695	0.1053	0.1488	0.1167	0.1004	0.0135	-0.0101
Gemcitabine	0.0056	0.1088	0.0584	0.0855	0.0658	0.1338	0.1020	0.0251	0.0538
Leucovorin	0.0145	0.0450	0.0768	0.0738	0.0043	0.1128	-0.0306	-0.0061	0.0622
Paclitaxel	-0.0175	0.0582	0.0807	0.1621	0.1429	0.1313	0.1071	-0.0022	-0.0518

Table 12. MCC of isomiR omics for different drugs and methods

Drug	CART	RF	XGB	LR	EN	MLP	AE	DOR	SVM
Carboplatin	0.0716	0.2374	0.1612	0.2144	0.2203	0.2410	0.2639	0.2559	0.1433
Cisplatin	0.0733	0.1710	0.0921	0.1594	0.1629	0.2609	0.2362	0.2779	0.0891
Cyclophosphamide	0.1013	0.1517	0.0556	0.1147	0.1461	0.0454	0.0750	0.0736	-0.0186
Docetaxel	0.1788	0.2655	0.2613	0.2005	0.2462	0.2773	0.3076	0.2790	0.0988
Doxorubicin	0.2877	0.4269	0.3190	0.3784	0.3445	0.3441	0.4247	0.4176	0.3046
Etoposide	0.2864	0.5478	0.4468	0.4858	0.5005	0.4196	0.5173	0.4814	0.3053
Fluorouracil	0.1578	0.2050	0.1597	0.2539	0.1930	0.1901	0.2250	0.2015	0.0179
Gemcitabine	0.0309	0.0545	0.0902	0.1699	0.1657	0.0834	0.1404	0.1872	0.0218
Leucovorin	-0.0518	0.0508	0.0044	0.1836	0.1112	0.0485	0.0341	0.0554	-0.0470
Paclitaxel	0.0559	0.1456	0.0844	0.2227	0.2087	0.1164	0.1388	0.2313	0.0780

Table 13. MCC of CNV omics for different drugs and methods

Drug	CART	RF	XGB	LR	EN	MLP	AE	DOR	SVM
Carboplatin	0.0250	-0.0247	-0.0131	-0.0758	-0.0357	0.0336	-0.0282	0.0010	0.0114
Cisplatin	0.0444	0.1372	0.2143	0.1235	0.1022	0.1628	0.0611	0.1201	0.0741
Cyclophosphamide	0.0084	-0.0069	-0.0204	0.1177	0.0317	0.0055	0.0216	0.0306	-0.0026
Docetaxel	0.0597	0.1908	0.1234	0.1766	0.1933	0.3288	0.2746	–	0.1775
Doxorubicin	0.1424	0.3478	0.3105	0.2469	0.2302	0.3638	0.1638	0.3453	0.1488
Etoposide	0.2505	0.2892	0.2702	0.1817	0.1630	0.3121	0.2158	0.3049	0.1878
Fluorouracil	0.0361	0.1203	0.0863	0.0335	0.0544	0.1361	0.1539	0.1738	-0.0092
Gemcitabine	0.0060	-0.0088	-0.0335	0.0142	0.0030	-0.0646	-0.0287	-0.0224	-0.0016
Leucovorin	-0.0020	0.1053	-0.0487	-0.0560	-0.0220	-0.0523	-0.0117	-0.0363	-0.0578
Paclitaxel	-0.0286	0.0661	0.0906	0.0191	0.0226	0.0564	-0.0148	-0.0155	-0.0077

Table 14. ROC-AUC of CNV omics for different drugs and methods

Drug	CART	RF	XGB	LR	EN	MLP	AE	DOR	SVM
Carboplatin	0.5611	0.5180	0.5363	0.4573	0.4813	0.5361	0.4749	0.4956	0.5014
Cisplatin	0.5422	0.6401	0.6749	0.6125	0.5600	0.6632	0.5643	0.6238	0.5774
Cyclophosphamide	0.5536	0.5402	0.6119	0.5000	0.5116	0.5281	0.5523	0.5490	0.4460
Docetaxel	0.5249	0.6730	0.6223	0.6385	0.5985	0.7322	0.7263	–	0.6550
Doxorubicin	0.5829	0.7414	0.7475	0.6469	0.6319	0.7243	0.6725	0.7274	0.6186
Etoposide	0.6624	0.7522	0.7311	0.7160	0.5827	0.7800	0.6273	0.7237	0.5867
Fluorouracil	0.5593	0.6051	0.5972	0.5311	0.5292	0.6214	0.6258	0.6250	0.4903
Gemcitabine	0.4963	0.4785	0.4723	0.4960	0.5007	0.4549	0.4797	0.4774	0.4957
Leucovorin	0.4903	0.5217	0.4873	0.4726	0.4899	0.4809	0.4873	0.4921	0.4462
Paclitaxel	0.5170	0.5373	0.5538	0.5187	0.5103	0.5616	0.5040	0.5060	0.4699

Table 15. ROC-AUC of isomiR omics for different drugs and methods

Drug	CART	RF	XGB	LR	EN	MLP	AE	DOR	SVM
Carboplatin	0.5329	0.6488	0.6152	0.6848	0.6455	0.6909	0.6671	0.6738	0.6217
Cisplatin	0.5572	0.7080	0.6947	0.6993	0.6565	0.7296	0.7143	0.7340	0.5343
Cyclophosphamide	0.5240	0.6082	0.5872	0.5991	0.5677	0.5503	0.6083	0.5984	0.4742
Docetaxel	0.6064	0.7577	0.7530	0.6727	0.6493	0.7387	0.7373	0.7232	0.5809
Doxorubicin	0.6339	0.7748	0.7278	0.7027	0.6971	0.7249	0.7801	0.7649	0.6950
Etoposide	0.6665	0.8436	0.8204	0.8877	0.7696	0.8331	0.8587	0.8672	0.7937
Fluorouracil	0.5892	0.6706	0.6405	0.6834	0.6209	0.6818	0.6930	0.6783	0.4936
Gemcitabine	0.5129	0.5403	0.5581	0.6083	0.6099	0.5679	0.6055	0.6237	0.5088
Leucovorin	0.4830	0.4748	0.4854	0.5005	0.5496	0.5360	0.5530	0.5161	0.4847
Paclitaxel	0.5151	0.5951	0.5627	0.6198	0.6081	0.5800	0.6034	0.6380	0.5638

Table 16. ROC-AUC of mRNA omics for different drugs and methods

Drug	CART	RF	XGB	LR	EN	MLP	AE	DOR	SVM
Carboplatin	0.5701	0.6831	0.6200	0.6845	0.6400	0.7067	0.6631	0.7011	0.6669
Cisplatin	0.5985	0.7065	0.7220	0.7265	0.6260	0.7225	0.7040	0.7220	0.6312
Cyclophosphamide	0.5242	0.6051	0.5893	0.6098	0.5402	0.4755	0.6213	0.6396	0.4575
Docetaxel	0.5861	0.7503	0.7203	0.7765	0.6508	0.7433	0.7600	0.7584	0.7562
Doxorubicin	0.6200	0.7647	0.7433	0.7485	0.6922	0.7301	0.7672	0.7618	0.7507
Etoposide	0.6660	0.8455	0.7697	0.8203	0.7328	0.8179	0.8501	0.8358	0.8431
Fluorouracil	0.5334	0.6018	0.6109	0.6078	0.5516	0.6212	0.6294	0.6383	0.5937
Gemcitabine	0.5233	0.5861	0.5702	0.5627	0.5633	0.5845	0.5944	0.5825	0.5048
Leucovorin	0.5016	0.4997	0.4659	0.5460	0.5575	0.5602	0.5072	0.5561	0.4520
Paclitaxel	0.5118	0.6092	0.5791	0.5976	0.6160	0.5790	0.6480	0.6562	0.6016

Table 17. ROC-AUC of miRNA omics for different drugs and methods

Drug	CART	RF	XGB	LR	EN	MLP	AE	DOR	SVM
Carboplatin	0.5334	0.6196	0.5922	0.6499	0.6573	0.6464	0.6814	0.6514	0.6243
Cisplatin	0.5904	0.7325	0.6790	0.7133	0.6465	0.7086	0.6808	0.6932	0.6904
Cyclophosphamide	0.5647	0.6369	0.6365	0.7673	0.6379	0.7940	0.6664	0.6627	0.6251
Docetaxel	0.6070	0.7551	0.7302	0.7107	0.6714	0.6925	0.7594	0.7758	0.6936
Doxorubicin	0.6878	0.8058	0.7944	0.7635	0.7258	0.7655	0.7703	0.7701	0.7493
Etoposide	0.6932	0.8654	0.8219	0.8691	0.7825	0.8574	0.8799	0.8595	0.8500
Fluorouracil	0.5665	0.6845	0.6627	0.7241	0.6382	0.6630	0.6913	0.6806	0.6704
Gemcitabine	0.4908	0.5816	0.5098	0.5499	0.5556	0.5288	0.5584	0.5841	0.5223
Leucovorin	0.4740	0.4661	0.3946	0.5766	0.5693	0.6208	0.5184	0.5171	0.4947
Paclitaxel	0.5074	0.5816	0.6095	0.5772	0.5906	0.5944	0.6275	0.6261	0.5211

Table 18. ROC-AUC of methylation omics for different drugs and methods

Drug	CART	RF	XGB	LR	EN	MLP	AE	DOR	SVM
Carboplatin	0.5231	0.5782	0.5782	0.6679	0.6060	0.5804	0.5477	0.4935	0.5558
Cisplatin	0.6108	0.6923	0.7019	0.7346	0.6185	0.6849	0.6164	–	0.5956
Cyclophosphamide	0.5234	0.6321	0.6089	0.6045	0.5441	0.6522	0.5464	0.5758	0.4837
Docetaxel	0.5797	0.6862	0.6317	0.7093	0.6431	0.6739	0.6293	0.5426	0.6374
Doxorubicin	0.6177	0.7749	0.7304	0.7970	0.7084	0.7839	0.7774	0.7633	0.7686
Etoposide	0.6393	0.8010	0.7974	0.7795	0.7064	0.7985	0.8257	0.7927	0.8167
Fluorouracil	0.5464	0.6056	0.5933	0.6252	0.5708	0.6149	0.5991	0.5119	0.5005
Gemcitabine	0.4992	0.5765	0.5570	0.5703	0.5517	0.6030	0.5817	0.5069	0.5341
Leucovorin	0.5093	0.4986	0.5217	0.4654	0.5074	0.5741	0.4824	0.5062	0.5214
Paclitaxel	0.5125	0.5943	0.5848	0.6021	0.5636	0.5977	0.5915	0.4984	0.5006