BMC Bioinformatics

RESEARCH Open Access

Application of transfer learning for cancer drug sensitivity prediction

Saugato Rahman Dhruba[1], Raziur Rahman[1], Kevin Matlock[1], Souparno Ghosh[2] and Ranadip Pal[1*]

From The International Conference on Intelligent Biology and Medicine (ICIBM) 2018
Los Angeles, CA, USA. 10-12 June 2018

Abstract

Background: In precision medicine, scarcity of suitable biological data often hinders the design of an appropriate predictive model. In this regard, large scale pharmacogenomics studies, like CCLE and GDSC hold the promise to mitigate the issue. However, one cannot directly employ data from multiple sources together due to the existing distribution shift in data. One way to solve this problem is to utilize the transfer learning methodologies tailored to fit in this specific context.

Results: In this paper, we present two novel approaches for incorporating information from a secondary database for improving the prediction in a target database. The first approach is based on latent variable cost optimization and the second approach considers polynomial mapping between the two databases. Utilizing CCLE and GDSC databases, we illustrate that the proposed approaches accomplish a better prediction of drug sensitivities for different scenarios as compared to the existing approaches.

Conclusion: We have compared the performance of the proposed predictive models with database-specific individual models as well as existing transfer learning approaches. We note that our proposed approaches exhibit superior performance compared to the abovementioned alternative techniques for predicting sensitivity for different anti-cancer compounds, particularly the nonlinear mapping model shows the best overall performance.

Keywords: Drug sensitivity prediction, Pharmacogenomic studies, CCLE, GDSC, Transfer learning, Nonlinear mapping, Latent variable, Cost optimization

Background

A consistent challenge in precision medicine is to design appropriate models for predicting the sensitivity of a tumor to an anti-cancer compound with high accuracy. In this aspect, large-scale pharmacogenomic studies of cancer genomes have provided unprecedented insights for studying anti-cancer therapeutics to determine putative prediction of drug sensitivity. The Genomics of Drug Sensitivity in Cancer (GDSC) [1] of the Cancer Genome Project and the Cancer Cell Line Encyclopedia (CCLE) [2] from the Broad Institute are two such studies where drug sensitivity profiles and genomic information across hundreds of compounds and cancer cell lines have been systematically gathered. There exists significant overlaps between the two databases which can further be utilized in designing more accurate sensitivity predictive models. Biological data for designing suitable predictive models are frequently scarce and therefore the availability of a secondary dataset often holds the promise for a better model development. However, majority of the machine learning approaches used in drug sensitivity prediction follow the inherent assumption that both training data and test data are in the same feature space with the same distribution. But, when training and test data, despite being in the same feature space, exhibit different distributions, one need to take the distribution shift into account. This is where transfer learning (TL) methodologies come into play [3].

*Correspondence: ranadip.pal@ttu.edu
[1] Department of Electrical and Computer Engineering, Texas Tech University, 1012 Boston Ave, 79409 Lubbock, TX, USA
Full list of author information is available at the end of the article

Often in TL environment, the source and target domains can be considered as linked subspaces as part of a high-level common domain space [4]. We, therefore need to assume that there exists some consistency between the different datasets to be utilized in TL. Haibe-Kains et al. [5] at first pointed out that, although the gene expression from CCLE and GDSC databases are well correlated between themselves, unexpectedly the measured pharmacological drug responses using common estimators such as IC_{50} and the area under the curve (AUC) measures are highly discordant. In response, the CCLE and GDSC investigators performed their own analysis [6] and presented results opposing the conclusions in [5]. They pointed out that in majority of the drugs, the exhibited AUC and IC_{50} distributions are dominated by drug insensitive lines with a much smaller number of outliers, and postulated that the differences in cell line biology between studies have resulted in the poor correlation. Considering these facts, they have demonstrated significant improvement in correlation between most of the drugs. In any event, the fact that both the databases are providing information about the same biological process, make them suitable candidates for applying transfer learning methodologies.

In case of inconsistent data with different distributions for training and test sets, various TL approaches [3] have been attempted for dataset shift. Unsupervised methods such as INSPIRE (INferring Shared modules from multiPle gene expREssion datasets) [7] is primarily focused on the expression datasets to extract a low-dimensional representation and predicts tumor phenotypes using regularized regression approaches. Inductive transfer learning (ITL) approaches, as in [8], tackle the issue of prediction for scarce primary data using a secondary dataset through importance sampling *i.e.*, reweighting the secondary distribution to the primary. While the primary data size is assumed to be significantly smaller than secondary data, for large number of unlabeled data, one has to adapt to covariate shift along with ITL. Boosting based approaches such as Dynamic-TrAdaBoost [9] applies ensemble methods to both source and target instances and then employs an update mechanism incorporating only the source instances useful for target task, with an additional dynamic correction factor. Kernel based ITL methods [10, 11] focus on finding an appropriate kernel for the newly available data, modeling the difference with existing data as a problem of finding the suitable bias.

The previous approaches for transfer learning work well under the assumption that the datasets are closely related (such as 9 ovarian cancer datasets in INSPIRE) and the number of samples are significantly larger than the number of features ($n > p$). However, the scenario is frequently reversed in the case of genomic (or proteomic) data *i.e.*, we usually have tens of thousands of genes and a small number of cell lines. Additionally, the previous methods for TL often involve removing the distribution shift via weighting without any explicit domain transfer. In our work, we have proposed two different TL approaches that consider mapping the data from two different databases to either a common space or to each other's domain, inherently taking care of the $n << p$ problem. The inherent assumptions here for each pair of similar datasets from CCLE and GDSC are – (i) The datasets are monotonically changing in the same direction, and (ii) There exists a functional relationship between them. To build an appropriate prediction model, we utilize the gene expression as the predictors and the drug sensitivity (specifically AUC) as the output. Considering the application of TL on these datasets, the proposed approaches in this paper can be classified into two categories, as illustrated in Fig. 1.

- Cost optimization based approach where we employ latent variable models to extract the underlying variables between different datasets. In this case, TL can be applied to only the output (Fig. 1(a)), as in parameter transfer approach [12, 13] or to both model input and output (Fig. 1(b)), as in [14, 15].
- Domain transfer approach where we design maps between databases to transfer data from primary domain to secondary and utilize the secondary data to improve the prediction model. Here, TL is applied to both input and output (Fig. 1(c)), as in instance transfer approach [14, 15].

To summarize, the key contributions of the paper is – we have implemented two TL based approach, where the target (primary) data is either transferred to a common latent variable space along with the source (secondary) data, or to the source domain through nonlinear mapping to improve the prediction of limited primary data employing the available secondary data.

Results

To evaluate the performance of our transfer learning algorithms, we have initially retrieved the data common to both CCLE and GDSC. From GDSC ($v6.0$) and CCLE, there are 15,664 common genes available in 623 common cell lines along with 15 common drugs. We have performed a drug-wise analysis and found that the number of cell lines decreases from 623 after incorporating the available drug sensitivity values, resulting in datasets with cell lines between $91 - 310$ along with 15,664 genes and corresponding sensitivity measures. For analysis involving gene expression, we have used ReliefF [16] to select the top 200 genes from each dataset and taken the intersection as the final feature set. For drug sensitivity measure, we have used the AUC values as they have more concordance between databases (median $\rho_s = 0.34$) than

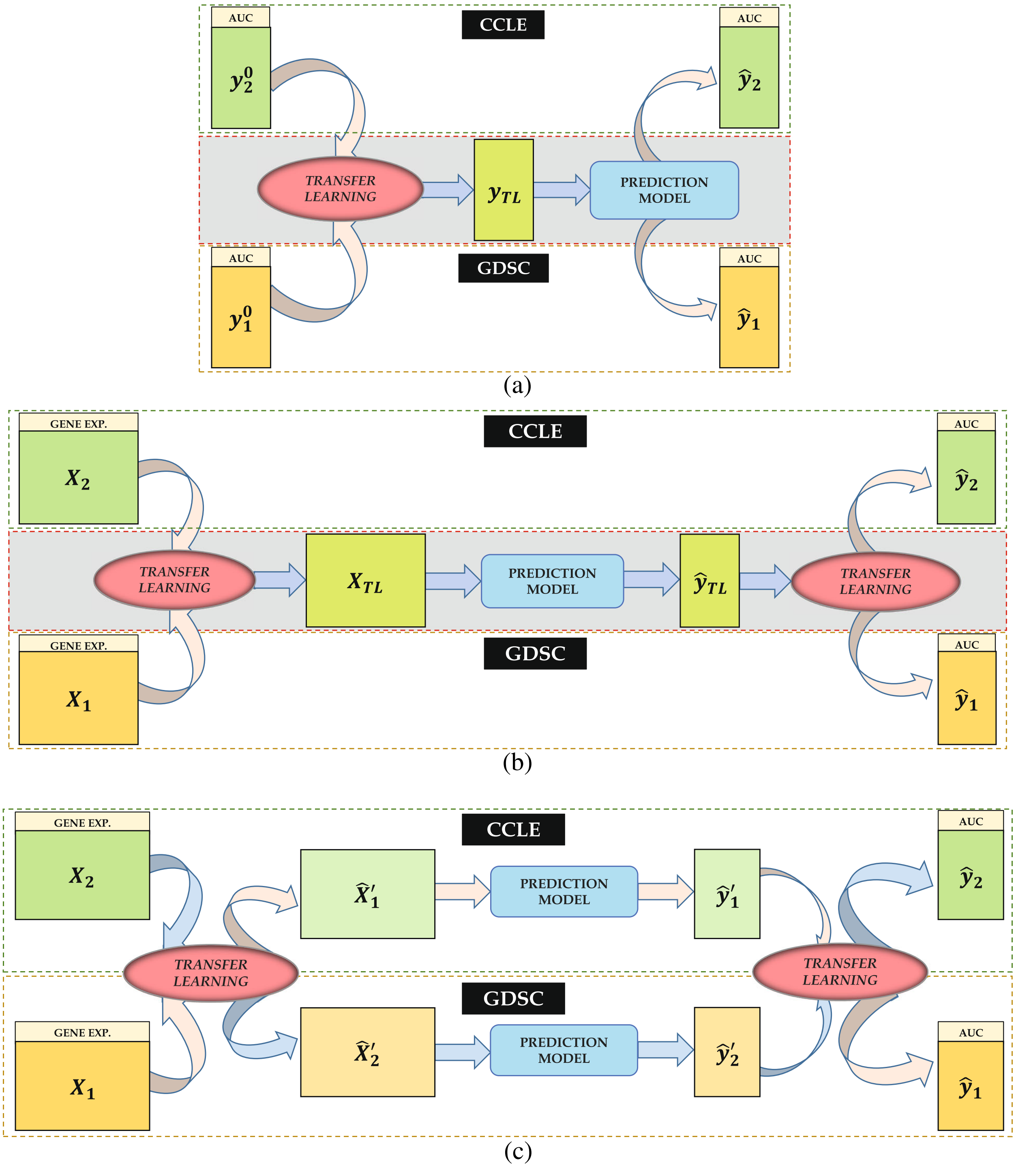

Fig. 1 Illustration of the proposed transfer learning approaches: (**a**) Optimization based TL, applied on output only, (**b**) Optimization based TL, applied on both input and output, and (**c**) Domain transfer TL, applied on both input and output

IC_{50} (median $\rho_s = 0.28$) [5]. Note that in spite of our discussion on inconsistencies between databases, the main goal here is to consider the scenario where a small portion of database 1 (*i.e.*, GDSC) is available while data for the entire database 2 (*i.e.*, CCLE) is available and we would like to use database 2 to improve the prediction performance for the rest of database 1. Thus, for evaluation, we will use the GDSC experimental AUCs as the *gold standard* and compare with the predicted AUCs.

Latent variable cost optimization approach

We have performed drug sensitivity prediction using the three latent variable cost optimization based approaches – Latent Regression Prediction (LRP), Latent-Latent Prediction (LLP), Combined Latent Prediction (CLP) (described in the "Methods" section) for 7 common drugs with sufficient cell lines ($n > 200$). For each method, subsets of 50 randomly chosen GDSC cell lines (X_{11} & y_{11} in Figs. 2 & 3) are used for the cost optimization in training and the rest (y_{12}) are predicted along with the known CCLE data (X_2 & y_2 in Figs. 2 & 3). Table 1 illustrates the comparison of prediction performance for all three methods with *Direct prediction (DP)* for K-fold cross-validation, where DP is defined as training on the 50 available cell lines and predicting for the rest. Here, the number of folds is found as $K = \frac{n}{50}$, where 1 fold (containing ∼50 samples) is used for training and the remaining $(K - 1)$ folds are used for testing.

Domain transfer approach

We have performed the Mapped Prediction (MP) approach (described in the "Methods" section) for predicting GDSC sensitivities for 7 common drugs with sufficient cell lines ($n > 200$) and different levels of database consistency. Figure 4 demonstrates the effect of first-order polynomial mapping for a representative gene expression set, while Fig. 5 illustrates the effect of second-order polynomial mapping for a representative drug sensitivity vector. Again, we used random subsets of 50 cell lines (G_{11}, d_{11} & G_{21}, d_{21} in Fig. 6) to retrieve the mapping functions and sensitivities for the rest (d_{12}) are predicted using the known CCLE data (G_{22}, d_{22}). Table 2 shows the comparison of prediction performance for MP approach for all 7 drugs with two other methods – Direct Prediction (DP) and *CCLE model prediction (CP)* for K-fold cross-validation, as defined above (*i.e.*, $K = \frac{n}{50}$ and 1 fold is used for training and $(K-1)$ folds for testing). For CP approach, the model is built using the available CCLE data directly and prediction is performed using the GDSC expression data. For prediction of AUC values using gene expression data, we have used a Bias-corrected Random Forest (BC-RF) [17–19] model.

Discussion

From Table 1, it is evident that the CLP method yields the best performance. Additionally, even though the LLP method often yield better results than DP, it frequently underperforms than LRP. Overall, 6 drugs out of 7 yield the best performance for CLP method while only Nilotinib performs the best with LRP. The prediction performance is similar in the reverse direction (*i.e.*, CCLE as the primary set and GDSC as secondary) where 5 out of 7 drugs show best performance for CLP.

For the Domain Transfer approach, it is evident from Table 2 that the MP approach performs significantly better than the both CP and DP. Furthermore, the performance of the CP approach is much worse compared to either MP or DP, which can be attributed to the existing distribution shift between CCLE and GDSC data in general. Note that among the 7 drugs, 17-AAG and PD-0325901 has moderate concordance ($0.5 \le \rho_s < 0.6$) while AZD6244, Nutlin-3 and PD-0332991 have poor concordance ($\rho_s < 0.4$) between databases. For PLX4720 and Nilotinib, there exist moderate to high consistency in terms of Pearson correlation ($\rho = 0.57$ and $\rho = 0.88$ respectively), although the rank correlation is low ($\rho_s = 0.29$ and $\rho_s \approx 0.1$ respectively). We have also implemented a model that uses the ensemble of available CCLE and GDSC data directly

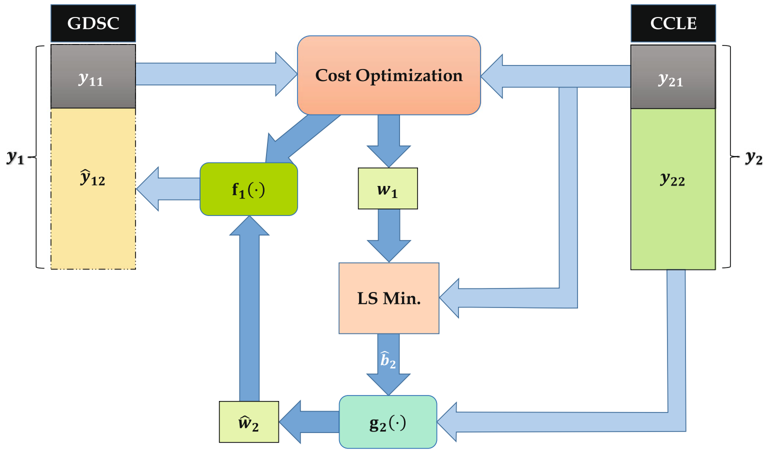

Fig. 2 Illustration of Latent Regression Prediction. Here, unknown set of GDSC AUC values, y_{12}, is predicted using the underlying latent vector, w_2, calculated from corresponding CCLE AUC set, y_{22}

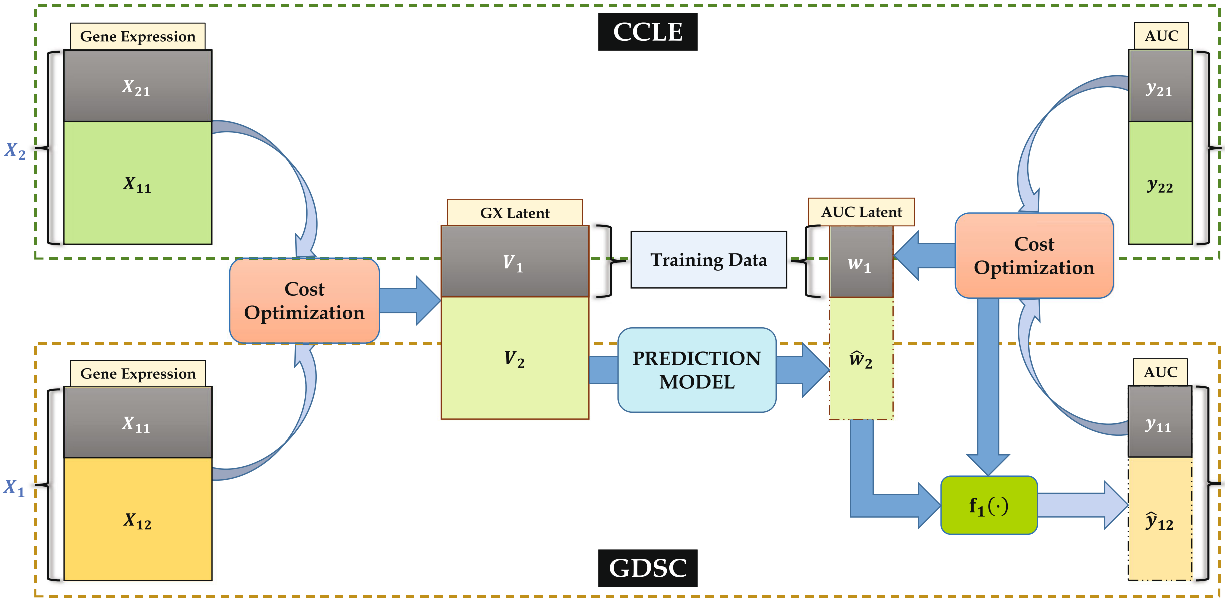

Fig. 3 Illustration of Latent-Latent Prediction. Here, unknown set of GDSC AUC values y_{12} is predicted using the underlying latent variables V and w_1, calculated from X_1, X_2 and y_{11}, y_{21}. V_1 & w_1 are used for training while V_2 is used to predict w_2 and then y_{12}

for training and predicts for the unlabeled GDSC expression data, referred as the *Combined Model Prediction*. An additional section provides a detailed description and comparative analysis of this model with the MP approach [see Additional file 1].

Comparison with inductive transfer learning

We have compared the results from the Mapped Prediction approach with an existing transfer learning approach, namely the *Importance-weighted Direct Inductive Transfer Learning (DITL)* proposed by Garcke et al. [8]. In DITL, the primary and secondary datasets are assumed to be related in a way so that in some parts of the domain, the two distributions can be similar, and therefore, one can employ the secondary dataset with primary via importance sampling (*i.e.*, reweighting the secondary distribution to the primary so that the secondary data points with positive effect on primary data will have greater weights). For prediction, DITL uses weighted Kernel Ridge regression (KRR) with *Gaussian* kernels, dubbing the whole approach as DITL-KRR [8]. Table 3 shows the comparison of prediction performance for DITL-KRR approach with MP and DP approaches for 4 representative drugs. Unlike the MP approach, DITL follows the $n > p$ assumption of machine learning and therefore, we used the intersection of top 50 genes from both datasets as the feature set while 50 cell lines were used for training. From Table 3, we can conclude that MP has a superior performance compared

Table 1 Comparison of *K*-fold cross-validation performance for 4 GDSC drug sensitivity prediction approaches – Latent Regression Prediction (LRP), Latent-Latent Prediction (LLP), Combined Latent Prediction (CLP) and Direct Prediction (DP), using data from CCLE

Drug	Pearson Correlation				NRMSE			
	LRP	LLP	CLP	DP	LRP	LLP	CLP	DP
17-AAG	0.5441	0.4691	**0.6382**	0.4591	0.2117	0.2147	**0.1930**	0.2164
AZD6244	0.3988	0.4155	**0.4524**	0.4008	0.1833	0.1718	**0.1684**	0.1703
Nilotinib	**0.9053**	0.3886	0.8768	0.4524	**0.0728**	0.1295	0.0888	0.1242
Nutlin-3	0.4093	0.5473	**0.5646**	0.5108	0.1965	0.1756	**0.1745**	0.1799
PD-0325901	0.6448	0.4502	**0.6606**	0.4465	0.1614	0.1870	**0.1585**	0.1878
PD-0332991	0.2497	0.0912	**0.2540**	0.0884	0.1695	0.1729	**0.1672**	0.1733
PLX4720	0.5682	0.5040	**0.6384**	0.5001	0.1237	0.1290	**0.1173**	0.1291

Bold values indicate the best performance

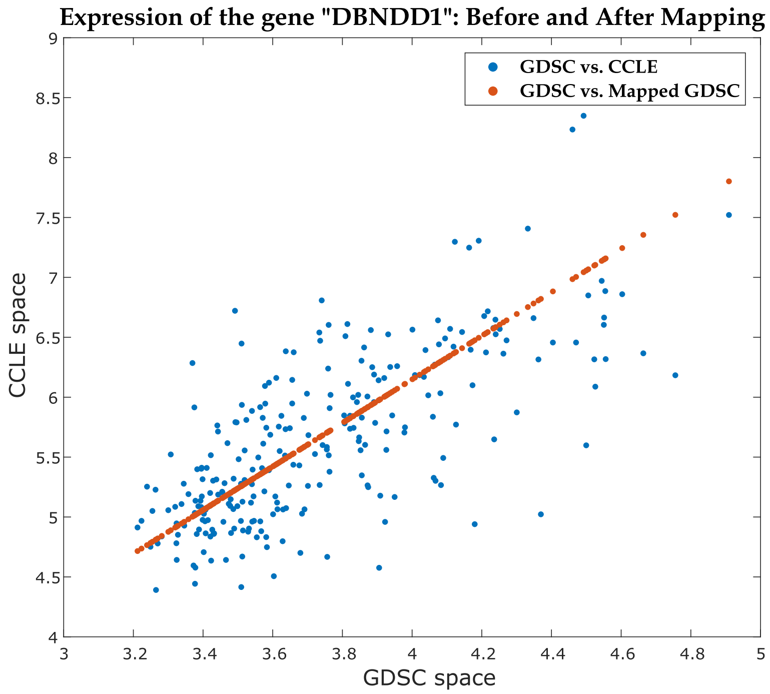

Fig. 4 Scatter plot of gene expression association between GDSC and CCLE spaces before and after applying the polynomial mapping for the gene "DBNDD1"

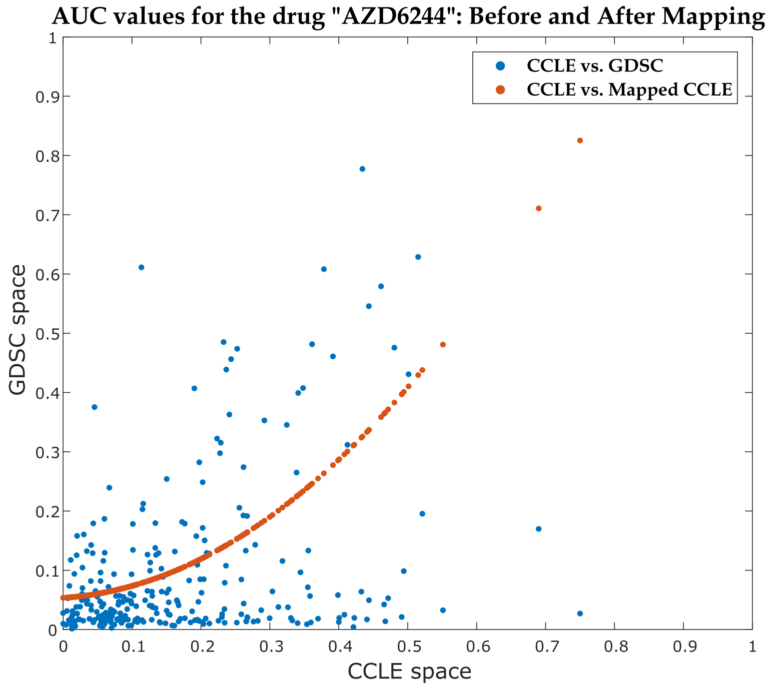

Fig. 5 Scatter plot of AUC association between GDSC and CCLE spaces before and after applying polynomial mapping for the drug "AZD6244" ($\rho_s = 0.26$)

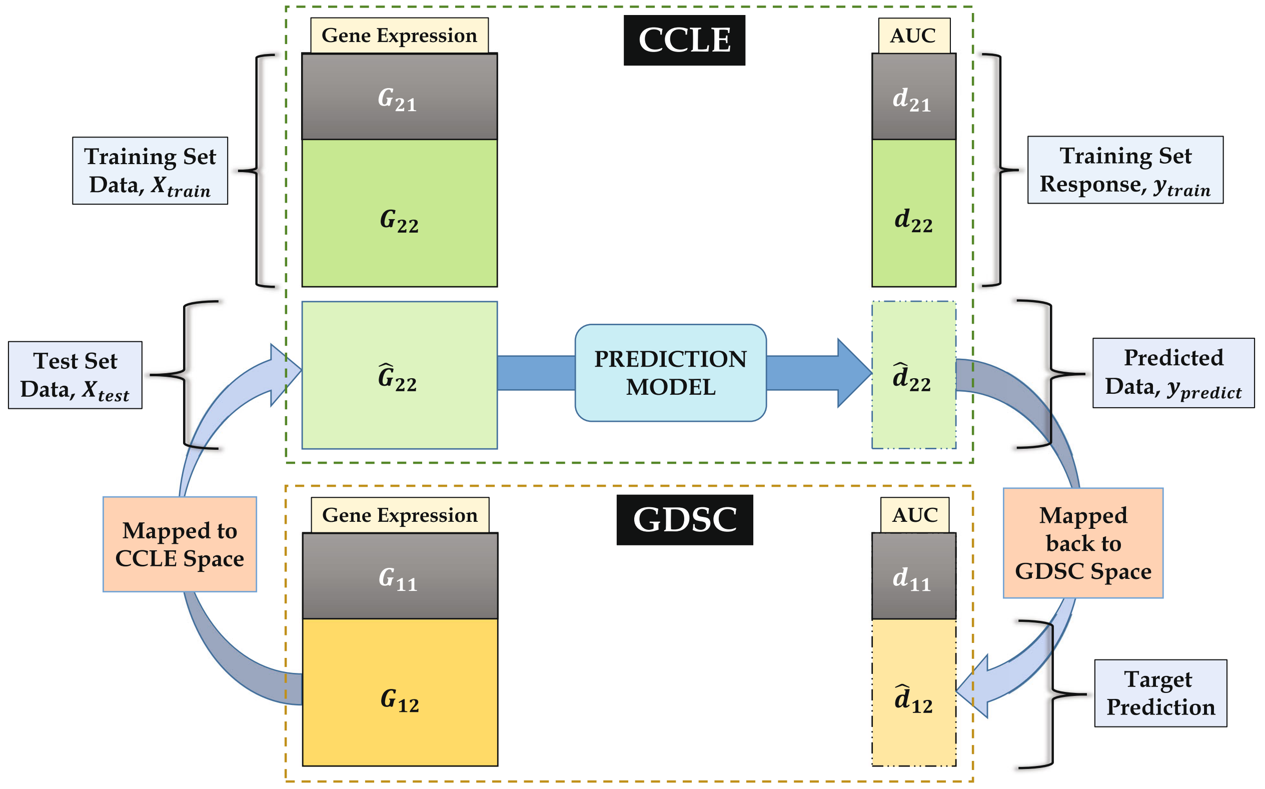

Fig. 6 Illustration of drug sensitivity prediction for GDSC using the nonlinear mapping between CCLE and GDSC

to the other approaches even when the number of features (therefore, information) is reduced to < 50.

Conclusions

In precision medicine, data from multiple large pharmacological studies can be utilized to design better predictive models. In this regard, transfer learning is employed to eliminate the distribution shift between the primary and secondary datasets. In this paper, we have proposed two different TL approaches to incorporate data from two large studies *i.e.*, CCLE and GDSC for designing a better predictive model. In the first approach, we have used a latent variable approach and then optimized the appropriate cost functions to get a pertinent prediction model. The second method uses a nonlinear mapping between both genomic and sensitivity data to transfer the primary data to secondary domain space and perform prediction utilizing the secondary datasets. Both methods show marked

Table 2 Comparison of *K*-fold cross-validation performance for three GDSC drug sensitivity prediction approaches – Mapped Prediction (MP), CCLE model Prediction (CP) and Direct Prediction (DP) using data from CCLE

Drug	Pearson Correlation			NRMSE		
	MP	CP	DP	MP	CP	DP
17-AAG	**0.6062**	0.4354	0.4591	**0.2112**	0.3073	0.2164
AZD6244	**0.4692**	0.3580	0.3579	**0.1683**	0.2173	0.1743
Nilotinib	**0.8698**	0.7957	0.4524	**0.1093**	0.1323	0.1242
Nutlin-3	**0.5606**	0.3102	0.5114	0.1852	0.2180	**0.1808**
PD-0325901	**0.6132**	0.5731	0.4224	**0.1689**	0.1875	0.1865
PD-0332991	**0.0923**	0.0305	0.0802	**0.1748**	0.1764	0.1755
PLX4720	**0.6335**	0.6135	0.5001	**0.1242**	0.159	0.1291

Bold values indicate the best performance

Table 3 Comparison of prediction performance for DITL-KRR approach with Mapped Prediction (MP) and Direct Prediction (DP) approaches for 4 common drugs

Drug	Number of features	Pearson Correlation			NRMSE		
		MP	DP	DITL-KRR	MP	DP	DITL-KRR
17-AAG	47	**0.6319**	0.4749	-0.2885	**0.1942**	0.2167	0.4056
AZD6244	49	**0.4407**	0.4016	-0.1468	**0.1554**	0.1570	0.2042
Nilotinib	35	**0.9338**	0.4674	-0.1701	**0.1003**	0.1257	0.1410
Nutlin-3	48	**0.5921**	0.5207	-0.1500	**0.1881**	0.1903	0.2697

Here, intersection of top 50 genes is taken as the feature set. Bold values indicate the best performance

improvement in drug sensitivity prediction compared to direct prediction and existing TL approaches, while the mapping approach shows the best overall performance.

We have faced a couple of issues during implementation. The LRP approach utilizes the underlying latent variable between the sensitivity datasets and generate the latent variable corresponding to unknown primary sensitivity data. However, to do so, it uses the available secondary data inferring that the prediction can be only performed for matched pair of datasets. Although the LLP approach overcomes this limitation, it often underperforms than LRP. In Table 4, we have presented the applicability of the sensitivity prediction approaches discussed in this paper for matched vs. unmatched pairs of datasets.

Furthermore, in Mapped Prediction, drug sensitivity mapping between databases using polynomials is drug-dependent and thus vulnerable to a user-fault. One potential new step can be modeling the map to be robust against the outliers. Another development can be investigating the effect of model stacking using the proposed approaches.

Methods

Latent variable cost optimization approach

In this section, our goal is to analyze the transfer learning approach from the viewpoint of a cost function optimization. Here, the assumption is that– if there exists such a way to transfer data from both CCLE and GDSC to a common space, then the information available in both databases can be incorporated together to result in a better overall performance [3]. Therefore, it can be inferred that in a suitable common space, the individual concordance between the common set (*i.e.*, underlying latent variable) and each dataset will be maximized and the reconstruction errors from the common set will be minimized. This is the rationale behind the cost function optimization approach.

Table 4 Applicability of Drug Sensitivity Prediction approaches for Matched and Unmatched Pairs of sets between Databases

Prediction Approach	Applicability	
	Matched	Unmatched
Direct Prediction	Yes	Yes
Latent Regression Prediction	Yes	No
Latent-Latent Prediction	Yes	Yes
Combined Latent Prediction	Yes	No
Mapped Prediction (Domain Transfer)	Yes	Yes
Direct Inductive Transfer Learning	Yes	Yes

Drug sensitivity prediction via cost optimization of sensitivity data

In this section, we have deployed cost function optimization of CCLE and GDSC sensitivity data to utilize the underlying latent vector for improving the sensitivity prediction to an anti-cancer drug. The hypothesis is that if both CCLE and GDSC sensitivity vectors can be represented as functions of a common latent variable, then this variable can be utilized along with a known set of CCLE sensitivity values to predict the unknown GDSC sensitivity or vice versa. This approach is regarded as the *Latent Regression Prediction (LRP)*, as the final prediction is performed using a regression model on the latent vector. For this method, only the drug sensitivity values (namely AUC) from the two databases are employed without any use of genomic characteristics data. Figure 2 illustrates the use of LRP method for drug sensitivity prediction. Assume that only a small portion, $(y_{11})_{n_1\times 1}$ of GDSC AUC set, $(y_1)_{n\times 1}$, is known, where $n_1 < n$. Then, the corresponding AUC set, $(y_{21})_{n_1\times 1}$, in CCLE can be used with y_{11} to perform a cost optimization to retrieve the optimum weight vector c for the latent variable, $(w_1)_{n_1\times 1}$, as follows (An additional section provides the detailed development of the cost function [see Additional file 1])

$$\min_{c} \frac{\left\|y_{11} - W_1 a_1\right\|_2^2 + \left\|y_{21} - W_1 a_2\right\|_2^2}{\rho(y_{11}, w_1) + \rho(y_{21}, w_1)} \tag{1}$$

$$\text{subject to} \quad \begin{aligned} &-1 \le c_0 \le 1, \\ &0 \le c_1, c_2 \le 1, \\ &c_1 + c_2 = 1 \end{aligned}$$

where $W_1 = \left[\,\vec{1}\ w_1\,\right]$, $c = \left[\,c_0\ c_1\ c_2\,\right]^T$ and $\vec{1}$ denotes a vector-of-one. Here, w_1 is the latent vector corresponding to y_{11} & y_{21} and assuming linear relationships, c_1 & c_2 are the weights of y_{11} & y_{21} in w_1 (while c_0 is the offset), defined as

$$w_1 = c_0 + c_1 y_{11} + c_2 y_{21} + \varepsilon = \left[\,\vec{1}\ y_{11}\ y_{21}\,\right] c + \varepsilon \tag{2}$$

Now, a_1 & a_2 are the regression coefficients for reconstruction of y_{11} & y_{21} from w_1 and can be obtained from the Least Squares (LS) minimizations of the reconstruction errors (ε).

$$\begin{aligned} y_{11} &= \mathrm{f}_1(w_1) = W_1 a_1 + \varepsilon_1 \\ y_{21} &= \mathrm{f}_2(w_1) = W_1 a_2 + \varepsilon_2 \end{aligned} \tag{3}$$

$$\text{where } a_1 = \begin{bmatrix} a_{10} \\ a_{11} \end{bmatrix},\ a_2 = \begin{bmatrix} a_{20} \\ a_{21} \end{bmatrix}$$

Solving (1), the weight vector, c, and, in turn, a_1, a_2 are found. From (3), it can be inferred that w_1 is also expressed as a linear function of y_{11} or y_{21} alone, *i.e.*

$$w_1 = \begin{cases} \mathrm{g}_1(y_{11}) = \left[\,\vec{1}\ y_{11}\,\right] b_1 + \varepsilon_1' = Y_{11} b_1 + \varepsilon_1' \\ \mathrm{g}_2(y_{21}) = \left[\,\vec{1}\ y_{21}\,\right] b_2 + \varepsilon_2' = Y_{21} b_2 + \varepsilon_2' \end{cases} \tag{4}$$

$$\text{where } b_1 = \begin{bmatrix} b_{10} \\ b_{11} \end{bmatrix},\ b_2 = \begin{bmatrix} b_{20} \\ b_{21} \end{bmatrix}$$

We assume that both CCLE and GDSC sensitivity vectors maintain individual functional relationships with the latent variable, and therefore, the coefficients a_1, a_2, b_1, b_2 will remain the same for the whole response sets (y_1 & y_2 in Fig. 2). Using w_1 and the known CCLE AUC set, y_{21}, the coefficient b_2 in (4) can be retrieved using LS minimization.

$$\min_{b_2} \|w_1 - Y_{21} b_2\|_2^2 \quad \text{which results in } \hat{b}_2 = Y_{21}^{+} w_1 \tag{5}$$

where $(\cdot)^{+}$ denotes the Moore-Penrose pseudoinverse. Using the rest of known CCLE AUC set, $(y_{22})_{n_2\times 1}$, the underlying latent vector, $(w_2)_{n_2\times 1}$, can be retrieved following (4)

$$\hat{w}_2 = \mathrm{g}_2(y_{22}) = \left[\,\vec{1}\ y_{22}\,\right] \hat{b}_2 = Y_{22} \hat{b}_2 \tag{6}$$

Finally, utilizing the coefficient a_1 found initially from solving (1), the unknown GDSC AUC values can be predicted following (3), as

$$\hat{y}_{12} = \mathrm{f}_1(\hat{w}_2) = \left[\,\vec{1}\ \hat{w}_2\,\right] a_1 = \hat{W}_2 a_1 \tag{7}$$

If only a part of CCLE drug sensitivity response is known along with a bigger portion of GDSC sensitivity set, then this whole process can be utilized for the prediction of CCLE responses by interchanging the GDSC and CCLE values.

We have also implemented a *k*NN regression based transfer learning approach for sensitivity prediction [see Additional file 1], which is computationally inexpensive to implement but often underperforms the LRP approach. We then applied an iterative update scheme to improve the performance of *k*NN approach and combined the updated *k*NN model with the LRP model [see Additional file 1]. The combined model shows similar performance to LRP model.

Drug sensitivity prediction via cost optimization of genomic and sensitivity data

In this section, we have utilized both gene expression and AUC data in cost optimization to improve the drug sensitivity prediction. Here, the goal is to establish a relationship between the two underlying latent variables corresponding to gene expression and AUC datasets respectively, and then exploiting this relationship for the prediction of unknown AUC values. This method is regarded as the *Latent-Latent Prediction (LLP)* since it involves the prediction of one latent variable from another. Figure 3 illustrates the use of LLP method for drug sensitivity prediction. Again, we assume that only a small portion, y_{11}, of GDSC AUC set, y_1, is known. Then, the corresponding CCLE AUC set, y_{21}, in CCLE is used with y_{11} to perform the cost optimization in (1) to generate the latent vector w_1 and the regression coefficients a_1, a_2.

Similar to the AUC optimization, the latent vector, $(v_k)_{n\times 1}$, corresponding to the expression vectors, $(x_{1k})_{n_1\times 1}$ & $(x_{2k})_{n_1\times 1}$ of gene k in GDSC & CCLE (where $k = 1, 2, \cdots, p$) can be found as follows (An additional section provides the detailed development of the cost function [see Additional file 1])

$$\min_{\lambda_k} \frac{\|x_{1k} - V_k\alpha_{1k}\|_2^2 + \|x_{2k} - V_k\alpha_{2k}\|_2^2}{\rho(x_{1k}, v_k) + \rho(x_{2k}, v_k)} \tag{8}$$

$$\text{subject to} \quad \begin{aligned} &-1 \le \lambda_{k0} \le 1, \\ &0 \le \lambda_{k1}, \lambda_{k2} \le 1, \\ &\lambda_{k1} + \lambda_{k2} = 1 \end{aligned}$$

where $V_k = \left[\,\vec{1}\ v_k\,\right]$ and $v_k = \left[\,\vec{1}\ x_{1k}\ x_{2k}\,\right] \lambda_k$.

Again, assuming linear relationships, $\lambda_k = \left[\,\lambda_{k0}\ \lambda_{k1}\ \lambda_{k2}\,\right]^T$ is the weight vector of latent v_k corresponding to the expression vectors x_{1k} & x_{2k}, *k*-th columns of the matrices $(X_1)_{n\times p}$ & $(X_2)_{n\times p}$, respectively and α's are the corresponding regression coefficients. The complete latent matrix, $V_{n\times p}$ is found performing this optimization for all p genes and concatenating the individual latent vectors, *i.e.*

$$V = \left[\,v_1\ v_2\ \cdots\ v_p\,\right] \tag{9}$$

For training, the latent matrix $(V_1)_{n_1\times p}$ corresponding to X_{11} and X_{21} is used as model input and w_1 as the corresponding output. The remaining latent, $(V_2)_{n_2\times p}$, is utilized for prediction of the latent vector, $(w_2)_{n_2\times 1}$. The unknown AUC values $(y_{12})_{n_2 \times 1}$ are predicted using (7) again.

$$\hat{w}_2 = \mathcal{M}(V_2) \tag{10}$$

$$\hat{y}_{12} = \mathrm{f}_1(w_2) = \left[\, \vec{1}\ \ \hat{w}_2 \,\right] a_1 = \hat{W}_2 a_1 \tag{11}$$

We have used Random Forest (RF) [18, 20] as our prediction model here. If only a part of CCLE drug sensitivity response is known along with a bigger portion of GDSC sensitivity set, then this whole process can be utilized for the prediction of CCLE responses by interchanging the GDSC and CCLE values.

Combined latent drug sensitivity prediction

To improve the predictive performance of the LLP model and utilize the available CCLE data more effectively, we have incorporated the two latent variable based models together. Here, we combine the predicted latent variables from the two models *i.e.*, w_2^{LRP} from (6) and w_2^{LLP} from (10) via simple averaging and generate the final prediction as before.

$$\hat{w}_2 = \frac{\hat{w}_2^{LRP} + \hat{w}_2^{LLP}}{2} \tag{12}$$

$$\hat{y}_{12} = \left[\, \vec{1}\ \ \hat{w}_2 \,\right] a_1 = \hat{W}_2 a_1 \tag{13}$$

The whole process is depicted as the *Combined Latent Prediction (CLP).* Comparisons among the three optimization based approaches yield that the combined method performs the best while the LLP approach often underperforms than LRP.

Domain transfer approach

In this section, our goal is to analyze whether the dependency structure between CCLE and GDSC can be modeled using a common mapping across different cell lines. The hypothesis is that– if there exists such a common mapping so that the data from one domain can be shifted to the other, then the additional information available in the second database can easily be transferred to the first to produce an overall better performance [3]. For analysis, we have considered a *polynomial regression mapping* [21] and selected the polynomial order by utilizing the Spearman rank correlation (ρ_s) between each pair of datasets from the two databases. This infers a high concordance for gene expression data between databases but poor consistency for drug sensitivity measures such as AUC or IC_{50} [5].

Gene expression mapping

Between GDSC and CCLE, there exist 15,664 common genes in 623 cell lines. Since the rank correlation between CCLE and GDSC gene expression is high (median $\rho_s = 0.86$), we have applied a gene-wise first-order polynomial regression mapping. Assume that $(g_{1i})_{n\times 1}$ and $(g_{2i})_{n\times 1}$ denote the expressions of the i-th gene in GDSC and CCLE, respectively (where $i = 1, 2, \cdots, p$). Then, for each individual gene, the expression mapping from GDSC space to CCLE space

$$\hat{g}_{2i} = \alpha_0^{(i)} + \alpha_1^{(i)} g_{1i} + \varepsilon^{(i)} \tag{14}$$

where $\hat{g}_{2i}$ denotes the mapped gene expression for i-th gene and α's are the regression coefficients quantifying the strength of the association. For the total $n \times p$ gene expression matrices, the equation becomes

$$\begin{aligned} \left[\, \hat{g}_{21}\ \ \hat{g}_{22}\ \cdots\ \hat{g}_{2p} \,\right] &= \left[\, \alpha_0^{(1)} + \alpha_1^{(1)} g_{11}\ \ \alpha_0^{(2)} + \alpha_1^{(2)} g_{12}\ \cdots\ \alpha_0^{(p)} + \alpha_1^{(p)} g_{1p} \,\right] \\ &\quad + \left[\, \varepsilon^{(1)}\ \ \varepsilon^{(2)}\ \cdots\ \varepsilon^{(p)} \,\right] \\ \text{or,}\quad \hat{G}_2 &= \overset{\leftrightarrow}{1} A_0 + G_1 A_1 + \mathcal{E} \end{aligned} \tag{15}$$

where $(A_0)_{p\times p}$ and $(A_1)_{p\times p}$ are two diagonal matrices containing the regression coefficients and $\mathcal{E}_{n_1 \times p}$ is the mapping error. Here, $\overset{\leftrightarrow}{1}$ denotes a matrix-of-one.

$$\begin{aligned} A_0 &= \mathrm{diag}\left(\alpha_0^{(1)}, \alpha_0^{(2)}, \cdots, \alpha_0^{(p)}\right) \\ A_1 &= \mathrm{diag}\left(\alpha_1^{(1)}, \alpha_1^{(2)}, \cdots, \alpha_1^{(p)}\right) \end{aligned} \tag{16}$$

We have performed a drug-wise analysis so that only data corresponding to a single drug is available at a time. Therefore, only a subset of the common 623 × 15664 gene expression matrix is used for each drug, corresponding to the available cell line responses. We used ReliefF [16] to select top 200 genes from both CCLE and GDSC datasets for each drug and took the intersection as the final feature set. Figure 4 illustrates the effect of the mapping for a single gene "DBNDD1". For analysis, we have randomly selected a small subset (*i.e.*, 50 cell lines) of available GDSC samples to get the mapping from the corresponding CCLE data and evaluated the performance on the remaining cell lines. Table 5 shows the correlation between the mapped GDSC expression set with corresponding CCLE set compared to the correlation

Table 5 Comparison of performance of gene expression mapping for two common drugs

Drug	Number of genes	Number of Test cell lines	Pearson Correlation with CCLE		Reconstruction MSE
			Original GDSC	Mapped GDSC	
17-AAG	371	259	0.8729	0.9406	0.8256
AZD6244	383	245	0.8486	0.9405	0.6297

Each result is a mean result for $n = 3$ independent trials between the actual GDSC and CCLE sets for two common drugs and the mean square errors (MSE) for reconstruction. From the correlation and MSE values, it can be inferred that the mapping function successfully captures the interrelationship between CCLE and GDSC gene expression sets.

Drug sensitivity mapping

For drug sensitivity measure, we used the AUC values again. The overall concordance for AUC between databases is poor (median $\rho_s = 0.34$), and therefore, we have considered a drug-wise second-order polynomial regression mapping. Assume that $(d_{1j})_{n\times 1}$ and $(d_{2j})_{n\times 1}$ denote the AUC vectors for the j-th drug in GDSC and CCLE, respectively. Then, for each drug, the drug sensitivity mapping from CCLE space to GDSC space

$$\hat{d}_{1j} = \beta_0 + \beta_1 d_{2j} + \beta_2 d_{2j}^2 + \varepsilon = D_{2j}\beta + \varepsilon, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \tag{17}$$

where $\hat{d}_{1j}$ denotes the mapped drug sensitivity dataset for j-th drug, $D_{2j} = \begin{bmatrix} \vec{1} & d_{2j} & d_{2j}^2 \end{bmatrix}$ is the design matrix, β contains the regression coefficients quantifying the strength of the association and $\varepsilon_{n\times 1}$ is the mapping error.

Note that, out of the 15 common drugs, 3 of the drugs have moderate consistency ($0.5 \le \rho_s < 0.6$) between databases, 3 have fair consistency ($0.4 \le \rho_s < 0.5$) and the rest have poor consistency ($\rho_s < 0.4$). Figure 5 illustrates the effect of the mapping of AUC values from CCLE to GDSC space for the drug AZD6244 with poor consistency between databases ($\rho_s = 0.26$).

For analysis, again we have randomly selected 50 cell lines to get the mapping and evaluated the performance on the rest. Table 6 shows the correlation between the mapped GDSC AUC set with corresponding CCLE set compared to the correlation between the actual GDSC and CCLE sets for two common drugs and MSE for reconstruction. From the correlation and MSE values, it can be inferred that the mapping function captures the interrelationship between CCLE and GDSC drug sensitivity sets satisfactorily.

Drug sensitivity prediction using nonlinear mapping

In this section, we have exploited the interrelationships between CCLE and GDSC through the mapping functions discussed in the previous sections. By using the mapping, we have integrated data from both databases to improve drug sensitivity prediction. Figure 6 illustrates the drug sensitivity prediction procedure using nonlinear mapping. We have performed a drug-wise analysis so that data is available for a single drug at a time. Assume that the GDSC and CCLE gene expression data are expressed as two $n \times p$ matrices, G_1 and G_2, respectively. Furthermore, only a small portion of G_1 *i.e.*, $(G_{11})_{n_1\times p}$, is available with the corresponding AUC values, $(d_{11})_{n_1\times 1}$ where $n_1 < n$, while the whole G_2 matrix is available with the AUC response, $(d_2)_{n\times 1}$. The goal is to predict the unknown AUC values, $(d_{12})_{n_2\times 1}$, for the larger GDSC subset, $(G_{12})_{n_2\times p}$. The CCLE datasets, G_{21} & d_{21}, corresponding to G_{11} & d_{11}, can be utilized in this regard to acquire the individual mapping functions h & f, generated from the polynomial mapping in (15) & (17), respectively.

$$G_{21} = h(G_{11}) = \overset{\leftrightarrow}{1} A_0 + G_{11}A_1 \tag{18}$$

$$d_{11} = f(d_{21}) = \begin{bmatrix} \vec{1} & d_{21} & d_{21}^2 \end{bmatrix} \beta = D_{21}\beta \tag{19}$$

where A_0, A_1 are defined from (16).

To predict the AUC for G_{12}, we map it to CCLE space using the mapping h as $(\hat{G}_{22})_{n_2\times p}$, as in Fig. 6. One can now utilize the additional information in the CCLE space by employing the complete CCLE data G_2 & d_2 for training the prediction model $\mathcal{M}$ while the mapped GDSC set, $\hat{G}_{22}$, is used to predict the sensitivity, $(\hat{d}_{22})_{n_2\times 1}$, in CCLE space. The desired prediction is then obtained by mapping it back to the GDSC space using f.

$$\hat{G}_{22} = h(G_{12}) = \overset{\leftrightarrow}{1} A_0 + G_{12}A_1 \tag{20}$$

$$\hat{d}_{22} = \mathcal{M}\left(\hat{G}_{22}\right) \tag{21}$$

$$\hat{d}_{12} = f\left(\hat{d}_{22}\right) = \begin{bmatrix} \vec{1} & \hat{d}_{22} & \hat{d}_{22}^2 \end{bmatrix} \beta = \hat{D}_{22}\beta \tag{22}$$

The whole process is referred as the *Mapped Prediction (MP)* of GDSC data. Furthermore, if only a part of CCLE gene expression data is available with corresponding drug sensitivity values along with a bigger portion of labeled GDSC data, then this whole process can be utilized for the prediction of CCLE sensitivity by interchanging the GDSC and CCLE values. For prediction using gene expression, we have used a Bias Corrected Random Forest (BC-RF) [19, 22] model where the effect of bias correction is measured using the residual angle [23].

Table 6 Comparison of performance of drug sensitivity (AUC) mapping for two common drugs

Drug	Number of Test cell lines	Pearson Correlation with GDSC		Reconstruction MSE
		Original CCLE	Mapped CCLE	
17-AAG	259	0.5176	0.5232	0.0330
AZD6244	245	0.4022	0.3267	0.0177

Each result is a mean result for $n = 3$ independent trials

Additional file

Additional file 1: Supplementary information to application of transfer learning for cancer drug sensitivity prediction. **Figure S1.** Illustration of kNN image regression prediction for unknown GDSC AUC dataset using the available CCLE data. **Figure S2.** Illustration of change in performance for a single validation set with change in the number of nearest neighbors. **Figure S3.** Illustration of prediction for a single iteration for the Updated kNN image regression prediction. **Figure S4.** Illustration of shift between GDSC and CCLE AUC distributions. **Table S1.** Comparison of MSE for reconstruction and corresponding cost function value for both optimized latent vector and mean latent vector. **Table S2.** Comparison of Pearson correlation and NRMSE among kNN Image Regression Prediction, Latent Regression Prediction and Direct Prediction of GDSC sensitivity using CCLE data. **Table S3.** Comparison of Pearson correlation and NRMSE among kNN Image Regression Prediction, Latent Regression Prediction and Direct Prediction of CCLE sensitivity using GDSC data. **Table S4.** Comparison of Pearson correlation and NRMSE among combined Latent Regression & updated kNN Image Regression Prediction, kNN Image regression Prediction, Latent Regression Prediction and Direct Prediction of GDSC drug sensitivity using CCLE data. **Table S5.** Comparison of Pearson correlation and NRMSE among combined Latent Regression & updated kNN Image Regression Prediction, kNN Image regression Prediction, Latent Regression Prediction and Direct Prediction of CCLE drug sensitivity using GDSC data. **Table S6.** Comparison of K-fold cross-validation results for 4 GDSC drug sensitivity prediction approaches using CCLE – Mapped Prediction, CCLE model Prediction, Combined Model Prediction and Direct Prediction. (PDF 648 kb)

Abbreviations

AUC: Area under the curve; CCLE: Cancer cell line encyclopedia; CLP: Combined latent prediction; GDSC: Genomics of drug sensitivity in cancer; LLP: Latent-latent prediction; LRP: Latent regression prediction; MP: Mapped prediction; NRMSE: Normalized root mean squared error; RF: Random forest; TL: Transfer learning

Acknowledgments

Not applicable.

Funding

This work was supported by NIH grant R01GM122084-01. The publication costs of this article was funded by NIH grant R01GM122084.

Availability of data and materials

For the analysis of transfer learning, the MATLAB codes are available in the following link: https://github.com/dhruba018/Transfer_Learning_Precision_Medicine, while the primary and secondary gene expression and area under the curve data are from the Genomics of Drug Sensitivity in Cancer repository, http://www.cancerrxgene.org/ and Cancer Cell Line Encyclopedia https://portals.broadinstitute.org/ccle, respectively.

About this supplement

This article has been published as part of *BMC Bioinformatics Volume 19 Supplement 17, 2018: Selected articles from the International Conference on Intelligent Biology and Medicine (ICIBM) 2018: bioinformatics*. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-19-supplement-17.

Authors' contributions

SRD, RR, SG and RP conceived of and designed the experiments. SRD and RR performed the experiments. SRD and RP analyzed the data. SRD, RR, KM, SG and RP wrote the paper. All authors have read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Department of Electrical and Computer Engineering, Texas Tech University, 1012 Boston Ave, 79409 Lubbock, TX, USA. [2]Department of Mathematics and Statistics, Texas Tech University, 1108 Memorial Circle, 79409 Lubbock TX, USA.

Published: 28 December 2018

References

1. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR, et al. Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41(D1):955–61.
2. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603–7.
3. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2010;22(10):1345–59.
4. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data. 2016;3(1):9.
5. Haibe-Kains B, El-Hachem N, Birkbak N. J, Jin AC, Beck AH, Aerts HJ, Quackenbush J. Inconsistency in large pharmacogenomic studies. Nature. 2013;504(7480):389–93.
6. Consortium CCLE, of Drug Sensitivity in Cancer Consortium G, et al. Pharmacogenomic agreement between two cancer cell line data sets. Nature. 2015;528(7580):84–87.
7. Celik S, Logsdon BA, Battle S, Drescher CW, Rendi M, Hawkins RD, Lee S-I. Extracting a low-dimensional description of multiple gene expression datasets reveals a potential driver for tumor-associated stroma in ovarian cancer. Genome Med. 2016;8(1):66.
8. Garcke J, Vanck T. Importance weighted inductive transfer learning for regression. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer; 2014. p. 466–81.
9. Al-Stouhi S, Reddy C. Adaptive boosting for transfer learning using dynamic updates. In: Proceedings of the 2011 European conference on Machine learning and knowledge discovery in databases - Volume Part I (ECML PKDD'11), Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis (Eds.), Vol. Part I. Berlin: Springer-Verlag; 2011. p. 60–75.
10. Rückert U, Kramer S. Kernel-based inductive transfer. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer; 2008. p. 220–33.
11. Sugiyama M, Kawanabe M. Machine Learning in Non-stationary Environments: Introduction to Covariate Shift Adaptation. Cambridge: MIT press; 2012, pp. 48–71.
12. Bonilla EV, Chai KM, Williams C. Multi-task gaussian process prediction. In: Advances in Neural Information Processing Systems. USA: Curran Associates Inc.; 2008. p. 153–60.
13. Gao J, Fan W, Jiang J, Han J. Knowledge transfer via multiple model local structure mapping. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2008. p. 283–91.
14. Jiang J, Zhai C. Instance weighting for domain adaptation in nlp. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, ACL, vol. 7. Prague: Association for Computational Linguistics; 2007. p. 264–71.
15. Liao X, Xue Y, Carin L. Logistic regression with an auxiliary data source. In: Proceedings of the 22nd International Conference on Machine Learning. New York: ACM; 2005. p. 505–12.
16. Kira K, Rendell LA. The feature selection problem: Traditional methods and a new algorithm. In: Proceedings of the 10th National Conference on Artificial Intelligence, AAAI, vol. 2. San Jose: AAAI Press / The MIT Press; 1992. p. 129–34.
17. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
18. Rahman R, Otridge J, Pal R. Integratedmrf: random forest-based framework for integrating prediction from different data types. Bioinformatics (Oxford, England). 2017;33(9):1407–1410.
19. Song J. Bias corrections for random forest in regression using residual rotation. J Korean Stat Soc. 2015;44(2):321–6.
20. Rahman R, Haider S, Ghosh S, Pal R. Design of probabilistic random forests with applications to anticancer drug sensitivity prediction. Cancer Informat. 2015;14(Suppl 5):57.
21. Draper NR, Smith H. Applied regression analysis. 1966;709(1):13.

22. Zhang G, Lu Y. Bias-corrected random forests in regression. J Appl Stat. 2012;39(1):151–60.
23. Matlock K, De Niz C, Rahman R, Ghosh S, Pal R. Investigation of model stacking for drug sensitivity prediction. In: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. ACM; 2017. p. 772.