Bioinformatics, 38(1), 2022, 204–210
doi: 10.1093/bioinformatics/btab580
Advance Access Publication Date: 15 August 2021
Original Paper

OXFORD

Data and text mining

A cross-level information transmission network for hierarchical omics data integration and phenotype prediction from a new genotype

Di He [1] **and Lei Xie**[1,2,3,*]

[1]PhD Program in Computer Science, Graduate Center, City University of New York, New York, NY 10016, USA, [2]Department of Computer Science, Hunter College, City University of New York, New York, NY 10065, USA and [3]Feil Family Brain & Mind Research Institute, Weill Cornell Medicine, Cornell University, New York, NY 10021, USA

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Received on March 31, 2021; revised on July 19, 2021; editorial decision on August 3, 2021; accepted on August 12, 2021

Abstract

Motivation: An unsolved fundamental problem in biology is to predict phenotypes from a new genotype under environmental perturbations. The emergence of multiple omics data provides new opportunities but imposes great challenges in the predictive modeling of genotype-phenotype associations. Firstly, the high-dimensionality of genomics data and the lack of coherent labeled data often make the existing supervised learning techniques less successful. Secondly, it is challenging to integrate heterogeneous omics data from different resources. Finally, few works have explicitly modeled the information transmission from DNA to phenotype, which involves multiple intermediate molecular types. Higher-level features (e.g. gene expression) usually have stronger discriminative and interpretable power than lower-level features (e.g. somatic mutation).

Results: We propose a novel Cross-LEvel Information Transmission (CLEIT) network framework to address the above issues. CLEIT aims to represent the asymmetrical multi-level organization of the biological system by integrating multiple incoherent omics data and to improve the prediction power of low-level features. CLEIT first learns the latent representation of the high-level domain then uses it as ground-truth embedding to improve the representation learning of the low-level domain in the form of contrastive loss. Besides, CLEIT can leverage the unlabeled heterogeneous omics data to improve the generalizability of the predictive model. We demonstrate the effectiveness and significant performance boost of CLEIT in predicting anti-cancer drug sensitivity from somatic mutations via the assistance of gene expressions when compared with state-of-the-art methods. CLEIT provides a general framework to model information transmissions and integrate multi-modal data in a multi-level system.

Availabilityand implementation: The source code is freely available at https://github.com/XieResearchGroup/CLEIT.

Contact: lxie@iscb.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Advances in next-generation sequencing have generated abundant and diverse omics data. They provide us with unparalleled opportunities to reveal the secrets of biology. An unsolved problem in biology is how to predict observable traits (phenotypes) given a new genetic constitution (genotype) under environmental perturbations. The predictive modeling of genotype-phenotype associations will answer not only fundamental questions in biology but also address urgent needs in biomedicine. A typical application is anti-cancer personalized medicine. Given a new cancer patient's genetic information, what is the best existing drug to treat this patient? Predicting phenotype from a new genotype is challenging due to the asymmetrical multi-level hierarchical organization of the biological system. Cell-, tissue- and organism-level phenotypes do not arise directly from DNAs but hierarchically through multiple intermediate molecular or cellular phenotypes characterized by protein interactions, gene expressions, etc. (Blois, 1984), as illustrated in Figure 1. In other words, in the information transmission process from DNA to RNA to protein to a biological pathway to the observed phenotype of interest, higher-level features (e.g. gene expression) usually have stronger discriminative and interpretable power than lower level features (e.g. somatic mutation) in a supervised learning task for predicting the phenotype, which is independent on the machine

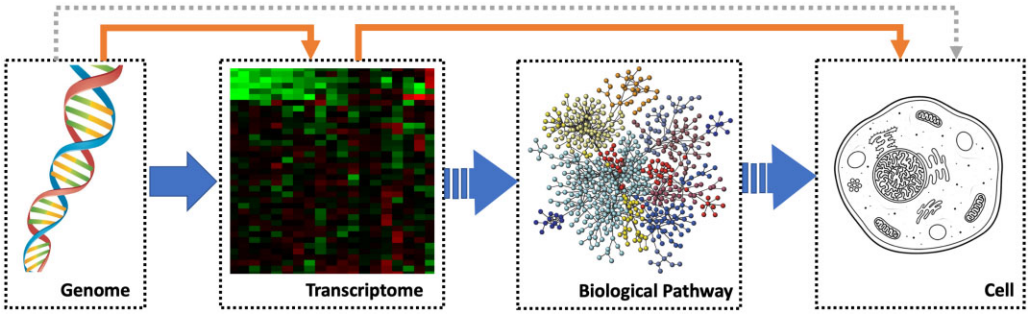

Fig. 1. Rationale of CLEIT. Cellular phenotypes rise from genotypes via multi-level intermediate molecular types hierarchically from DNA to RNA to protein to biological pathway (blue arrows). The predictive and interpretable power of the DNA-level features for the phenotype is weaker than that of the high-level features such as transcriptome and biological pathways. Instead of predicting the phenotype from the genotype directly by bypassing the intermediate molecular types (gray dashed arrow), we will include the information of intermediate molecular type and model the hierarchical organization of a biology system (orange solid arrows)

learning model applied. This premise is supported by multiple studies such as anti-cancer drug sensitivity prediction (Costello *et al.*, 2014), cancer drug combination (Menden *et al.*, 2019), microbiome (Lloyd-Price *et al.*, 2019) and empirical studies (Chiu *et al.*, 2019). Therefore, a multi-scale modeling approach is needed to simulate the asymmetrical hierarchical information transmission process for linking the genotype to the phenotype (Hart and Xie, 2016). It will, in turn, improve the interpretability of model predictions and facilitate clinical decisions. The interpretability of machine learning model is critical for the biomedical application. In principle, the multi-scale modeling of genotype-phenotype associations will facilitate opening the black box of machine learning (Yang *et al.*, 2019). For example, the embedding from the transcriptomics profile, directly or indirectly, can be used to elucidate biological pathways responsible for the synergy of drug combinations (Liu and Xie, 2021). In addition to the above fundamental challenge, the predictive modeling of genotype-phenotype associations faces several technical difficulties that hinder the application of existing machine learning methods. Firstly, omics data are often in an extremely high dimension. Secondly, the coherently labeled data are scarce compared with unlabeled data. Finally, it is not a trivial task to integrate heterogeneous omics data from different resources and modalities.

We develop a novel neural network-based framework: Cross-LEvel Information Transmission (CLEIT) network to address the aforementioned challenges. Inspired by domain adaptation techniques, CLEIT first learns to construct the low-dimensional latent representation that encodes signals indicative of tasks at hand from a high-level domain. Then, CLEIT uses the embedding from the high-level domain as ground-truth embedding to regularize the representation learning of the low-level domain in the form of a contrastive loss. In addition, we adopt a pre-training-fine-tuning approach, where pre-training enables the usage of unlabeled heterogeneous omics data to improve the generalizability of CLEIT, while fine-tuning is employed to enable more task-focused predictions given a specific labeled dataset.

As a demonstration of CLEIT's efficacy in a biological setting, we applied CLEIT to predicting anti-cancer drug sensitivity from somatic mutations. Precision anti-cancer therapy tailed to individual patients based on their genetic profile has gained tremendous interest in clinical (The American Cancer Society, 2020). Existing studies such as Ben-Hamo *et al.* (2019) and Mucaki *et al.* (2019) focused on inferring drug response based on the most salient mutation signatures. Although the drug response of several successful targeted therapies, e.g. kinase inhibitors, can be predicted from a few driver mutations harbored in patients, the percentage of US patients who can benefit from the targeted therapy is only about 4.9% (Marquart *et al.*, 2018). The choice of optimal therapy for most cancer patients remains a significant challenge (Adam *et al.*, 2020). It is well known that cancer acquires numerous mutations during its evolution. Both driver and passenger mutations collectively confer cancer phenotypes and are associated with drug responses (Aparisi *et al.*, 2019). Thus it is necessary to use the entire mutation profile of cancer to predict anti-cancer drug sensitivity in most cases. The machine learning models that can explicitly model hierarchical biological processes will undoubtedly facilitate the development of personalized medicine. In particular, we aim to build accurate predictive models solely using mutation data as inputs to mimic the practical clinical setting, where only patient mutation profiles are available for drug screening. In addition, we use a denoising Autoencoder as a building block and a pre-training-fine-tuning strategy to integrate noisy and sparse mutation, gene expression and protein-protein interaction data from different resources. Our extensive experiments show that CLEIT significantly outperforms other state-of-the-art methods in this regard.

2 Related work

CLEIT aims to develop a framework that constructs an indicative knowledge-abundant low-dimensional latent space from a high dimensional feature space of particular domains, which lacks salient discriminative information of tasks of interest. For example, although somatic mutation data undoubtedly posses biology-rich information, its sparsity and binary characteristics often make it extremely challenging to be utilized to build effective machine learning models for downstream predictive tasks. We resort to a domain adaptation-inspired approach to combat such data limitation issues.

Domain adaptation aims to transfer the knowledge gained on the source domain with sufficient labeled data to the target domain without or with limited labeled data when the source and target domains are of different data distributions. In particular, feature-based domain adaptation approaches (Weiss *et al.*, 2016) have gained popularity along with the advancement in deep learning techniques due to their power in feature representation learning. It aims to learn a shared feature representation by minimizing the discrepancy across different domains while leveraging supervised loss from labeled source domain samples to maintain trait space's discriminative power. Deep domain confusion (DDC) (Tzeng *et al.*, 2014) and CORAL (Sun *et al.*, 2015) focus on exploring proper statistical distribution discrepancy metrics. Domain adversarial neural network (DANN) (Ganin *et al.*, 2016) and adversarial discriminative domain adaptation (ADDA) (Tzeng *et al.*, 2017) intend to minimize the distribution difference across domains with adversarial training and generative adversarial network (Goodfellow *et al.*, 2014), respectively. Moreover, domain separation network (DSN) (Bousmalis *et al.*, 2016) was proposed to separate private representations for each domain and shared representations across domains explicitly. Although CLEIT borrowed some ideas from the domain adaptive transfer learning, there is a significant difference between CLEIT and those approaches. The goal of classic domain adaptation is to use the label information from the source domain data to boost the performance of supervised tasks in the target domain without abundant labels. The feature in the target domain usually has a similar discriminative power to that in the source domain. While in our case, we focus on resolving the inherent discriminative power

discrepancy between two hierarchical related domains. The feature of the high-level domain has higher discriminative power than that of the low-level domain. Moreover, the entity types of source and target domains are usually the same in conventional domain adaptation. In our case, they are of different types. Specifically, our goal for information transmission is to solely push the latent representation of the low-level domain to approximate the one of the high-level domain; that is, the feature representation learned from the high-level domain is fixed and used as ground-truth feature representation of the low-level domain. In this setting, the latent space where the CLEIT happens is no longer a symmetrical consensus from different domains. The high-level and low-level domain is used as an input and an output, respectively, to boost the discrimination power of the low-level domain. A mapping function is learned between them.

The multi-modal integration of somatic mutation and gene expression data has been utilized to improve predicting anti-cancer drug sensitivity, e.g. in Costello *et al.* (2014) and Sharifi-Noghabi *et al.* (2019). These methods assume that both labeled mutation data and labeled gene expression data are available during training and inference. In addition, they integrate omics data horizontally. In contrast, CLEIT only needs to use the mutation data as the input during the inference stage. During the training stage, the mutation and gene expression data can come from different data resources and be unlabelled. Thus, CLEIT is more practical than existing methods. Moreover, CLEIT explicitly models the hierarchical, asymmetrical information transmission in a biological system, as shown in Figure 1.

3 Contributions

CLEIT aims to address an important problem of multi-scale modeling of genotype-phenotype associations. The major contributions of this research are summarized as follows.

- We propose a novel neural network framework that can explicitly model asymmetrical CLEITs in a complex system to boost the discriminative power of the low-level domain. The multi-level hierarchical structure is the fundamental characteristic of the biological and ecological system. The proposed architecture is general and can be applied to model various machine learning tasks where two domains have different features.
- The proposed neural network framework provides a new approach to integrating multiple omics data vertically to represent the multi-level organization of a biological system. The integration of mutation, gene expression and protein-protein interaction data from different resources can help to address the heterogeneity problem.
- We design a pre-training-fine-tuning strategy to fully utilize both labeled and unlabeled omics data that are naturally noisy, high-dimensional and sparse. In particular, the incorporation of autoencoder alleviated the high-dimensionality challenge of omics data and brought in denoising effects. Furthermore, the effective usage of unlabeled data addressed the sparsity of labeled data.
- In terms of biomedical application, the CLEIT significantly improves personalized anti-cancer drug sensitivity prediction using only somatic mutation data. To the best of our knowledge, CLEIT is the first deep learning-based framework designed to perform drug sensitivity prediction tasks solely on whole-genome somatic mutation profiles, which achieves comparable performance to the model trained from gene expression profiles. The oncology panel of somatic mutations has been routinely performed in cancer treatment. The application of CLEIT may improve the effectiveness of cancer treatment and achieve personalized medicine.

4 Materials and methods

4.1 Problem formulation

The problem that we are interested in is to predict the phenotype of interest (e.g. cell viability following drug treatments) of a cell from its mutation profile. Due to the multi-level hierarchical organization of a biological system, RNA-level gene expression profile, can achieve superior performance to DNA-level mutation data for predicting phenotypes independent on machine learning models applied to them. Here, the performance difference is due to the nature of each data domain, instead of the volume of labeled samples as in a classical domain adaptation setting. However, although feature spaces of DNA and RNA domains are not the same, the entities cross the feature spaces are hierarchically related, i.e. the RNA converts the information stored in the DNA. Based on this realization, this work aims to utilize the knowledge learned from the gene expression data to boost the predictive power of the mutation profile. In other words, we want to achieve the similar prediction performance when only using the mutation data as features to that when using the gene expression data. Formally, we denote a data domain D as $D = \{\mathcal{X}, P(X)\}$, where $\mathcal{X}$ stands for the feature space and samples within domain D, $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$. $P(X)$ is the affiliated marginal distribution. In this work, we consider two domains $D_H = \{\mathcal{X}_h, P_h(X_h)\}$ and $D_L = \{\mathcal{X}_l, P_l(X_l)\}$, namely the high-level domain and low-level domain, where $\mathcal{X}_h \neq \mathcal{X}_l$, $P_h(X_h) \neq P_l(X_l)$. In our benchmark experiments, the gene expression is used as D_H, while the somatic mutation is specified as D_L.

4.2 CLEIT framework

To use the knowledge learned from D_H to boost the performance of D_L, we propose a Cross-LEvel-Information Transmission (CLEIT) framework. The strategy of CLEIT is to encode the data from both domains into certain latent features. The embedded latent feature has the direct implication of the task of interests and achieves the CLEIT through transferring knowledge via learned representations cross domains.

Figure 2 shows the overall framework of CLEIT. The training of CLEIT involves five steps: (i) learning an embedding of D_H from unlabeled data using standard autoencoder (AE) (Hinton and Zemel, 1994), (ii) fine-tuning the pre-trained embedding of D_H from step 1 using a multi-layer perceptron (MLP) in the setting of multi-task supervised learning, (iii) and (iv) learning an embedding of D_L from unlabeled data using AE along with the embedding regularization between D_L samples and corresponding D_H samples in the form of an MLP-based transmitter training (v) supervised learning of the final predictive model of D_L using an architecture that appends the pre-trained multi-task MLP (as a warm start) from step 2 as well as the pre-trained AE encoder and the transmitter of D_L from steps 3 and 4. We denote unlabeled D_H samples as $X_{H_u} = \{x_{H_u}^{(i)}\}_{i=1}^{N_{H_u}}$ and labeled samples as $X_{H_l} = \{(x_{H_l}^{(i)}, y^{(i)})\}_{i=1}^{N_{H_l}}$, where $N_{H_\cdot}$ stands for the number of samples in corresponding datasets. Furthermore, $z_{H_\cdot}$ is used to symbolize the latent vectors (embeddings) learned in different phases throughout the training. Samples from the D_L are similarly denoted. For details of methods, see 'Supplementary Method' in Supplementary Material.

5 Experiments

5.1 Experiment set-up

5.1.1 Datasets

We evaluate the performance of CLEIT on a real-world problem: predicting anti-cancer drug sensitivity given the mutation profile of cell lines. The mutation profile (oncology panel) has been implemented in the clinic but has weaker discriminative power for drug sensitivity prediction than the gene expression profile that is not a clinical standard yet. During the training stage, we use unlabeled mutation and gene expression data for unsupervised pre-training, and a small set of labeled data for supervised training. During the testing stage, we only use mutation data from cell lines that do not

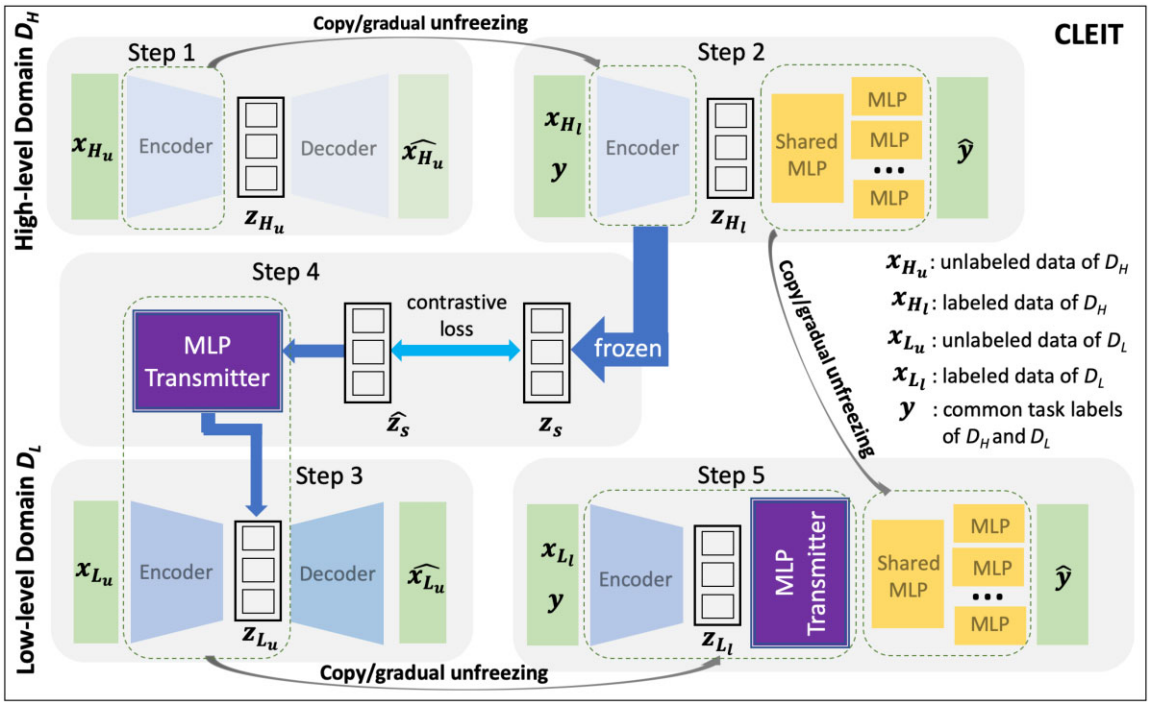

Fig. 2. CLEIT Framework. The training of CLEIT involves five steps. First, the encoder of D_H is learned from an autoencoder and fine-tuned by a supervised multi-task MLP in steps 1 and 2. Then, the embedding of D_L is encoded from an autoencoder in step 3, and the difference between it and that of D_H is minimized via an MLP transmitter in step 4 as measured by contrastive loss. In step 5, the supervised model of D_L is fine-tuned by the model that appends the pre-trained multi-task MLP of D_H in step 2 and the regularized encoder of D_L in step 3

overlap with those used in the training stage to evaluate the performance. Specifically, we collected and integrated data from several diverse resources: cancer cell line data from CCLE (Ghandi *et al.*, 2019), pan-cancer data from TCGA (Goldman *et al.*, 2018), drug sensitivity data from GDSC (Yang *et al.*, 2013) and gene-gene interactions from STRING (Szklarczyk *et al.*, 2019). CCLE includes 1305 and 1697 cancer cell line samples with the gene expression profile and the somatic mutation profile, respectively. The pan-cancer datasets include 9808 and 9093 tumor samples with the gene expression profile and the somatic mutation profile, respectively. Moreover, we only keep the mutation profiles of samples with matched gene expression profiles in our unlabeled mutation dataset. All gene expression data are metricized by the standard transcripts per million base for each gene, with additional log transformation. For the somatic mutation data, we kept only non-silent genes and assembled as a binary-valued sparse vector. Furthermore, we applied pyNBS (Huang *et al.*, 2018), a random walk with restart algorithm, to transform the binary-valued mutation profile into continuous valued features by performing mutation score propagation on STRING gene-gene interaction network. The network-regularized mutation profile will not only reduce the sparsity of features but also significantly boost its prediction power (Huang *et al.*, 2018). We selected the top 1000 varied genes measured by the percentage of unique values in gene expression samples for cancer cell lines and tumor tissue samples separately. The use of 2000 genes achieves the best prediction performance, but is not significantly different from the use of 1000 genes (Supplementary Table S1). Then we combined the two sets of top 1000 varied genes as the input features. The union has 1424 unique genes in total. In addition, we only kept the genes present in the mutation profiles as our final raw feature sets, although CLEIT does not require it. We did so for a fair comparison to other domain adaptation methods since all other methods in comparison consist of a shared encoder component that requires the same number of input features across domains. The final feature set consists of 1407 genes. Furthermore, we matched the omics data of CCLE cell lines against the GDSC drug sensitivity score measured by the Area Under Drug Response Curve (AUC), which is presented as the fraction of the total area under the drug response curve between the highest and lowest screening concentration in GDSC (Yang *et al.*, 2013). The AUC, a continuous-valued drug sensitivity measurement, is used across our experiments as the dependent variable for the supervised fine-tuning. In total, we assembled 680 CCLE cell lines with both mutation and gene expression, which are associated with 93 anti-cancer drugs after removing drugs that have more than 10% missing drug sensitivity measurements within these cell line samples. These 680 cell lines and 59 203 drug sensitivity data were used as training data in the fine-tuning stage. Additional non-overlapping 278 cell lines that have only mutation information were used as hold-out testing data in our study. By combining both TCGA and CCLE datasets, 11 113 and 9743 samples that do not have measured drug sensitivities were used as unlabeled data in the pre-training stage. The gene expression profile is considered as D_H, while the mutation is D_L. A summary of the pre-processed data are shown in Table 1.

5.1.2 Training, validation and testing procedure

To demonstrate CLEIT's stable performance in the given anti-cancer drug sensitivity prediction task, we repeated the model training five times. First, we split the labeled dataset that has both gene expression and mutation profile into 5-folds. Then, in each repetition, we used four out of five folds as the labeled training set, the remaining one fold left as the validation set. The detailed training procedure of CLEIT is summarized as follows. In the D_H pre-training, we trained CLEIT for N epochs. With parameter grid search, N is selected based on the target task performance. While for the fine-tuning of D_H, we employed early stopping with validation labeled fold (only gene expression) as mentioned earlier in this section. For the pre-training of D_L, similar to pre-training of D_H, we specified the number of epochs based on the task-specific performance. In the fine-tuning of D_L, we employed early stopping with the same validation fold (only mutation) in the fine-tuning of D_H. The final trained model is used to make predictions on a labeled mutation-only test set. All other baseline models followed the same training and testing procedure.

Table 1. Summary of pre-processed data for training and testing

Category	Unlabeled (pre-training)	Labeled (fine-tuning)	Labeled (test)
Gene Expression (#samples)	11 113	680	NA
Somatic Mutation (#samples)	9743	680	278
Drug Sensitivity (#cell line-drug pairs)	NA	59203	23475

5.1.3 Performance evaluation

We evaluated CLEIT's performance by predicting drug sensitivity on a hold-out labeled mutation-only test data. We measured the regression performance using Pearson correlation, Spearman correlation, RMSE (root mean squared error). Note that there is a maximum of 93 drug sensitivity scores associated with each cell line sample. The results are shown with the average performance per cell line sample (sample-wise) and per drug (drug-wise). Besides, because of the incompleteness of the ground truth matrix, the prediction entries without a ground truth sensitivity score are filtered out in the calculation of each evaluation metric.

5.2 Baseline models

We compared CLEIT with the following base-line models: MLP without and with the AE pre-training for D_L as well as several of the most popular domain adaptation algorithms that are used to transfer the knowledge learned from D_H to D_L. They include Deep Domain Confusion (DDC) network (Tzeng *et al.*, 2014), Correlation Alignment (CORAL, Sun *et al.*, 2015), Domain Adversarial Neural Network (DANN, Ganin *et al.*, 2016), Adversarial Domain Adaptation Network (ADDA, Tzeng *et al.*, 2017) and Domain Separation Network (DSN, Bousmalis *et al.*, 2016). Specifically, we used the original architecture of baseline models but exactly same features and training/testing data and procedure as CLEIT so that we have a fair performance comparison between them. Specifically, for DDC, CORAL, DANN and ADDA, we followed their original approach and combined their respective domain adaptation objectives with drug response prediction in the supervised training (same dataset used in CLEIT fine-tuning). For DSN, we employed its MMD variant for the stability of training and adopted the same pre-training fine-tuning process as used in CLEIT. In the pre-training, we leveraged unlabeled data from both domains to pre-train the encoders with an autoencoder reconstruction task and its domain adaptation objective. In the fine-tuning, we further adapt the encoder and appended the predictor with the labeled drug response dataset. To evaluate the contribution of different components in CLEIT, we performed ablation studies by (i) removing the unlabeled pre-training process and incorporating the cross-level transmission loss into the labeled training, (ii) removing the transmitter, (iii) changing the cross-level transmission loss function to Maximum Mean Discrepancy (MMD) loss (Gretton *et al.*, 2012) and Earth Mover distance approximated using Wasserstein-GAN (WGAN) (Gulrajani *et al.*, 2017). The latent dimension for hidden representation for all models is specified as 128, and all autoencoder frameworks share the same [512, 256, 128, 256, 512] architecture. Besides, all pre-trained encoders will be appended with a predictor module of the same architecture ([128] shared layer + [64,32] individual drug MLP) for the fine-tuning process.

6 Results and discussion

6.1 Gene expression feature has stronger predictive power than somatic mutation-based feature

Consistent with extensive performance evaluations from blind tests in a DREAM challenge (Costello *et al.*, 2014), and other studies (Chiu *et al.*, 2019), the gene expression feature has more substantial predictive power than the mutation-based feature. As shown in Supplementary Figure S1(a), the model trained with only labeled gene expression data has a 6.45% performance gain over the model trained with corresponding labeled somatic mutation data when evaluated using a sample-wise average. With the additional utilization of unlabeled pre-training, models trained with only gene expression data and only mutation data both showed slightly better performance, while their performance gap is around 6.8%. In terms of drug-wise average, as shown in Supplementary Figure S1(b), the performance gap between models built on mutation-only and expression only data is even more apparent. The multi-modal learning that combines the mutation and gene expression features fails to improve the performance (Supplementary Fig. S1). These results confirmed that the gene expression is more predictive than the somatic mutation for predicting the anti-cancer drug sensitivity.

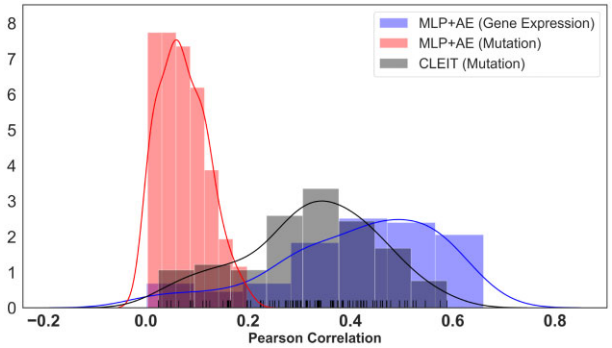

Fig. 3. Drug-wise Pearson correlation on validation dataset

6.2 CLEIT can transfer the knowledge learned from gene expression features to the model with mutation features

To demonstrate that CLEIT can transfer the knowledge learned from the gene expression feature to the model that uses the mutation-only data, we compared the drug-wise Pearson correlation distribution of CLEIT with those of the MLP+AE models trained with only gene expression or mutation data. Figure 3 shows the histogram of Pearson correlations of 93 drugs for three models. CLEIT using the mutation data shifts the performance distribution close to the model trained using the gene expression data with a False Discovery Rate (FDR) of 1.0 based on Kolmogorov-Smirnov (KS) test. It significantly outperforms the MLP+AE model using the mutation data (FDR = 3.47e-37 of KS test). It is aligned with our primary goal in this work. Note that the histograms in Figure 3 were from the validation data. Next, we evaluate the performance of CLEIT in a hold-out mutation-only test data.

6.3 CLEIT significantly outperforms state-of-the-art models to predict anti-cancer drug sensitivity using mutation-only data

Given that gene expression data have stronger predictive power than somatic mutation data, we evaluate if CLEIT can use the gene expression to boost the performance for predicting anti-cancer drug sensitivity when only the somatic mutation data are available as the input. The results for both drug-wise and sample-wise evaluation are shown in Tables 2 and 3. As seen in those tables, models that consist of unlabeled pre-training processes generally outperform the models trained with only labeled data, indicating the importance of leveraging unlabeled data. The models trained with domain adaptation methods with unlabeled pre-training (DSN or CLEITs) or only

Table 2. Evaluation results on test data (drug-wise)

Method	Pearson	Spearman	RMSE
MLP (mutation-only)	0.0591 ± 0.0069	0.0532 ± 0.0066	0.0233 ± 0.0018
MLP+AE (mutation-only)	0.0681 ± 0.0085	0.0629 ± 0.0108	0.0151 ± 0.0001
DDC	0.0633 ± 0.0087	0.0621 ± 0.0087	0.0150 ± 0.0006
CORAL	0.0580 ± 0.0105	0.0542 ± 0.0080	0.0164 ± 0.0005
DANN	0.0571 ± 0.0061	0.0516 ± 0.0038	0.0173 ± 0.0010
ADDA	0.0681 ± 0.0111	0.0685 ± 0.0142	0.0197 ± 0.0010
DSN	0.1003 ± 0.0186	0.0915 ± 0.0252	0.0147 ± 0.0007
CLEIT (w/o pre-training)	0.1005 ± 0.0236	0.0924 ± 0.0216	0.0147 ± 0.0005
CLEIT (w/o transmitter)	0.2587 ± 0.0126	0.2254 ± 0.0348	0.0124 ± 0.0006
CLEIT (MMD)	0.1758 ± 0.0086	0.1421 ± 0.0200	0.0148 ± 0.0009
CLEIT (WGAN)	0.0795 ± 0.0083	0.0821 ± 0.0106	0.0150 ± 0.0009
CLEIT	**0.2770** ± 0.0086	**0.2482** ± 0.0243	**0.0121** ± 0.0006

Note: The best results are shown in bold.

Table 3. Evaluation results on test data (sample-wise)

Method	Pearson	Spearman	RMSE
MLP (mutation-only)	0.7390 ± 0.0017	0.6957 ± 0.0022	0.0235 ± 0.0017
MLP+AE (mutation-only)	0.7450 ± 0.0003	0.6984 ± 0.0004	0.0150 ± 0.0001
DDC	0.7449 ± 0.0017	0.7010 ± 0.0010	0.0151 ± 0.0004
CORAL	0.7439 ± 0.0013	0.7002 ± 0.0010	0.0165 ± 0.0004
DANN	0.7428 ± 0.0017	0.6995 ± 0.0019	0.0174 ± 0.0008
ADDA	0.7315 ± 0.0053	0.6891 ± 0.0010	0.0199 ± 0.0008
DSN	0.7470 ± 0.0002	0.7024 ± 0.0004	0.0148 ± 0.0004
CLEIT (w/o pre-training)	0.7467 ± 0.0003	0.7023 ± 0.0004	0.0149 ± 0.0004
CLEIT (w/o transmitter)	0.7569 ± 0.0081	0.7172 ± 0.0070	0.0125 ± 0.0005
CLEIT (MMD)	0.7443 ± 0.0018	0.7003 ± 0.0009	0.0147 ± 0.0009
CLEIT (WGAN)	0.7465 ± 0.0005	0.7022 ± 0.0008	0.0152 ± 0.0009
CLEIT	**0.7640** ± 0.0094	**0.7233** ± 0.0063	**0.0122** ± 0.0005

Note: The best results are shown in bold.

labeled training outperform their non-domain adaptation counterparts. It implies that D_L will benefit from the knowledge transfer from D_H. Furthermore, CLEIT models significantly outperform all other models in consideration (t-test P-value < 0.05). The best-performed model is the CLEIT that uses contrastive loss. Compared with the best performed state-of-the-art model (DSN), the accuracy of CLEIT, when measured by Pearson correlation, improves 277% and 2.2% for the drug-wise and the sample-wise test, respectively. Similar results can be seen in terms of Spearman correlation and RMSE. The performance gain of CLEIT over MLP and MLP+AE is 3.4% and 2.5%, respectively, in the sample-wise setting. Yet in the drug-wise setting, the improved gap is enlarged to 469% and 407%. The much improved drug-wise performance achieved by CLEIT indicated a much higher quality drug-sensitivity prediction with the mutation-only data.

CLEIT models that incorporate MLP-transmission function show significantly better performance than those without, suggesting that the transmission function plays a role in CLEIT. Choice of the loss function in the information transmission is also important. It is clear that contrastive loss performs better than MMD and WGAN. It is noted that MMD is used in DSN. When CLEIT uses MMD as the loss function to measure the domain discrepancy, the major difference between CLEIT and MMD is that CLEIT treats the information transmission between two domains asymmetrical, while DSN considers domain adaptation symmetrical. The results in Tables 2 and 3 show that CLEIT-MMD outperforms DSN in drug-wise setting and perform similarly in sample-wise settings. It indicates that the explicit modeling of the hierarchical organization of D_L and D_H is important.

6.4 CLEIT outperforms state-of-the-arts for predicting top-ranked cell-line specific anti-cancer therapies

Furthermore, CLEIT can predict the best therapy for a new patient using only mutation data for personalized medicine. We compared the performance of different methods with the precision of top-k ($k = 1, 10$) predictions ranked by the AUC scores, which is defined as the ratio of drugs with top-k smallest predicted scores per cell line among the drugs with top-k ground-truth scores. Mutation-only test results can be found in Figure 4. Clearly, the CLEIT model also outperforms other models in this scenario. Compared with the second-best performed model DSN, CLEIT improves the performance by approximately 40% when $k = 1$.

7 Conclusion

This article proposed a novel machine learning framework CLEIT for the predictive modeling of genotype-phenotype associations by explicitly modeling the asymmetric CLEIT in the biological system. Using the anti-cancer drug sensitivity prediction with only mutation data as a benchmark, CLEIT clearly outperforms existing methods and demonstrates its potential in personalized medicine. Although we only study the knowledge transfer between DNA level and RNA level in this article, the same strategy can be applied to other levels in the biological system, for example, imputing proteomics data using transcriptomics data. Nevertheless, the performance of CLEIT could be further improved by incorporating domain knowledge. For example, an autoencoder module that can model gene-gene interactions and biological pathways will be greatly helpful. Under the framework of CLEIT, it is not difficult to integrate other omics data such as epigenomics and proteomics. They may further improve the

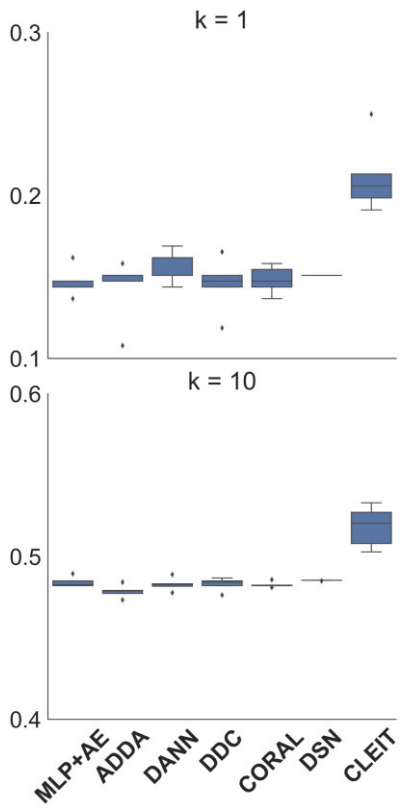

Fig. 4. Top K Precision on Mutation-only Test Dataset

performance of CLEIT. Another challenge in personalized medicine is to transfer knowledge from cell line data to patient tissue data (He and Xie, 2021). It will be interesting to develop new neural network architectures in the framework of CLEIT to address this problem.

Acknowledgement

The authors thank for reviewer's constructive comments.

Funding

This work was supported by the National Institute of General Medical Sciences of National Institute of Health [R01GM122845] and the National Institute on Aging of the National Institute of Health [R01AD057555].

Conflict of Interest: none declared.

References

Adam,G. *et al.* (2020) Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ. Precision Oncol.*, **4**, 19–10.

Aparisi,F. *et al.* (2019) Passenger mutations in cancer evolution. *Cancer Reports and Reviews*, **3, doi: 10.15761/CRR.1000188**.

Ben-Hamo,R. *et al.* (2019) Resistance to paclitaxel is associated with a variant of the gene bcl2 in multiple tumor types. *NPJ. Precision Oncol.*, **3**, 12–11.

Blois,M.S. (1984) *Information and Medicine: The Nature of Medical Descriptions*. University of California Press, Berkeley, CA.

Bousmalis,K. *et al.* (2016) Domain separation networks. *Adv. Neural Inf. Process. Syst.*, **29**, 343–351.

Chiu,Y.-C. *et al.* (2019) Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Med. Genomics*, **12**, 18.

Costello,J.C. *et al.*; NCI DREAM Community. (2014) A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.*, **32**, 1202–1212.

Ganin,Y. *et al.* (2016) Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, **17**, 2096–2030.

Ghandi,M. *et al.* (2019) Next-generation characterization of the cancer cell line encyclopedia. *Nature*, **569**, 503–508.

Goldman,M. *et al.* (2018) The ucsc xena platform for cancer genomics data visualization and interpretation. *BioRxiv*, doi: 10.1101/326470.

Goodfellow,I. *et al.* (2014) Generative adversarial nets. *Adv. Neural Inf. Process. Syst.*, **27**, 2672–2680.

Gretton,A. *et al.* (2012) A kernel two-sample test. *J. Mach. Learn. Res.*, **13**, 723–773.

Gulrajani,I. *et al.* (2017) Improved training of wasserstein gans. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 5767–5777. Long Beach, CA, USA.

Hart,T. and Xie,L. (2016) Providing data science support for systems pharmacology and its implications to drug discovery. *Exp. Opin. Drug Discov.*, **11**, 241–256.

He,D. and Xie,L. (2021) CODE-AE: a coherent de-confounding autoencoder for predicting patient-specific drug response from cell line transcriptomics, arXiv:2102.00538.

Hinton,G.E. and Zemel,R.S. (1994) Autoencoders, minimum description length, and Helmholtz free energy. *Adv. Neural Inf. Process. Syst.*, **6**, 3–10.

Huang,J.K. *et al.* (2018) pynbs: a python implementation for network-based stratification of tumor mutations. *Bioinformatics*, **34**, 2859–2861.

Liu,Q. and Xie,L. (2021) Transynergy: mechanism-driven interpretable deep neural network for the synergistic prediction and pathway deconvolution of drug combinations. *PLoS Comput. Biol*, 17, e1008653.

Lloyd-Price,J., IBDMDB Investigators. *et al.* (2019) Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, **569**, 655–662.

Marquart,J. *et al.* (2018) Estimation of the percentage of us patients with cancer who benefit from genome-driven oncology. *JAMA Oncol.*, **4**, 1093–1098.

Menden,M.P. *et al.*; AstraZeneca-Sanger Drug Combination DREAM Consortium. (2019) Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat. Commun.*, **10**, 2674–2617.

Mucaki,E.J. *et al.* (2019) Predicting responses to platin chemotherapy agents with biochemically-inspired machine learning. *Signal Transduct. Targeted Ther.*, **4**, 1–12.

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

Sun,B. *et al.* (2015) Return of frustratingly easy domain adaptation. *arXiv Preprint arXiv, 1511, 05547, arXiv:1511.05547*.

Szklarczyk,D. *et al.* (2019) String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.

The American Cancer Society. (2020) Precision or personalized medicine. www.cancer.org/treatment/treatments-and-side-effects/treatment-types/precision-medicine (18 March 2021, date last accessed).

Tzeng,E. *et al.* (2014) Deep domain confusion: maximizing for domain invariance. *arXiv Preprint arXiv, 1412, 3474*.

Tzeng,E. *et al.* (2017) Adversarial discriminative domain adaptation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7167–7176, Honolulu, HI.

Weiss,K. *et al.* (2016) A survey of transfer learning. *J. Big Data*, **3**, 9.

Yang,J.H. *et al.* (2019) A white-box machine learning approach for revealing antibiotic mechanisms of action. *Cell*, **177**, 1649–1661.

Yang,W. *et al.* (2013) Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, **41**, D955–D961.