Biomedicine & Pharmacotherapy 177 (2024) 117070

Contents lists available at ScienceDirect

Biomedicine & Pharmacotherapy

journal homepage: www.elsevier.com/locate/biopha

scDrug+: predicting drug-responses using single-cell transcriptomics and molecular structure

Yih-Yun Sun [a,b], Chiao-Yu Hsieh [b], Jian-Hung Wen [b,c], Tzu-Yang Tseng [a,d], Jia-Hsin Huang [b], Yen-Jen Oyang [a], Hsuan-Cheng Huang [c,*], Hsueh-Fen Juan [a,b,d,e,f,**]

[a] Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taiwan
[b] Taiwan AI Labs, Taipei 10351, Taiwan
[c] Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, Taipei 11221, Taiwan
[d] Department of Life Science, National Taiwan University, Taipei 106, Taiwan
[e] Center for Computational and Systems Biology, National Taiwan University, Taipei 106, Taiwan
[f] Center for Advanced Computing and Imaging in Biomedicine, National Taiwan University, Taipei 106, Taiwan

ARTICLE INFO

Keywords:
Drug-responses
Single-cell transcriptomics
Machine learning
Precision medicine

ABSTRACT

Predicting drug responses based on individual transcriptomic profiles holds promise for refining prognosis and advancing precision medicine. Although many studies have endeavored to predict the responses of known drugs to novel transcriptomic profiles, research into predicting responses for newly discovered drugs remains sparse. In this study, we introduce scDrug+, a comprehensive pipeline that seamlessly integrates single-cell analysis with drug-response prediction. Importantly, scDrug+ is equipped to predict the response of new drugs by analyzing their molecular structures. The open-source tool is available as a Docker container, ensuring ease of deployment and reproducibility. It can be accessed at https://github.com/ailabstw/scDrugplus.

1. Background

Precision medicine, also known as "personalized medicine," is a rapidly advancing field that aims to improve the effectiveness of medical treatments by taking into account the unique genetic, environmental, and lifestyle factors of individual patients [1]. While there are plenty of medications designed for the average patient, they may not effectively treat the specific conditions of some patients. In contrast, precision medicine is a computational approach that integrates patient medical data to tailor treatments to the specific needs of each patient [1]. In oncology, cancers often result from the accumulation of gene mutations that can lead to differential gene expression and contribute to cell death resistance and sustained proliferative signaling, which are hallmarks of cancer [2,3]. High-throughput sequencing (HTS) technologies have made it easier for people to access their own omics data, such as genomic and transcriptomic data, which can help determine precise genomic profiling and provide more information to make a diagnosis or identify customized treatment options. With the aid of genomic and transcriptomic data, we can capture the characteristics of an individual and potentially reveal whether heterogeneity exists in a tumor [4]. Clonal diversity in cancer cells can lead to drug resistance and poor prognosis due to the presence of drug-tolerant cells [5]. Therefore, identifying heterogeneity in cancer and making early inferences about drug response based on expression profiles are crucial for improving prognosis and achieving personalized medicine [6]. Although it is crucial to verify whether a drug is suitable for a specific genomic profile, such verification currently relies on time-consuming and costly wet lab experiments. To avoid investing in inactive molecules, it would be advantageous to screen for potential drug candidates. Consequently, various studies have explored computational methods for predicting drug response, which are collectively known as drug response prediction (DRP) methods, based on large-scale drug-screening data from cell line profiles [7–10]. These methods can facilitate the identification of promising drug candidates.

There are two commonly used measurements of drug response, namely the half-maximum inhibitory concentration (IC_{50}) and the area under the dose-response curve (AUC). IC_{50} is a measure of the potency of a substance in inhibiting a specific biological or biochemical function by 50 %. The use of AUC as a drug-sensitivity metric was first introduced by Fallahi-Sichani *et al.* in 2013 [11]. To obtain the shape of the

* Corresponding author.
** Corresponding author at: Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taiwan.
E-mail addresses: hsuancheng@nycu.edu.tw (H.-C. Huang), yukijuan@ntu.edu.tw (H.-F. Juan).

https://doi.org/10.1016/j.biopha.2024.117070
Received 19 April 2024; Received in revised form 18 June 2024; Accepted 29 June 2024
Available online 3 July 2024

dose-response curve, cell viability under different dosages is fitted to a four-parameter logistic (4PL) regression model, as shown in the following equation:

$$y = E_{\inf} + \left(\frac{E_0 - E_{\inf}}{1 + \left(\frac{D}{EC_{50}} \right)^{HS}} \right), \quad (1)$$

where y is the cell viability at a specific dosage D; E_0 and $E_{\inf}$ indicate the upper and lower asymptotes of the dose-response curve, respectively, and represent the drug efficacy; EC_{50} is the concentration at half-maximum effect, which is equal to IC_{50} when $E_0 = 1$ and $E_{\inf} = 0$, and indicates the potency of the drug; and the Hill slope (HS) is the slope parameter, similar to the Hill coefficient. The integration of the fitted dose-response curve is defined as the AUC, which combines the potency and efficacy of a drug into a single indicator [11]. From a biological standpoint, when a drug has a higher IC_{50} or AUC value in a cell line, it is more likely to be inactive in that cell line. Although IC_{50} is typically used in large-scale analysis of cellular response to anti-cancer drugs, it assumes that 100 % effect is achieved, which means all cells died in the proliferation assay at a high concentration. However, there are an increasing number of anti-cancer therapies that focus on cytostatic drugs rather than cytotoxic drugs. Cytostatic drugs are considered less harmful to surrounding cells because they inhibit the cell cycle and offer potential for long-term management. Despite these benefits, they present challenges such as limited efficacy and the development of resistance. In contrast, cytotoxic drugs induce cell death, providing faster therapeutic effects. The maximal drug effect ($A_{\max}$) of the dose-response curves of cytostatic drugs is less than 100 %, and thus the use of IC_{50} might not be appropriate for all anti-cancer drugs [12]. It should be noted that the AUC values of a drug can be compared between different cell lines under the same range of concentration, but the comparison of different drugs might be inaccurate, owing to the variability of dose ranges.

In computational chemistry, the precise representation of molecules in machine-readable formats is crucial for effective downstream analysis and prediction. Various molecular representations exist, from string types like registry systems (e.g., CAS Registry Number and PubChem CID) [13] [14] [15] and structure-based notations like Simplified Molecular Input Line Entry System (SMILES) [16] and the International Chemical Identifier (InChI) [17,18]. SMILES and InChI provide different levels of detail and readability. Feature-based representations of molecules, known as molecular descriptors (MD), are numerical representations of physicochemical information that can be divided into two categories: experimental measurements with physicochemical properties and theoretical molecular descriptors [19]. Molecular descriptors, numerical representations of physicochemical data, span from Zero-dimensional (0-D) descriptors, such as atomic weight, and one-dimensional (1-D), to 3-D descriptors that include molecular conformation. 1-D descriptors consider the local structure of molecules and are usually represented as fingerprints, developed based on different segmentation rules, such as path-based fingerprints like the atom-pair (AP) [20] and RDKit Daylight-like fingerprints [21] and circular fingerprints like the extended connectivity fingerprint (ECFP) [22]. A previous study showed that twelve different fingerprints were highly correlated with each other and had no statistically significant difference [23]. These descriptors, while powerful, sometimes overlap or collide in information. To aid computational operations, continuous representations within latent spaces have been proposed over traditional discrete ones. This continuous approach allows for more flexibility in molecular navigation. Advances in computer-learned molecular representations have given rise to encoders using autoencoders (AE) or natural language processing (NLP) methods, with models like SMILES2vec [24] and SMILESVec [25]. Many pre-trained models, built on architectures like Transformer and GNN graph neural networks (GNN) [26], utilize datasets like ZINC [27], QM9 [28], PubChem [29], CheMBL [30], and GEOM [31]. These models find application in drug discovery and molecular property prediction. However, the lack of a reliable and realistic benchmark makes the evaluation of these molecular pre-trained models unreliable [32].

Many large-scale pharmacogenetic databases such as The Cancer Genome Atlas (TCGA) [33] and the Cancer Cell Line Encyclopedia (CCLE) are now accessible, offering genomic and transcriptomic data from cancer patients or cancer cell lines. Some databases, like the Genomics of Drug Sensitivity in Cancer (GDSC) database [34], the Cancer Therapeutics Response Portal (CTRP) database [35], and the PRISM drug repurposing resource [36], also contain extensive drug screening results. These datasets, which provide drug sensitivity profiles, enable the exploration and prediction of drug responses by merging various omics data using computational techniques. They are critical tools for both drug discovery and personalized medicine. As the number of these large-scale pharmacogenetic datasets grow, several computational strategies for predicting drug response have emerged. These include linear regression models and their extensions, matrix-factorization-based (MF) methods, and machine-learning (ML) techniques. Drug-response prediction is a supervised learning problem that aims to compute the responses of three types including (i) a known drug in a new cell line, (ii) a new drug in a known cell line, or (iii) a new drug in a new cell line. Such predictions center on the idea that drugs with analogous characteristics in a specific cell line will likely yield comparable results. Similarly, drugs are anticipated to show parallel responses in cell lines with matching genomic profiles. Techniques like ridge regression [37] and elastic net [9] have been applied to assume a linear relationship between drug response and the genomic profile. Models like the similarity-regularized matrix-factorization (SRMF) method [10] are influenced by matrix factorization and integrate the similarities among cell lines and drugs, along with L2 regularization terms. CaDRReS model [38], developed by Suphavilai et al., estimates drug response by factoring in additional gene expression similarity data. Other models utilize ML methods, such as MolecularGNN_smiles [39], DrugGCN [8], and Chawla et al.'s approach which integrates a deep neural network (DNN) with gene set variational analysis (GSVA) score and molecular embedding [7]. Notably, most models use IC_{50} as the primary measure for drug response prediction. A unique approach, scDrug [40], created by our team, uses single-cell RNA sequencing (scRNA-seq) for malignant cluster identification and employs the gene expression profiles of these clusters for drug response prediction. The model in scDrug modifies the CaDRReS model [38] and predicts drug responses of recognized molecules based on normalized AUC. By using the scDrug framework, users can analyze scRNA-seq data to identify molecules that can target malignant clusters. Nonetheless, the existing scDrug drug-response prediction model only predicts known molecules, which limits its drug-repurposing potential. In this study, we have developed scDrug+ to extend the predictive capabilities of the original scDrug framework, advancing from type-i to type-iii response predictions by incorporating molecular structural information. This enhancement enriches scDrug with a drug discovery functionality, and is expected to facilitate more accurate predictions tailored to unique gene expression profiles and novel drugs. The expansive scope of our study, along with the detailed molecular representations and computational methodologies utilized, are graphically depicted in Fig. 1 and Fig. 2, respectively.

2. Results

2.1. Data preparation, model selection, and performance evaluation

In this study, we derived clean datasets for drug responses from the PRISM and GDSC databases. The refined PRISM dataset comprised 1441 molecules and 476 cell lines, leading to 616,017 cell–drug pairs, whereas the GDSC dataset included 125 molecules and 291 cell lines, totaling 29,813 cell–drug pairs (Supplementary Table S1).

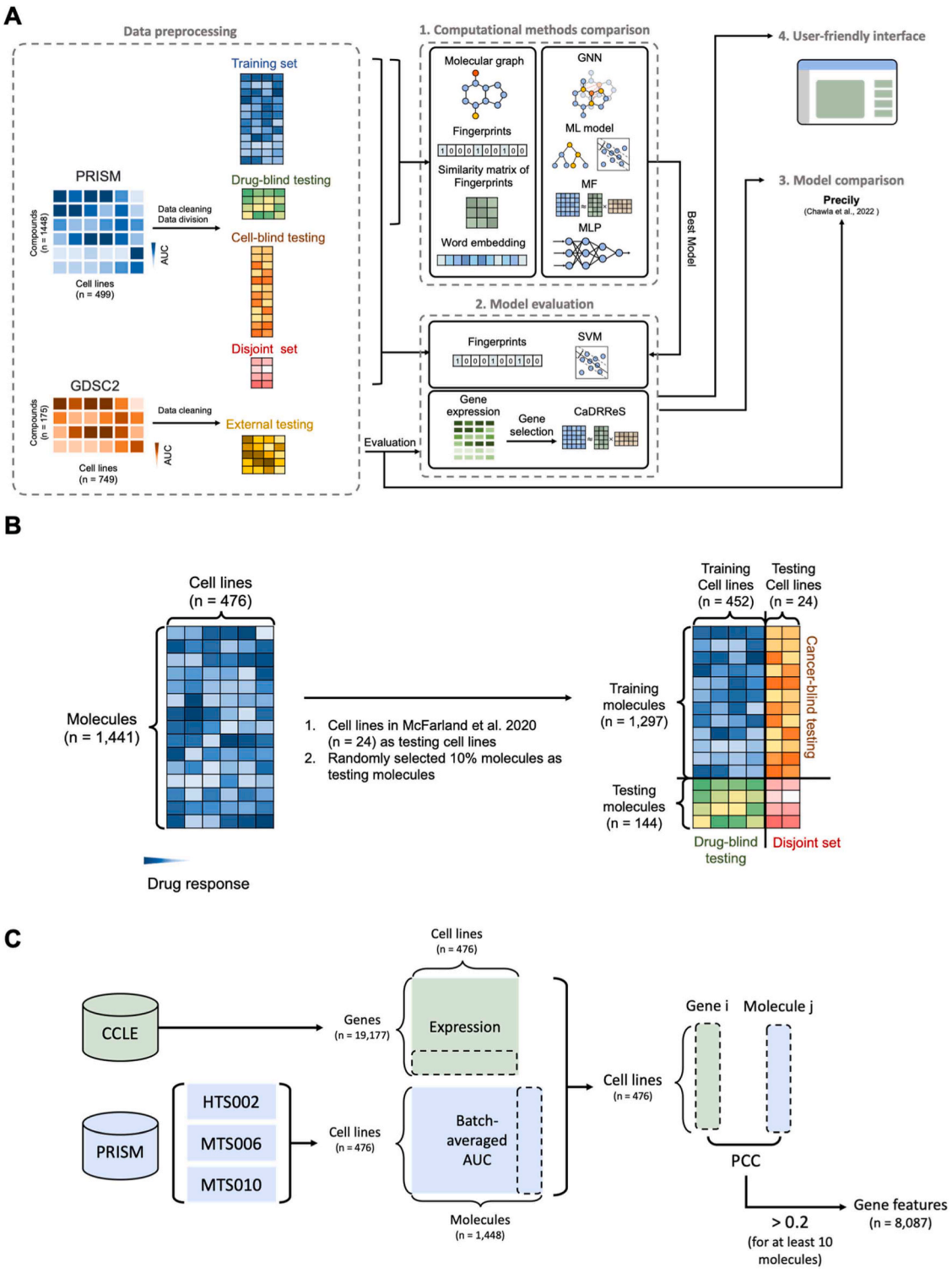

Fig. 1. Overview of this study. (A) The workflow of this study. Data cleaning and division procedures were first performed on the PRISM and GDSC datasets. The training set and drug-blind testing set from PRISM were used to train and evaluate different combinations of molecular representations and computational methods. The best-performing model was then used to construct a combined model with the modified CaDRReS model and trained on the training set from PRISM. Finally, the disjoint set from PRISM and the external testing set from GDSC2 were used to evaluate the combined model. (B) Data division procedure of the PRISM dataset. After undergoing a data cleaning procedure, the drug response data from PRISM was divided into 4 sets: a training set, a drug-blind testing set, a cell-blind testing set, and a disjoint set. Twenty-four cell lines from McFarland *et al.* [50] were used as testing cell lines, and 10 % of the molecules were randomly selected for testing. (C) The feature selection process of gene expression profile. As per the feature selection process for gene expression described by Hsieh *et al.* [40], batch-averaged AUC values were calculated when duplicates were present in different batches of the PRISM dataset (HTS002, MTS006, MTS010). Gene expression data for 476 cell lines per gene were obtained from the CCLE dataset, and the AUC values for 476 cell lines per molecule were used to calculate PCC. Genes with an absolute PCC over 0.2 for at least 10 molecules were retained as feature genes, resulting in 8087 genes being selected.

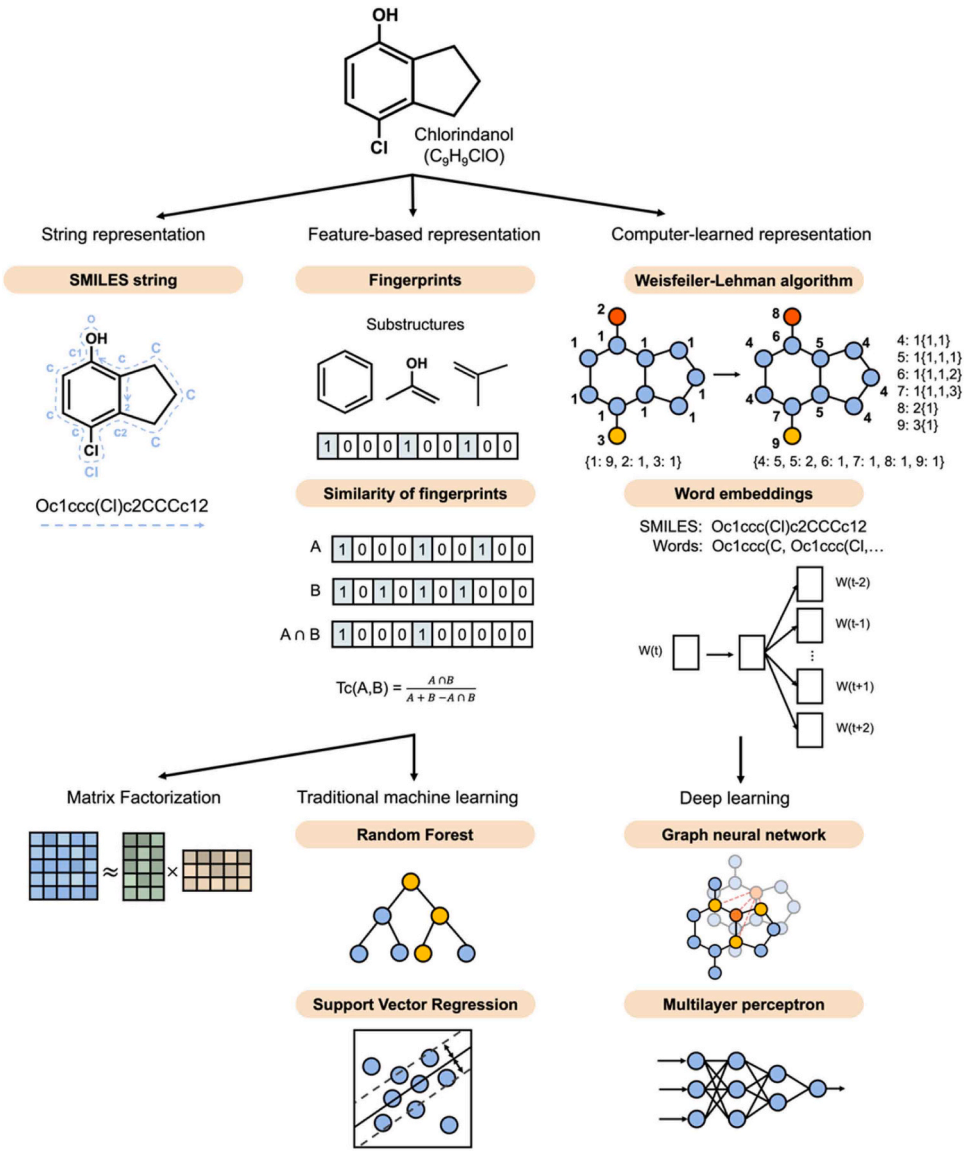

Fig. 2. Overview of molecular representations and computational methods used in this study. A molecule, such as Chlorindanol, will first be converted into a SMILES string (Oc1ccc(Cl)c2CCCc12) based on its molecular structure. The SMILES string of the molecule is then represented using various formats. We utilize the fingerprint, a feature-based representation, which is a vector indicating the presence or absence of specific substructures in the molecule. Some computational representations were also utilized in this study. The Weisfeiler–Lehman algorithm was applied to the molecular graph, where blue nodes represent carbon atoms, red nodes represent oxygen atoms, and yellow nodes represent chlorine atoms in this example. Labels were assigned to each node based on the neighboring nodes in the molecular graph. Furthermore, the Word2vec model was employed to convert the SMILES string into word embeddings. Molecular similarity was assessed using Tanimoto similarity and cosine similarity. Different molecular representations can be used as inputs for distinct computational methods. For instance, fingerprints and molecular similarity can be used as inputs for traditional machine-learning models and matrix-factorization methods. The molecular graphs can be served as inputs for graph neural networks (GNN), while word embeddings can be utilized as inputs for multilayer perceptron (MLP) models.

We then divided the PRISM dataset into a training set (1297 molecules, 452 cell lines), a drug-blind testing set (144 molecules, 452 cell lines), a cell-blind testing set (1297 molecules, 24 cell lines), and a disjoint testing set (144 molecules, 24 cell lines) (Fig. 1**(B)**, Supplementary Table S2).

After performing the feature selection procedure described in ***Methods*** and shown in Fig. 1 **(C)**, we obtained gene expression profiles for 8087 genes of cell lines from the CCLE dataset. The training set and drug-blind testing set were used to train and evaluate the models mentioned in ***Methods*** to determine the most appropriate model for constructing the combined model. Supplementary Table S3 shows the hyperparameters of the models.

To compare the performance of ML models with a default set of parameters, we evaluated four combinations: (1) Support Vector Machine (SVM) with RDKit Daylight-like fingerprints as input features, (2) SVM with similarity of RDKit Daylight-like fingerprints as input features, (3) Random Forest (RF) with RDKit Daylight-like fingerprints as input features, and (4) RF with similarity of RDKit Daylight-like fingerprints as input features. Table 1 shows the model performance comparison results of machine learning methods. Regression evaluation metrics were used to assess the model predictions of the four models, as shown in Fig. 3**(A)**. Among the four models, the SVM model using RDKit Daylight-like fingerprints as input features demonstrated better cell-wise Pearson's correlation coefficients (PCC) performance than the other models. However, no significant differences between the four models were observed based on drug-wise PCCs. Therefore, we chose to further compare the SVM model with RDKit Daylight-like fingerprints as input features against other types of models.

Table 1
Model performance comparison of machine learning methods.

Model	Cell-wise		Drug-wise	
	Median PCC	Median MSE	Median PCC	Median MSE
SVM_RDKfp	0.466	0.021	0.213	0.012
SVM_simMat	0.436	0.021	0.180	0.010
RF_RDKfp	0.460	0.020	0.215	0.012
RF_ simMat	0.350	0.023	0.230	0.010
MF	0.404	0.023	0.235	0.007
1WL_GNN	0.344	0.028	0.234	0.011
SMILESVec	0.269	0.032	0.223	0.010

The performance of the SVM_RDKfp model was then compared with three other models mentioned in ***Methods***, namely the 1-WL GNN model, the CaDRReS-like model, and the model using a multi-layer perceptron with SMILESVec embeddings as input features, referred to in this study as the 1-WL model, MF model, and SMILESVec model, respectively. The results of the comparison of the four models are presented in Fig. 3**(B)**. The SVM_RDKfp model outperformed the other models in cell-wise evaluation, with a higher distribution of PCCs and a lower distribution of mean squared error (MSE). Although there was no significant difference in the drug-wise PCCs between the four models, the SVM_RDKfp model had the lowest outlier value of MSE in the drug-wise evaluation, suggesting that its predictions were more robust and stable. Therefore, the SVM_RDKfp model was used to construct the combined model for predicting the drug response of new drugs and new cell lines.

2.2. In-depth analysis of molecular similarity, data integrity, and their impacts on drug response model performance

In this section, we further conduct analysis to investigate the factors influencing model performance in drug-response prediction. Based on the assumption which states that similar drugs would have a similar effect on the same cell line, we initially calculated the similarity of corresponding drug response (AAC) vectors which is defined in ***Methods*** between different molecules to determine the similarity of drug response. To address missing values in the AAC vectors, we first padded them with 0 before applying cosine similarity to the padded AAC vectors. Additionally, we used different molecular encoders in the models to encode the molecules into computationally readable embeddings, which were then used to compute the similarity of molecules. Specifically, we utilized fingerprints with a dimensionality of 420 from the 1-WL algorithm, embeddings of dimensionality 50 from the 1-WL fingerprints that underwent GNN processing, RDKit Daylight-like fingerprints of dimensionality 2048, and word embeddings of dimensionality 100 from the SMILESVec pre-trained model to compute the similarity of molecules.

To compute the similarity of 1-WL fingerprints, we first converted them into one-hot encoding. Tanimoto similarity was used to calculate the molecular similarity of RDKit Daylight-like fingerprints, while cosine similarity was used for the other molecular representations. The

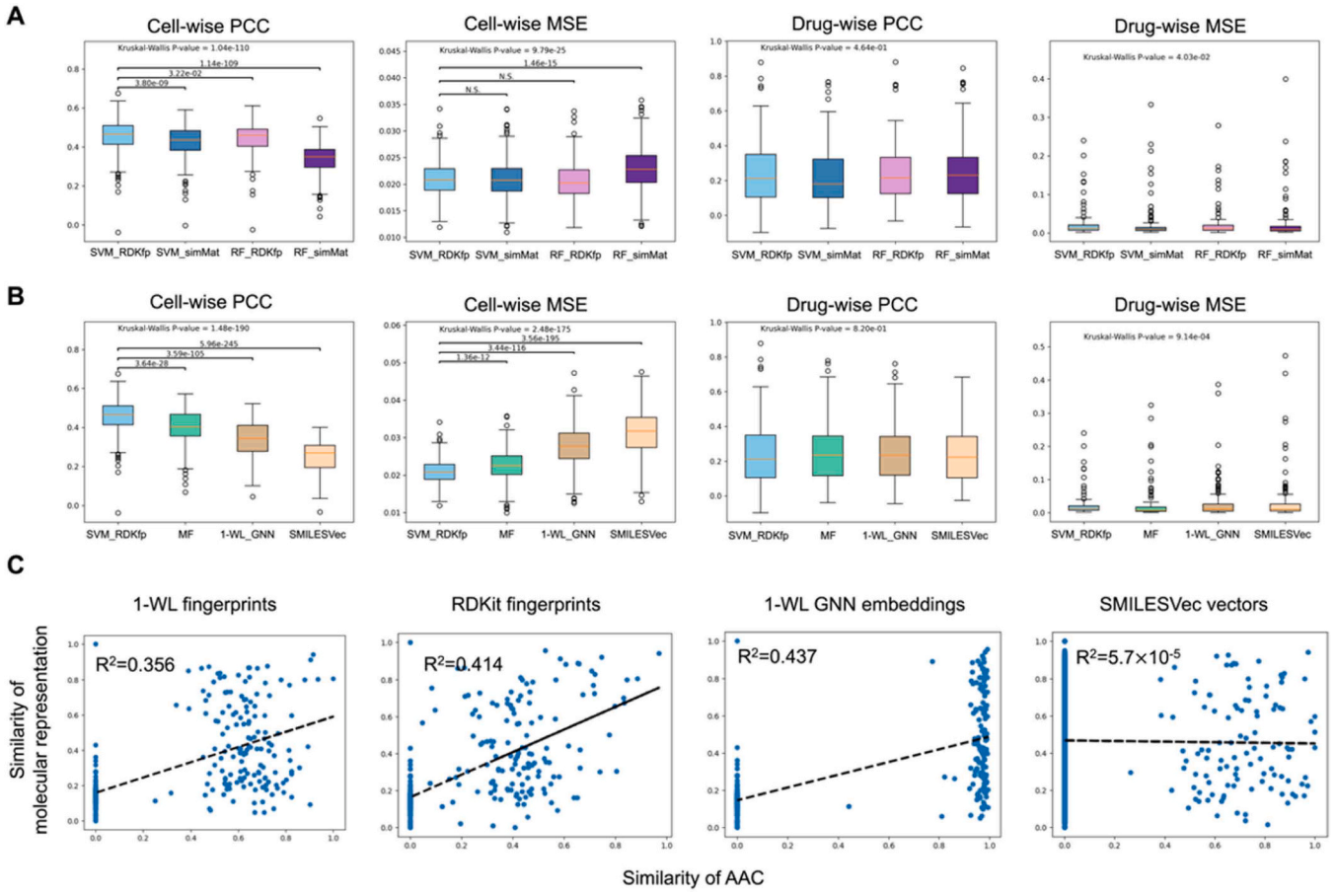

Fig. 3. Comparison of the performance of computational methods on drug-response prediction for new drugs on known cell lines. (A) Evaluation results of the ML models. The cell-wise PCC demonstrates that the SVM_RDKfp model outperformed other models with a low MSE, while all four ML models showed similar performance and MSE distribution from a drug-wise perspective. (B) Evaluation results of different types of models in this study. The cell-wise PCC demonstrates that the SVM_RDKfp model outperformed other models with a low MSE, while all four types of models showed similar performance and MSE distribution from a drug-wise perspective. (C) Scatter plots of the similarity of molecular representation and the corresponding drug-response (AAC) similarity. In the four molecular representations, the molecular similarity from 1-WL fingerprints and RDKit Daylight-like fingerprints were found to be linearly correlated with drug-response similarity.

upper triangular matrix of each molecular similarity matrix was extracted and the molecule with maximal similarity was selected. The selected molecular similarity and its corresponding drug response similarity were displayed on a scatter plot, and linear regression was used to determine if there was a linear relationship between the molecular similarity and drug response similarity.

Fig. 3**(C)** displays the scatter plot of the four molecular representations. The molecular similarity from 1-WL fingerprints and RDKit Daylight-like fingerprints were found to be linearly correlated with drug-response similarity, with R^2 values of 0.356 and 0.414, respectively. However, the molecular similarity from embeddings of the SMILESVec pre-trained model did not exhibit a linear relationship that aligned with the drug-response assumption. While the molecular similarity of embeddings from the GNN seems to be linearly correlated with the drug-response similarity, the GNN model might overclassify molecules into similar or unsimilar representations. In general, the 1-WL fingerprints and the RDKit Daylight-like fingerprints appear to be more appropriate for use in this study to support the assumption that similar molecules have similar drug responses.

Based on the cell-wise evaluation shown in Fig. 3**(A)** and **(B)**, we observed that some cell lines had inaccurate predictions. We hypothesized that the proportion of ineffective molecules (AAC = 0) in these cell lines in the training set could influence the predictions of the models. To investigate this, we generated scatter plots to depict the proportion of ineffective molecules in cell lines and the corresponding cell-wise PCCs derived from the predictions of the models, as shown in Supplementary Figure S2**(A)**. The results indicated that the cell-wise PCCs of most models in this study had a slightly negative linear correlation with the proportion of ineffective molecules, except for the 1-WL GNN model. However, it should be noted that the R^2 values for all models in this study were close to 0, suggesting that the correlation was not strong.

2.3. Impact of missing data on model performance and comparative analysis in drug response predictions

In addition to the proportion of ineffective molecules, we also hypothesized that the proportion of missing values in cell lines in the training set would impact the performance of the models. To investigate this hypothesis, scatter plots were generated to examine the relationship between the proportion of missing values in cell lines and the corresponding cell-wise PCC derived from the predictions of the models, as shown in Supplementary Figure S2**(B)**. The scatter plots indicated that the cell-wise PCC from the MF model, 1-WL GNN model, and SMILESVec model were linearly correlated with the proportion of missing values in the cell lines, with R^2 values of 0.307, 0.138, and 0.124, respectively. However, the SVM_RDKfp model appeared to be unaffected by the proportion of missing values in cell lines in the training set.

Based on the aforementioned results, we used the SVM_RDKfp model to construct a combined model for predicting drug response in new molecules and new cell lines. We first trained the CaDRReS model using the training set to obtain predictions for the cell-blind testing set, which included new cell lines and known molecules. These predictions were then used to train the SVM_RDKfp model for predicting drug response in new cell lines and new molecules, *i.e.*, the disjoint set. We compared the performance of the combined model to the CaDRReS model trained with the concatenation of the training set and drug-blind testing set, as well as the SVM_RDKfp model trained with the cell-blind testing set. The comparison results, as shown in Supplementary Figure S2**(C)** and Supplementary Table S4, indicate that the performance of the combined model was close to that of the SVM_RDKfp model, with a median cell-wise PCC of 0.389 and 0.444, respectively, while the CaDRReS model had a median cell-wise PCC of 0.833. And the median cell-wise MSE of the combined model and the SVM_RDKfp model was slightly higher than the CaDRReS model (0.021 and 0.023 vs. 0.007). The median drug-wise PCC was approximately the same in all three models (median PCC = 0.131, 0.251, and 0.154 for the combined model, SVM_RDKfp model, and CaDRReS model, resprectively), and the median drug-wise MSE of the combined model and the SVM_RDKfp model was higher than that of the CaDRReS model (0.021 and 0.013 vs. 0.007).

2.4. Comparative assessment of the combined model and precily on drug response prediction from GDSC and PRISM datasets

We further tested the stability of the combined model using the GDSC external testing set, and the evaluation results of the GDSC external set and the PRISM disjoint set are shown in Fig. 4**(A)** and Table 2. For regression evaluation metrics, only the SCC was computed to evaluate the performance of the models on the two datasets. As shown in Fig. 4**(B)**, the median SCC of the PRISM disjoint set (median cell-wise SCC = 0.273, median drug-wise SCC = 0.084) was better than that of the GDSC external testing set (median cell-wise SCC = 0.270, median drug-wise SCC = −0.043), and the distribution of cell-wise SCCs of the GDSC external testing set (IQR = 0.653) was wider than that of the PRISM disjoint set (IQR = 0.091). For classification evaluation metrics, the cell-wise F1 score of the PRISM disjoint set (median cell-wise F1 score = 0.746) was much better than that of the GDSC external testing set (median cell-wise F1 score = 0.229) and the drug-wise F1 score of the PRISM disjoint set (median drug-wise F1 score = 0.696) was also better than that of the GDSC external testing set (median drug-wise F1 score = 0.351).

Drug-response prediction assumes that cell lines with similar gene expression profiles would have similar drug responses when treated with the same molecule, and similar molecules would have similar drug responses on the same cell lines. To confirm this hypothesis, we investigated the cell-line similarity and molecular similarity in the PRISM dataset and the GDSC external dataset. First, we employed t-SNE to analyze molecular similarity derived from the RDKit Daylight-like fingerprint, as well as the cell line kernels obtained from the CaDRReS model, which were the cell line similarity of the cell lines in the two testing sets and the cell lines in the PRISM training set. The results are shown in Fig. 4**(C)** and Supplementary Figure S3**(A)**. The t-SNE and k-means clustering results illustrated in Supplementary Figures S3**(A)** and **(B)** reveal that molecules from the PRISM training set, the PRISM disjoint set, and the GDSC external testing set were all intermingled, with no distinct clustering observed for any of the testing sets. Furthermore, the violin plot of the KNN (k-nearest neighbors) class probabilities for the neighbors of testing molecules to be training molecules was depicted in Supplementary Figure S3**(C)**. This plot indicates that the class probabilities of the testing molecules were high, with all of them exceeding 50 %. This suggests that the testing molecules bore similarity to the training molecules. On the other hand, the t-SNE results illustrated in Fig. 4**(C)** show that, based on the computed kernels, the cell lines in the PRISM training set, the PRISM disjoint set, and the GDSC external testing set could be clustered into one major and two small clusters. One of the small clusters was composed of only the cell lines from the GDSC external testing set, which indicates that some of the cell lines in the GDSC external testing set were different from the cell lines in the PRISM training set. This might be the reason the combined model failed to perform well on the GDSC external testing set.

We further analyzed the lineage of the cell lines in the three sets. Supplementary Figure S3**(D)** shows the composition of the types of lineages in the PRISM training set, with the lung lineage providing the most cell lines. The cell lines in the PRISM disjoint set and the GDSC external testing set were then classified by their lineages. Fig. 4**(D)** illustrates the t-SNE plot of the cell-line kernels colored by the lineages they belonged to and shows that three lineages, plasma_cell, lymphocyte and blood, did not contain the PRISM training set, but were present in the GDSC external testing set. The cell-wise SCC and cell-wise F1 score values of the PRISM disjoint set and the GDSC external testing set were classified by lineages and are illustrated in Supplementary Figures S3**(F) and (G)** and Fig. 4**(D) and** Supplementary Figures S3**(E)**, respectively. In Fig. 4**(D)** and Supplementary Figure S3**(E)**, the three lineages that

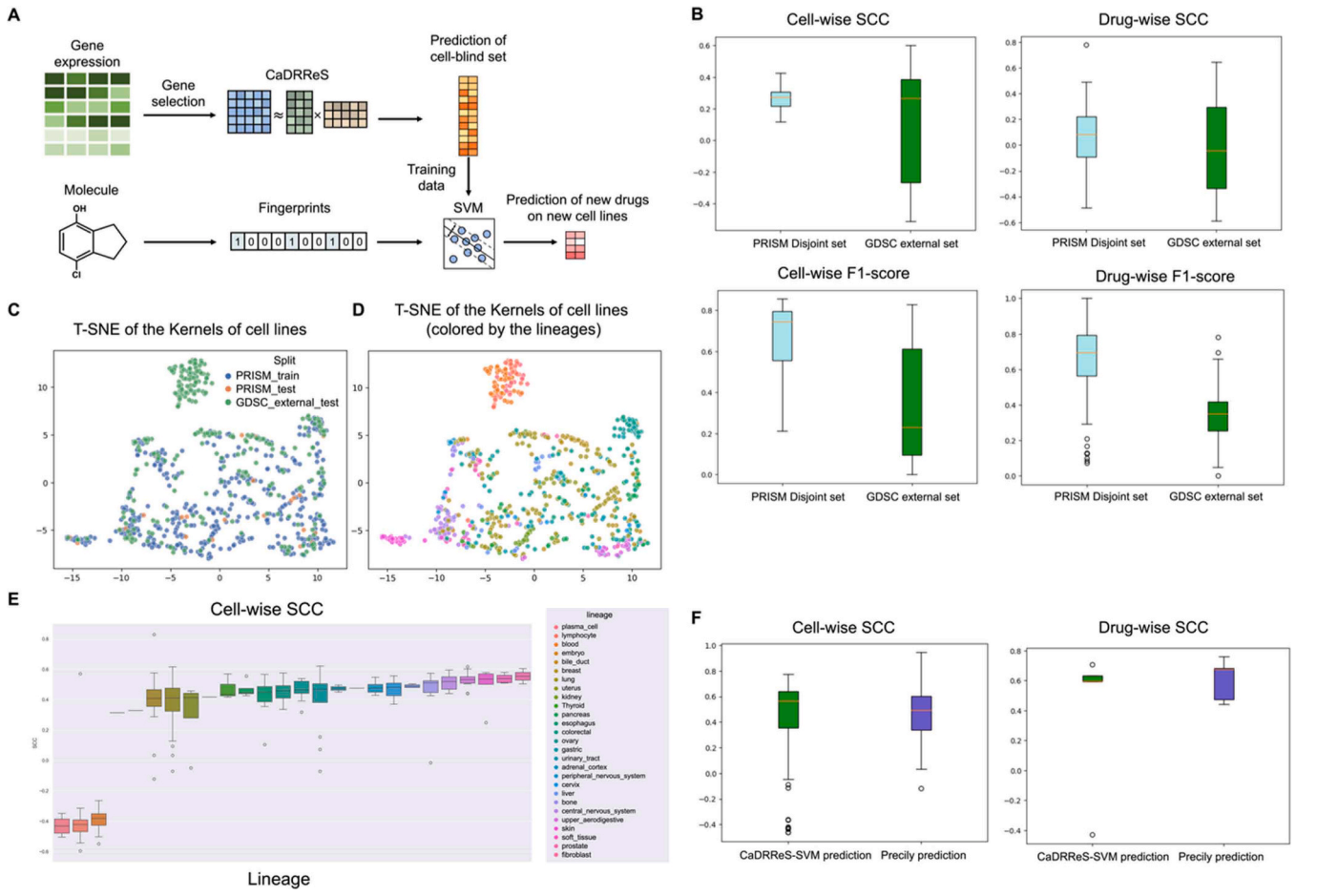

Fig. 4. Performance evaluation of drug response prediction for new drugs on new cell lines and lineage analysis of cell lines. (A) Model architecture of the combined model. The gene expression profile of the cell lines first underwent the gene feature selection process, and the kernels of cell lines were computed based on the gene expression of the remaining genes. The kernels of cell lines were then used to map the features of cell lines into a latent space in the CaDRReS model and obtain the drug-response prediction of existing drugs on new cell lines, which would be used as the training data of the following SVM model. On the other hand, the SMILES strings of the molecules were converted into fingerprints and used as the input feature of the SVM model. Finally, through the SVM model, the drug responses of new drugs on new cell lines could be predicted based on the molecular structure of the molecules and the gene expression profile of the cell lines. (B) Evaluation of the model on the PRISM disjoint set and the GDSC external dataset. The cell-wise SCC shows that the combined model performed better on the PRISM disjoint set than on the GDSC external testing set, whereas the drug-wise SCC indicates that the combined model could predict the rank of cell lines in the GDSC external set more accurately. Additionally, the cell-wise F1 score and drug-wise F1 score suggest that the classification of molecules in the PRISM disjoint set was more precise than in the GDSC external testing set. (C) The t-SNE plot of kernels of cell lines. There are one major cluster and two small clusters, and one of the small clusters was all composed of the cell lines from the GDSC external testing set, showed that these cell lines are different from the cell lines in the PRISM training set. (D) The t-SNE plot illustrating the distribution of kernels of the cell lines, with dots colored according to their respective lineages. The result showed that the cell lines in the small cluster were from “lymphocyte”, “plasma cell”, and “blood” lineages. (E) The boxplots show the cell-wise SCC of the GDSC external testing set classified by the cell line lineages. Cell-wise SCC values of cell lines in the three lineages (“lymphocyte”, “plasma cell”, and “blood”) were lower than those in other cell lines. (F) Comparison of the performance of our model and Precily on the GDSC dataset. The cell-wise SCC (on the left) shows that our model performed comparably to Precily, except for some of the cell lines. The drug-wise SCC (on the right) indicates that only one of the molecules had an inaccurate rank prediction.

Table 2
Model performance of the combined model on the PRISM testing set and the GDSC external testing set.

Dataset	Cell-wise		Drug-wise	
	Median SCC	Median F1-score	Median SCC	Median F1-score
PRISM	0.273	0.746	0.084	0.696
GDSC2	0.270	0.229	-0.043	0.351

were different from the cell lines contained in the PRISM training set were associated with low cell-wise SCC and F1 scores. This is consistent with a previous study that showed that similar cell lines and similar molecules would have similar drug responses [3].

Moreover, the distributions of F1-score values per cell exhibited greater variance compared to those of SCC values per cell, suggesting that the accuracy and precision of the classification of molecular status could be influenced by the varied dose ranges used to obtain observed drug-response values from experiments. To evaluate the potential impact of dose-range differences on the performance of the combined model, we examined the dose ranges of molecules screened in the PRISM training set. We found that most of the molecules had minimum and maximum dosages of approximately 0.00015 and 10 μM. Consequently, we set the dose-range threshold at 9.999 μM to filter out molecules with similar dose ranges. After filtering, 1225 and 142 molecules remained in the PRISM training set and disjoint set, respectively. These molecules, which were screened with the same dose range, were used to retrain the combined model. The results of the original and retrained combined models are presented in Supplementary Table S4. The comparison showed that the average SCC values per cell slightly improved for the PRISM disjoint set and the GDSC dataset. The cell-wise average F1-score values slightly improved for the PRISM disjoint set, but slightly

worsened for the GDSC dataset. These findings indicate that training and testing the model with molecules screened within the same dose range might increase the comparability of drug responses between different molecules in the same cell line. However, the prediction of drug response for molecules screened with different dose ranges, such as those in the GDSC external testing set, may deviate from observed values.

We conducted a comparative analysis between our model, trained on the PRISM dataset, and Precily [7], a drug-response prediction model capable of predicting drug response on new cell lines and molecules. Precily, which is trained on the GDSC dataset, employs a two-step process. First, it transforms the gene expression profiles of the cell lines and the SMILES string representations of the molecules into vectors using GSVA scores and word embeddings, respectively. Next, it concatenates these vectors and then feeds them into a DNN for further processing. To assess the performance of both our model and Precily, we selected five molecules that were not included in the training of Precily. These molecules were used to evaluate the predictive capabilities of both models. It is worth noting that the cell lines on which the five molecules were screened had been previously used in the training of Precily but not in our model. The comparison results, depicted in Fig. 4**(F)**, indicate that our model performs on par with Precily, as evidenced by median cell-wise SCC values of 0.564 and 0.494 for our model and Precily, respectively. A comprehensive overview of the median cell-wise SCC and drug-wise SCC values for both models is presented in Table 3. The final drug-response prediction model, which combines matrix factorization and SVM, has been integrated into the scDrug pipeline [40]. The open-source docker tool can be accessed at https://github.com/yysun0116/scDrugplus.git.

3. Discussion

Precision medicine relies on the accurate prediction of drug response based on a specific gene expression profile. Our team has developed scDrug [40], which is an adaptation of the CaDRReS model, designed to forecast the drug reaction (AUC) for recognized drugs based on the gene expression patterns from malignant cell clusters found in single-cell RNA sequencing. In this study, we've expanded scDrug's capability from predicting type-i drug reactions (a known drug in a new cell line) to type-iii (a new drug in a new cell line) by devising a composite model. Fig. 3**(A)** and **(B)** present the evaluation findings of the machine learning and deep learning models applied, with testing done on the PRISM blind-drug dataset. The outcomes indicate that the SVM model, using the standard parameter set and RDKit Daylight-like fingerprints as input (SVM_RDKfp model), surpassed its counterparts. This enhanced efficacy could be linked to the existence of "activity cliffs" in the drug response and structure association. This enhanced efficacy could be linked to the existence of "activity cliffs" in the drug response and structure association [41].

Fig. 3**(C)** illustrates that RDKit Daylight-like fingerprints effectively represent the linear correlation between molecular similarity and drug-response similarity, with an R^2 value of 0.414). This suggests that such a molecular depiction is more suited for tasks involving drug-response prediction. An earlier research indicated that various molecular representations, including SMILES strings, fingerprints, molecular graphs, and those based on deep learning, have comparable performances in predicting drug-response [42]. While linear notations and fingerprints are straightforward and yet potent, there might be a necessity for pre-training and multi-task learning for enhancing the performance of graph-based molecular representations [43].

Previous studies [7,44–46] have addressed the issue of anti-cancer drug response prediction, primarily conducting experiments on the GDSC and CTRP datasets and focusing on missing drug responses in known cell lines or for known drugs, similar to our cancer-blind and drug-blind testing. In contrast, our key achievement is the capability to predict responses for unknown drugs on unknown cell lines. Another advantage of our approach is the use of the PRISM dataset to build our models. This dataset covers a much larger set of tested drugs compared to GDSC and CTRP. This extensive coverage allows us to investigate drug responses for novel drugs on previously unseen cell lines, providing a broader scope for discovery and enhancing the robustness and applicability of our model. While other works [47,48] have incorporated drug features such as targets, enzymes, chemical substructures, pathways, and mono side effects to predict drug responses, these frameworks are limited in their ability to predict responses for novel drugs due to the issue of missing features. Our work, on the other hand, utilizes only the chemical structure, enabling the prediction of responses for novel drugs.

Fig.entary Figures S2(A) and **(B)** demonstrate that the SVM_RDKfp model's performance (measured by cell-wise PCC) remains unaffected by the percentage of non-reactive molecules and the fraction of absent values in cell lines within the training dataset, with R^2 values being 0.038 and 0.023, respectively. If the cell-wise PCCs have a stronger correlation with the percentage of non-reactive molecules, indicating that the number of non-reactive molecules in the cell lines affects the optimization of the model performance, negative sampling would be suggested to enhance the training process of the model. The outcomes of the integrated model when applied to both the PRISM disjoint dataset and the GDSC external evaluation set are depicted in Fig. 4**(B)**. While there was room for improvement in the correlation between actual and forecasted AUC values at the molecular level, the molecules were aptly categorized into three categories: "inactive", "ambiguous", and "promising". For classification on the PRISM disjoint testing set, the model registered a median cell-wise F1 score of 0.746. Additionally, the model effectively discerned the molecules' relative response, achieving a median cell-wise SCC of 0.273. However, the combined model's efficacy on the GDSC external test set wasn't as commendable as on the PRISM disjoint test set, as detailed in the ***Results*** section.

This study presents several constraints that warrant consideration. Firstly, the forecasted AUC values for the molecules might not mirror the accuracy of the actual observed values. Secondly, the trustworthiness of the projected AUC ranking and the categorization tiers for the molecules is primarily restricted to those analyzed within a dosage spectrum of roughly 0.00015–10 μM. Thirdly, there's potential for disparity between the projected AUC rank and the actual rank, especially if the gene expression patterns of the predicted cell lines vary considerably from those used during the model's training phase. Lastly, while our model is principally designed to forecast drug reactions based on gene expression data and molecular configurations, integrating further data points, such as cell-specific copy-number values, might potentially bolster the precision of the drug-response predictions.

Drug combinations have been a widely adopted approach to overcome drug resistance and treat complex diseases such as cancer. Our original scDrug enables us to tailor combination therapies by simultaneously targeting multiple pathways involved in the disease. By selecting drugs that reverse the gene expression signature of a disease, we can identify potential drug combinations. However, since the algorithm relies on the observed gene expression signatures of known drugs, it cannot be directly extended to novel drugs in scDrug+. An important future direction is to develop a strategy for predicting the gene expression signatures of novel drugs. This would allow us to find better combination treatments that include the most sensitive drugs, which can also reverse the disease signature and target different genes and pathways

Table 3
Model performance comparison of our model and Precily.

Model	Predicted measurement	Median cell-wise SCC	Median drug-wise SCC
Precily (Chawla et al., 2022)	IC50	0.494	0.676
Our model	AUC	0.564	0.597

involved in the disease-related gene signature.

scDrug+ benefits the field of biomedicine by promoting personalized medicine and expediting drug discovery. It enables the creation of personalized treatment plans by predicting drug responses based on individual transcriptomic profiles at the single-cell level. Additionally, scDrug+ improves the drug discovery processes by forecasting responses to new drugs, thereby saving time and resources in drug development. Furthermore, scDrug+ integrates large-scale pharmacogenetic databases and increases accessibility through its open-source platform, facilitating widespread use and collaborative improvements within the research community.

4. Conclusions

Computational techniques are garnering increasing interest for predicting drug responses, especially within the precision medicine paradigm. This research took a departure from the widely-adopted IC_{50}, choosing instead the area under the drug-response curve (AUC). This shift aids in predicting both cytotoxic and cytostatic drugs more efficiently. We used matrix-factorization methods and an SVM with RDKit Daylight-like fingerprints to frame our prediction model, which is adept at predicting drug reactions for novel molecules in the realm of oncology. Our exploration evaluated different molecular representations for their effectiveness in drug prediction. Remarkably, fingerprint-based models demonstrated their prowess in highlighting the linear ties between molecular similarity and drug-response similarity, standing toe-to-toe with graph and deep learning models, all without needing pre-training or simultaneous multi-task learning. Our scrutiny of various computational techniques revealed the SVM approach, when employed with standard parameters, to excel beyond other predictive methods. Augmenting the Cancer Drug Response via a Recommender System (CaDRReS) model and integrating it with the SVM model rooted in RDKit Daylight-like fingerprints, we crafted a combined model. The combined model showed promise in classifying molecules as either "inactive," "unclear," or "potential" and captured the nuanced relationships between molecules in each cell lineage. In essence, our endeavor furnishes not just a potent tool for drug response prediction on novel molecules and diverse gene expression patterns but also offers a comparative analysis of molecular representations and computational drug prediction methods. Our model's assessment sets a benchmark for future endeavors in type-iii drug response prediction, employing AUC as a drug response indicator. To heighten our model's efficacy, we suggest incorporating pharmacogenetic databases from varied sources. Such integration can leverage a more extensive dataset, potentially refining prediction accuracy. Employing adjusted AUC values, as elucidated by Pozdeyev *et al.* [12], from multiple repositories might also hone the predicted outcomes' precision. Furthermore, some prior studies [49] have used molecular dynamics in drug-response prediction specifically for lung cancer, suggesting that the addition of information regarding target proteins could enhance drug response predictions for specific genetic expressions. Hence, there's ample scope to amplify the precision of drug-response prediction models in future research.

5. Methods

5.1. Data integration and preprocessing: harnessing the PRISM, GDSC, and CCLE datasets for drug response predictions

In this study, we utilized three datasets: the PRISM drug repurposing resource [36], the GDSC database [34], and the CCLE database [28]. The drug response matrix from the PRISM dataset incorporated the AUC for its measurements. Initially, the PRISM repurposing dataset (version 19Q4) evaluated 4518 compounds across 578 cell lines using a single dosage. Those compounds yielding positive results underwent a secondary screening, leading to 1448 compounds being assessed against 489 cell lines in triplicate across eight doses levels, ranging from 610 pM to 10 μM. A four-parameter logistic regression model was then applied to the cell viability data from these eight doses to craft drug-response curves. The GDSC2 dataset, a subset of the GDSC dataset (Release 8.1, Oct. 2019), screened 175 compounds against 794 cell lines, providing with two metrics: AUC and IC_{50}. We primarily utilized the the PRISM dataset to formulate the drug-response prediction model, while deploying the GDSC2 dataset as our external testing set. We derived the gene expression profiles of the cell lines from the CCLE dataset (version 21Q3). Further sections will delve into the intricacies of our data preprocessing approach. For reference, all the datasets harnessed in this research were sourced from the Dependency Map (DepMap) portal website (https://depmap.org/portal/download/all/).

Fig.mentary Figure S1(A) delineates the data purification process for the PRISM dataset. Prioritizing the integrity of cell line identity, we initially removed drug-response entries from cell lines that didn't pass the short tandem repeat (STR) profiling test. Acting on guidelines from the official DepMap forum, we preserved the entries with the most substantial screen ID for drugs subjected to multiple screenings, as the more recent screenings are generally superior in quality. We also observed that some AUC values exceeded 1, possibly due to data discrepancies; consequently, we adjusted these values to a maximum of 1. Post adherence to these steps, the remaining drug-response data included 1448 drugs against 480 cell lines.

For effective drug response prediction, it's imperative to have the gene expression profiles of the cell lines. Hence, we matched the cell lines from our refined PRISM dataset with those in the CCLE dataset, basing our comparison on the DepMap IDs of the cell lines. This led to the exclusion of four cell lines not listed in CCLE, leaving us with 476 cell lines represented in both datasets. We also removed compounds that had the symbol "." in their SMILES string, as it denotes a dissociation context. After these adjustments, a total of 1441 molecules were evaluated against 476 cell lines, resulting in 69,899 unique cell line-compound pairs post the data refinement process. Subsequently, we earmarked 10 % of the compounds for testing purposes and selected the same 24 testing cell lines from McFarland *et al.* [50], which were utilized in scDrug, as our test cell lines. As showcased in Fig. 1(B), this classification segmented the PRISM dataset into four distinct categories: the training set (known compounds and known cell lines), the cell-blind testing set (known compounds and unknown cell lines), the drug-blind testing set (known compounds and known cell lines), and the disjoint testing set (unknown compounds and unknown cell lines).

Fig.mentary Figure S1(B) outlines the process of data refinement for the GDSC dataset. Initially, we focused on sifting through the data from the GDSC2 dataset. Using the "get_compounds" function provided by the PubChemPy module (version 1.0) in Python (version 3.10.6), we fetched the SMILES strings of the molecules present in the GDSC2 dataset. For compounds whose SMILES strings couldn't be identified through PubChemPy, a manual search was conducted. Any compounds still elusive post this manual probe were discarded. For compound-cell line pairings with multiple entries, we consolidated the drug response data by averaging their values. Post this comprehensive cleaning regimen, our dataset showcased drug response data for 167 molecules tested against 793 cell lines. Given the GDSC dataset's role as an external testing set, we pinpointed and retained the cell lines consistent in both GDSC and CCLE datasets, post comparison with the PRISM dataset. Conclusively, our external test set comprised drug-response data for 125 compounds across 291 cell lines.

5.2. Navigating molecular representations: from graph isomorphism to drug prediction models

Distinguishing between distinct molecular structures is pivotal for a drug-response prediction model, closely mirroring the challenges of the graph isomorphism problem. Given that this problem is categorized as NP-complete, scholars have crafted diverse strategies to estimate a solution, with some specifically tailored for molecular structures. In this

study, we delved into several molecular representations to pinpoint the most fitting one for our drug prediction framework. These chosen molecular representations, showcased in Fig. 2, will be detailed further in the upcoming subsections.

Fingerprints are a subset of molecular representation techniques that capture the unique substructures within various molecules. Crafting the RDKit Daylight-like fingerprint entails a three-step procedure. Initially, the molecule is divided into several subgraphs based on different bond path lengths, ranging from a minimum of 1 to a maximum of 7 in this study. Post fragmentation, each subgraph is allocated a unique raw ID through a hash function. Subsequently, a modulo operation is carried out on these raw IDs, and the derived remainders, symbolizing the fingerprint IDs, are translated into bit vectors to formulate the fingerprint. In this study, the RDKit Daylight-like fingerprint was adapted to the molecules using the "RDKFingerprint" function from the rdkit module (v2022.03.5) in Python (version 3.6.10), with a default fingerprint size of 2048 bits.

5.3. From SMILES strings to neural embeddings: expanding the horizons of molecular representation

The SMILES notation of a molecule can be used to reconstruct its corresponding molecular graph, which can then be analyzed as a graphical problem. The Weisfeiler–Lehman graph isomorphism test (WL test) stands out as a robust method for distinguishing between different graphs [51]. In the WL test, each node within a graph is endowed with a distinct label. The process then cyclically accumulates the labels from its neighboring nodes to create a multiset, which is then hashed into fresh distinct labels. By contrasting the label sets from two graphs, one can ascertain whether they're isomorphic. GNNs are a subset of message-passing neural networks (MPNNs) and can be segmented into three primary components: AGGREGATE, COMBINE, and READOUT. Given a graph $G = (V,E)$, where each node $v \in V$ has a feature X_v, GNNs systematically refresh a node's feature by considering its neighboring nodes. After k iterations, the node features capture the structural information within its k-hop neighborhood. The GNN's k-th layer can be articulated as follows:

$$a_v^{(k)} = AGGREGATE^{(k)}\left(\left\{h_u^{(k-1)}, \quad u \in N(v)\right\}\right)$$

and

$$h_v^{(k)} = COMBINE^{(k)}\left(h_v^{(k-1)}, \quad a_v^{(k-1)}\right), \tag{2}$$

where $h_v^{(k)}$ is the node feature of node v at the k-th iteration, which aggregates the node features of its neighbors and itself from the $(k-1)$-th iteration. If a representation of the entire graph is needed, then the READOUT layer is applied after the last GNN iteration:

$$h_G = READOUT\left(\left\{h_v^{(k)} \middle| v \in G\right\}\right). \tag{3}$$

The power of the isomorphism test of GNNs is comparable to that of the WL test [52]. MolecularGNN_smiles [39] operates as an r-th order WL (r-WL) GNN. Initially, it translates the SMILES string of a compound into substructural fingerprints via the r-WL algorithm. Each fingerprint subsequently undergoes conversion to embeddings. The ensuing GNN then ingests these embeddings along with a graph's adjacency matrix to produce predictions for the property in focus. In our analysis, we adopted the 1-WL GNN as a key molecular representation technique.

The SMILES notation of a molecule can be perceived as a sequence of characters, akin to a word string, making it suitable for analysis using NLP techniques. SMILESVec, a word2vec skip-gram model introduced by Ozturk *et al.* [25] and based on a foundational design [53], produces word-level molecular embeddings by training on a comprehensive molecular database. To formulate these embeddings, the SMILES string is initially split into specific subsequences or "words", determined by a set word size. The skip-gram model then anticipates a collection of contextual words based on an input word paired with a specified window size. The resulting SMILESVec embedding for a molecule is computed by summing the vectors of its individual words and then dividing by the total word count, as outlined below:

$$SMILESVec = \frac{\sum_{k=1}^{n} vector(subsequence_k)}{n}. \tag{4}$$

In this study, we utilized a pre-trained SMILESVec model, which underwent training using the PubChem database, encompassing about 2.3 million molecules. This model employed a word size of 8 and a window size of 20, producing embeddings with a dimensionality of 100. The file containing this pre-trained SMILEVec model, named "drug.l8.pubchem.canon.ws20", is accessible for download at https://cmpe.boun.edu.tr/~hakime.ozturk/smilesvec.html.

5.4. Deciphering molecular similarity and gene expression profiling: advanced methods for drug discovery

Using the aforementioned molecular representations, we determined molecular similarity. Numerous similarity coefficients can be employed to gauge the resemblance between two distinct molecules. Among these, the Tanimoto coefficient (Tc) stands out as a frequently adopted metric for fingerprint-based similarity measurements, essentially serving as an extended version of the Jaccard index. The equation for the Tanimoto coefficient is presented as follows:

$$Tc\left(fp_i, \quad fp_j\right) = \frac{fp_i \bullet fp_j}{||fp_i||^2 + \left||fp_j|\right|^2 - fp_i \bullet fp_j}. \tag{5}$$

The cosine similarity is another popular coefficient, gauging the likeness between two vectors within an inner product space. The equation representing cosine similarity is detailed below:

$$Cos\left(fp_i, \quad fp_j\right) = \frac{fp_i \bullet fp_j}{||fp_i||^2 \times \left||fp_j|\right|^2}. \tag{6}$$

In this study, we evaluated molecular similarity by applying the Tanimoto similarity to the RDKit Daylight-like fingerprint and using cosine similarity for other representations to contrast the likeness among molecules. Additionally, the molecular similarity measure derived from the RDKit Daylight-like fingerprint was incorporated as one of the molecular representation methods in this investigation.

In this study, we examined the gene expression profiles from the CCLE dataset (version 21Q3), comprising 1377 cell lines and 19,177 genes. To pinpoint pertinent features, we adopted the methodology delineated by Hsieh *et al.* [40] for assessing the gene expression profile. The step-by-step feature selection is depicted in Fig. 1**(C)**. Given that the PRISM dataset encompassed three screening batches (namely HTS002, MTS006, and MTS010), our initial step was to ascertain the average AUC across these batches. For every gene within the CCLE dataset and each molecule in the PRISM dataset, we gauged the correlation between the gene expression variance across different cell lines and the batch-averaged AUCs for different cell lines corresponding to the molecule. Subsequently, 8087 genes, which showcased absolute PCCs exceeding 0.2 in a minimum of ten molecules, were retained as feature genes. For the next phase, to derive the kernel feature incorporated in the PRISM model of scDrug, we began by determining the log2 expression fold-change based on the mean expression profile spanning cell lines and factoring in the above-mentioned feature genes. This set the stage for computing the similarity between cell lines, grounded on their fold-change profiles, utilizing PCCs.

5.5. Harnessing molecular representations and machine learning for advanced drug response predictions from RDKit to GNNs

In this study, we harnessed 1-AUC as our prediction metric, which represents the area situated above the dose-response curve, commonly termed as the AAC. As showcased in our workflow (refer to Fig. 1**(A)**), we employed an array of molecular representations and computational methodologies to predict how molecules respond across diverse cell lines, leveraging both the training and drug-blind testing sets. The most proficient model was subsequently integrated with the PRISM model from scDrug, resulting in our definitive type-iii drug-response prediction model. The subsequent sections delve deeper into the models we engaged in this study. Fig. 2 graphically presents the molecular representations and computational approaches utilized here. For a more granular perspective, the specific hyperparameters for the discussed models can be found in Supplementary Table S5.

In the initial phase, molecules from the PRISM dataset were transformed into RDKit Daylight-like fingerprints, each having a 2,048-dimensional space. Additionally, the Tanimoto similarity for these fingerprints was determined. For predicting drug responses, we leveraged ML models like random forest (RF) [54] and SVMs [55]. The RF and SVM implementations were facilitated using the "ensemble.RandomForestRegressor" and "svm" functions from the sklearn module (version 0.24.2) in Python (version 3.6.10), employing default parameters. Both the RDKit fingerprints and the Tanimoto similarity metric served as input for the RF and SVM models. We explored four ML model combinations: SVM using RDKit fingerprints, SVM utilizing the similarity metric, RF with RDKit fingerprints, and RF leveraging the similarity metric. The efficacy of these configurations was evaluated, and the most efficient model was selected for subsequent comparison with other model variants.

In this study, every molecule from the PRISM dataset was processed using the 1-WL algorithm. The ensuing fingerprint dictionary had a dimension of 420. It's crucial to highlight that this dictionary's size is contingent upon the 1-hop substructures of the provided molecules. These fingerprints were subsequently converted into 50-dimensional embeddings utilizing the "nn.embedding" function from the PyTorch module (version 1.12.1+cu102) in Python. The molecular graph's embeddings and adjacency matrix then served as inputs for a GNN model composed of six layers. This GNN model utilized the rectified linear unit (ReLU) function, symbolized as σ, as its activation function, with the global pooling operation executed via summation. The vertex transition within the GNN can be articulated as:

$$v_i^{(t+1)} = \sigma\left(v_i^{(t)} + \sum_{i \in N(i)} h_{ij}^{(t)}\right) \tag{7}$$

and

$$h_{ij}^{(t)} = f\left(W_{neighbor}\begin{bmatrix} v_j^{(t)} \\ e_{ij}^{(t)} \end{bmatrix} + b_{neighbor}\right) \tag{8}$$

and the edge transition steps in GNN are described by:

$$e_{ij}^{(t+1)} = \sigma\left(e_{ij}^{(t)} + g_{ij}^{(t)}\right) \tag{9}$$

and

$$g_{ij}^{(t)} = f\left(W_{side}\left(v_i^{(t)} + v_j^{(t)}\right) + b_{side}\right). \tag{10}$$

In these equations, $v_i^{(t)}$ represents the i-th vertex embedding at time step t, $h_{ij}^{(t)}$ denotes the hidden neighborhood vector derived from the vertex and edge embeddings for the neighbors of vertex i, $e_{ij}^{(t)}$ represents the edge embedding between the i-th and j-th vertices at time step t, $g_{ij}^{(t)}$ indicates the hidden neighborhood vector derived from the vertex embeddings for the neighbors of vertex i, and f represents the element-wise sigmoid function. The GNN model's output was subsequently channeled through a six-layer multilayer perceptron (MLP) to produce the predicted drug response.

5.6. Enhancing drug response predictions: integrating molecular embeddings and modifying CaDRReS matrix factorization approaches

CaDRReS [38] is an MF model that incorporates information of gene expression similarity of cell lines, calculated using Pearson's correlation. The CaDRReS model can be represented by:

$$\widehat{S}_{ui} = \mu + b_i^Q + b_u^P + q_i \bullet p_u = \mu + b_i^Q + b_u^P + \quad q_i\left(x_u W_p\right)^T, \tag{11}$$

where μ represents the average drug response; b_i^Q and b_u^P are the bias terms for drug i and cell line u, respectively; and q_i and p_u where $q_i,\ p_u \in \mathbb{R}^f$ denote the latent vectors for drug i and cell line u, respectively. The transformation matrix $W_p \in \mathbb{R}^{d \times f}$ projects the feature vectors $x_u \in \mathbb{R}^d$, which represent the cell line similarity vector, onto the latent space. However, in this study, we deviated from the original approach by employing the molecular similarity metric instead of the cell line similarity metric. The molecular similarity was computed from RDKit Daylight-like fingerprints using Tanimoto similarity. The adapted model can be represented as:

$$\widehat{S}_{ui} = \mu + b_i^Q + b_u^P + q_i \bullet p_u = \mu + b_i^Q + b_u^P + \left(x_i W_q\right) p_u, \tag{12}$$

where $W_q \in \mathbb{R}^{d \times f}$ is the transformation matrix that projects the feature vectors $x_i \in \mathbb{R}^d$, representing the molecular similarity vector of drug I, onto the latent space. It is worth noting that we used the model without bias terms and scaled the AAC to $AAC \times 240 - 80$, as in scDrug, to make the prediction and achieve improved performance. The concluding model under assessment employed the pre-established SMILESVec model [53] for molecular encoding, yielding molecular embeddings with a 100-dimensional span. These embeddings were subsequently fed into a 6-layer MLP to facilitate predictions. Beyond the ML models, every other model we explored had multi-output capabilities, allowing them to forecast the drug responses of all training cell lines for a designated test molecule in one predictive sweep. Across all these models, the mean squared error (MSE) served as the consistent loss function.

5.7. Comparative evaluation methods for drug response: cell-wise vs. drug-wise metrics

To evaluate the performance of the above-mentioned models, we employed two distinct perspectives: cell-wise evaluation and drug-wise evaluation. Drug-wise assessment focuses on the model's performance when tested across various cell lines but using a consistent drug. Conversely, cell-wise assessment measures the model's effectiveness when tested across a range of drugs but targeting a specific cell line. Leveraging these two evaluation techniques ensures a comprehensive understanding of the model's predictive capability for individual drugs and distinct cell lines.

In this study, predicting drug response was approached as a regression task. As a result, prevalent regression evaluation metrics, including PCC, Spearman's rank correlation coefficient (SCC), and MSE, were adopted to appraise the prediction accuracy. The respective formulas for these metrics are provided below:

$$PCC = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(\widehat{y}_i - \overline{\widehat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}\sqrt{\sum_{i=1}^{n}(\widehat{y}_i - \overline{\widehat{y}})^2}}, \tag{13}$$

$$SCC = \frac{\sum_{i=1}^{n}(R(y_i) - \overline{R(y)})(R(\widehat{y}_i) - \overline{R(\widehat{y})})}{\sqrt{\sum_{i=1}^{n}(R(y_i) - \overline{R(y)})^2}\sqrt{\sum_{i=1}^{n}(R(y_i) - \overline{R(\widehat{y})})^2}} \quad , \tag{14}$$

and

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{15}$$

where y_i and $\hat{y}_i$ represent the ground truth and predicted values, respectively, of the AAC value of molecule i, and $R(y_i)$ and $R(\hat{y}_i)$ denote the rank of the ground truth and predicted AAC value of molecule I, respectively. If a model has multiple outputs, then the PCC and MSE are replaced by the masked average PCC (masked aCC) and the masked MSE to address the problem of missing values. The formulas for masked aCC and masked MSE are as follows:

$$masked\_aCC = \sum_d \frac{\sum_c \left(Y_{dc} - \frac{Y_{dc}\odot M_{dc}}{\sum_{d,\ c} M_{dc}}\right)\left(\hat{Y}_{dc} - \frac{\hat{Y}_{dc}\odot M_{dc}}{\sum_{d,\ c} M_{dc}}\right)}{\sqrt{\sum_c \left(Y_{dc} - \frac{Y_{dc}\odot M_{dc}}{\sum_{d,\ c} M_{dc}}\right)^2 \sum_c \left(\hat{Y}_{dc} - \frac{\hat{Y}_{dc}\odot M_{dc}}{\sum_{d,\ c} M_{dc}}\right)^2}} \tag{16}$$

and

$$masked\_aCC = \sum_d \frac{\sum_c \left(Y_{dc} - \frac{Y_{dc}\odot M_{dc}}{\sum_{d,\ c} M_{dc}}\right)\left(\hat{Y}_{dc} - \frac{\hat{Y}_{dc}\odot M_{dc}}{\sum_{d,\ c} M_{dc}}\right)}{\sqrt{\sum_c \left(Y_{dc} - \frac{Y_{dc}\odot M_{dc}}{\sum_{d,\ c} M_{dc}}\right)^2 \sum_c \left(\hat{Y}_{dc} - \frac{\hat{Y}_{dc}\odot M_{dc}}{\sum_{d,\ c} M_{dc}}\right)^2}}, \tag{17}$$

where Y_{dc} and $\hat{Y}_{dc}$ represent the ground truth and prediction matrix of AAC values for d molecules in c cell lines, respectively. The mask matrix M_{dc} consists of 1 s and 0 s, where 0 represents the absence of drug response in the data and 1 otherwise.

Beyond the regression evaluation metrics, we also utilized classification metrics like recall, precision, and the F1 score to gauge the model's practical relevance. The AUC values were categorized into three distinct classes: "inactive," "unclear," and "potential." A molecule was labeled "inactive" if its AUC value surpassed 0.8, aligning with the inactivity threshold delineated by Corsello *et al.* [36]. The demarcation between "unclear" and "potential" was established at 0.4. The formulas to calculate recall, precision, and the F1 score are detailed below:

$$Recall = \frac{TP}{(TP + FN)}, \tag{18}$$

$$Precision = \frac{TP}{(TP + FP)}, \tag{19}$$

and

$$F1score = \frac{2 \times Precision \times Recall}{(Precision + Recall)}, \tag{20}$$

where TP, TN, FP, and FN represent the number of true positive, true negative, false positive, and false negative cases, respectively.

Funding

This work was financially supported by the National Science and Technology Council (NSTC 109–2221-E-002–161-MY3, NSTC 109–2221-E-010–011-MY3, 111–2327-B-002–017, NSTC 112–2221-E-A49–061-MY3 and NSTC 112–2321-B-002–013), Research Proposal for NTU Core Consortiums (NTU-113L8503) and the Center for Advanced Computing and Imaging in Biomedicine (NTU-113L900701) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

CRediT authorship contribution statement

Chiao-Yu Hsieh: Software, Data curation. **Yih-Yun Sun:** Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Tzu-Yang Tseng:** Validation. **Jian-Hung Wen:** Methodology, Data curation. **Jia-Hsin Huang:** Resources. **Hsuan-Cheng Huang:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Yen-Jen Oyang:** Methodology. **Hsueh-Fen Juan:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Data availability

https://github.com/ailabstw/scDrugplus

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.biopha.2024.117070.

References

[1] D.J. Duffy, Problems, challenges and promises: perspectives on precision medicine, Brief. Bioinform. 17 (2016) 494–504.
[2] R. Sager, Expression genetics in cancer: shifting the focus from DNA to RNA, Proc. Natl. Acad. Sci. USA 94 (1997) 952–955.
[3] N. Zhang, H. Wang, Y. Fang, J. Wang, X. Zheng, X.S. Liu, Predicting Anticancer Drug Responses Using a Dual-Layer Integrated Cell Line-Drug Network Model, PLoS Comput. Biol. 11 (2015) e1004498.
[4] Y. Wang, D. Zheng, The importance of precision medicine in modern molecular oncology, Clin. Genet 100 (2021) 248–257.
[5] G. Gambardella, G. Viscido, B. Tumaini, A. Isacchi, R. Bosotti, D. di Bernardo, A single-cell analysis of breast cancer cell lines to study tumour heterogeneity and drug response, Nat. Commun. 13 (2022) 1714.
[6] F. Feng, B. Shen, X. Mou, Y. Li, H. Li, Large-scale pharmacogenomic studies and drug response prediction for personalized cancer medicine, J. Genet Genom. 48 (2021) 540–551.
[7] S. Chawla, A. Rockstroh, M. Lehman, E. Ratther, A. Jain, A. Anand, A. Gupta, N. Bhattacharya, S. Poonia, P. Rai, et al., Gene expression based inference of cancer drug sensitivity, Nat. Commun. 13 (2022) 5680.
[8] S. Kim, S. Bae, Y. Piao, K. Jo, Graph convolutional network for drug response prediction using gene expression data, Mathematics 9 (2021) 772.
[9] K. Qiu, J. Lee, H. Kim, S. Yoon, K. Kang, Machine learning based anti-cancer drug response prediction and search for predictor genes using cancer cell line gene expression, Genom. Inf. 19 (2021) e10.
[10] L. Wang, X. Li, L. Zhang, Q. Gao, Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization, BMC Cancer 17 (2017) 513.
[11] M. Fallahi-Sichani, S. Honarnejad, L.M. Heiser, J.W. Gray, P.K. Sorger, Metrics other than potency reveal systematic variation in responses to cancer drugs, Nat. Chem. Biol. 9 (2013) 708–714.
[12] N. Pozdeyev, M. Yoo, R. Mackie, R.E. Schweppe, A.C. Tan, B.R. Haugen, Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies, Oncotarget 7 (2016) 51619–51625.
[13] D.S. Wigh, J.M. Goodman, A.A. Lapkin, A review of molecular representation in the age of machine learning, WIREs Comput. Mol. Sci. 12 (2022) e1603.
[14] P.G. Dittmar, N.A. Farmer, W. Fisanick, R.C. Haines, J. Mockus, The CAS ONLINE search system. 1. General system design and selection, generation, and use of search screens, J. Chem. Inf. Comput. Sci. 23 (1983) 93–102.
[15] E.W. Sayers, E.E. Bolton, J.R. Brister, K. Canese, J. Chan, D.C. Comeau, C. M. Farrell, M. Feldgarden, A.M. Fine, K. Funk, et al., Database resources of the National Center for Biotechnology Information in 2023, Nucleic Acids Res 51 (2023) D29–D38.
[16] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1988) 31–36.
[17] S.R. Heller, A. McNaught, I. Pletnev, S. Stein, D. Tchekhovskoi, InChI, the IUPAC International Chemical Identifier, J. Chemin.-. 7 (2015) 23.
[18] N. Schneider, R.A. Sayle, G.A. Landrum, Get your atoms in order–an open-source implementation of a novel and robust molecular canonicalization algorithm. J. Chem. Inf. Model 55 (2015) 2111–2120.

[19] P. Carracedo-Reboredo, J. Linares-Blanco, N. Rodriguez-Fernandez, F. Cedron, F. J. Novoa, A. Carballal, V. Maojo, A. Pazos, C. Fernandez-Lozano, A review on machine learning approaches and trends in drug discovery, Comput. Struct. Biotechnol. J. 19 (2021) 4538–4558.
[20] R.E. Carhart, D.H. Smith, R. Venkataraghavan, Atom pairs as molecular features in structure-activity studies: definition and applications, J. Chem. Inf. Comput. Sci. 25 (1985) 64–73.
[21] G. Landrum, 2016, RDKit: Open-Source Cheminformatics Software. 2016..
[22] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (2010) 742–754.
[23] S. Riniker, G.A. Landrum, Open-source platform to benchmark fingerprints for ligand-based virtual screening, J. Chemin.-. 5 (2013) 26.
[24] Goh G.B., Hodas N.O., Siegel C., Vishnu A.: SMILES2Vec: An Interpretable General-Purpose Deep Neural Network for Predicting Chemical Properties. pp. arXiv:1712.02034; 2017:arXiv:1712.02034.
[25] H. Ozturk, E. Ozkirimli, A. Ozgur, A novel methodology on distributed representations of proteins using their interacting ligands, Bioinformatics 34 (2018) i295–i303.
[26] Bongini P., Bianchini M., Scarselli F.: Molecular graph generation with Graph Neural Networks. pp. arXiv:2012.07397; 2020:arXiv:2012.07397.
[27] J.J. Irwin, K.G. Tang, J. Young, C. Dandarchuluun, B.R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield, R.A. Sayle, ZINC20-A Free Ultralarge-Scale Chemical Database for Ligand Discovery, J. Chem. Inf. Model 60 (2020) 6065–6073.
[28] R. Ramakrishnan, P.O. Dral, M. Rupp, O.A. von Lilienfeld, Quantum chemistry structures and properties of 134 kilo molecules, Sci. Data 1 (2014) 140022.
[29] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B.A. Shoemaker, P. A. Thiessen, B. Yu, et al., PubChem 2023 update, Nucleic Acids Res 51 (2023) D1373–D1380.
[30] D. Mendez, A. Gaulton, A.P. Bento, J. Chambers, M. De Veij, E. Felix, M. P. Magarinos, J.F. Mosquera, P. Mutowo, M. Nowotka, et al., ChEMBL: towards direct deposition of bioassay data, Nucleic Acids Res 47 (2019) D930–D940.
[31] S. Axelrod, R. Gomez-Bombarelli, GEOM, energy-annotated molecular conformations for property prediction and molecular generation, Sci. Data 9 (2022) 185.
[32] Xia J., Zhu Y., Du Y., Li S.Z.: A Systematic Survey of Molecular Pre-trained Models. pp. arXiv:2210.16484; 2022:arXiv:2210.16484.
[33] K. Tomczak, P. Czerwinska, M. Wiznerowicz, The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge, Conte Oncol. (Pozn.) 19 (2015). A68-77.
[34] W. Yang, J. Soares, P. Greninger, E.J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J.A. Smith, I.R. Thompson, et al., Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells, Nucleic Acids Res 41 (2013) D955–D961.
[35] M.G. Rees, B. Seashore-Ludlow, J.H. Cheah, D.J. Adams, E.V. Price, S. Gill, S. Javaid, M.E. Coletti, V.L. Jones, N.E. Bodycombe, et al., Correlating chemical sensitivity and basal gene expression reveals mechanism of action, Nat. Chem. Biol. 12 (2016) 109–116.
[36] S.M. Corsello, R.T. Nagari, R.D. Spangler, J. Rossen, M. Kocak, J.G. Bryan, R. Humeidi, D. Peck, X. Wu, A.A. Tang, et al., Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. Nat. Cancer 1 (2020) 235–248.
[37] P. Geeleher, N.J. Cox, R.S. Huang, Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines, Genome Biol. 15 (2014) R47.
[38] C. Suphavilai, D. Bertrand, N. Nagarajan, Predicting Cancer Drug Response using a Recommender System, Bioinformatics 34 (2018) 3907–3914.
[39] M. Tsubaki, K. Tomii, J. Sese, Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences, Bioinformatics 35 (2019) 309–318.
[40] C.Y. Hsieh, J.H. Wen, S.M. Lin, T.Y. Tseng, J.H. Huang, H.C. Huang, H.F. Juan, scDrug: From single-cell RNA-seq to drug response prediction, Comput. Struct. Biotechnol. J. 21 (2023) 150–157.
[41] D. van Tilborg, A. Alenicheva, F. Grisoni, Exposing the Limitations of Molecular Machine Learning with Activity Cliffs, J. Chem. Inf. Model 62 (2022) 5938–5951.
[42] D. Baptista, J. Correia, B. Pereira, M. Rocha, Evaluating molecular representations in machine learning models for drug response prediction and interpretability, J. Integr. Bioinform. 19 (2022).
[43] X. An, X. Chen, D. Yi, H. Li, Y. Guan, Representation of molecules for drug response prediction, Brief. Bioinform. 23 (2022).
[44] F. Yassaee Meybodi, C. Eslahchi, Predicting anti-cancer drug response by finding optimal subset of drugs, Bioinformatics 37 (2021) 4509–4516.
[45] A. Emdadi, C. Eslahchi, Clinical drug response prediction from preclinical cancer cell lines by logistic matrix factorization approach, J. Bioinform. Comput. Biol. 20 (2022) 2150035.
[46] F. Ahmadi Moughari, C. Eslahchi, ADRML: anticancer drug response prediction using manifold learning, Sci. Rep. 10 (2020) 14245.
[47] R. Masumshah, C. Eslahchi, DPSP: a multimodal deep learning framework for polypharmacy side effects prediction, Bioinform. Adv. 3 (2023) vbad110.
[48] R. Masumshah, R. Aghdam, C. Eslahchi, A neural network-based method for polypharmacy side effects prediction, BMC Bioinforma. 22 (2021) 385.
[49] R. Qureshi, S.A. Basit, J.A. Shamsi, X. Fan, M. Nawaz, H. Yan, T. Alam, Machine learning based personalized drug response prediction for lung cancer patients, Sci. Rep. 12 (2022) 18935.
[50] J.M. McFarland, B.R. Paolella, A. Warren, K. Geiger-Schuller, T. Shibue, M. Rothberg, O. Kuksenko, W.N. Colgan, A. Jones, E. Chambers, et al., Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action, Nat. Commun. 11 (2020) 4296.
[51] B. Weisfeiler, A.A. Lehman, A reduction of a graph to a canonical form and an algebra arising during this reduction, Nauchno-Tech. Inf. 2 (9) (1968) 12–16.
[52] Xu K., Hu W., Leskovec J., Jegelka S.: How Powerful are Graph Neural Networks? pp. arXiv:1810.00826; 2018:arXiv:1810.00826.
[53] Mikolov T., Chen K., Corrado G., Dean J.: Efficient Estimation of Word Representations in Vector Space. pp. arXiv:1301.3781; 2013:arXiv:1301.3781.
[54] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32.
[55] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.