Guvencpaltun, Betul; Kaski, Samuel; Mamitsuka, Hiroshi

DIVERSE: Bayesian Data IntegratiVE learning for precise drug ResponSE prediction

Published in:
IEEE/ACM Transactions on Computational Biology and Bioinformatics

DOI:
10.1109/TCBB.2021.3065535

Published: 11/04/2022

Document Version
Publisher's PDF, also known as Version of record

DIVERSE: Bayesian Data IntegratiVE Learning for Precise Drug ResponSE Prediction

Betül Güvenç Paltun, Samuel Kaski, and Hiroshi Mamitsuka

Abstract—Detecting predictive biomarkers from multi-omics data is important for precision medicine, to improve diagnostics of complex diseases and for better treatments. This needs substantial experimental efforts that are made difficult by the heterogeneity of cell lines and huge cost. An effective solution is to build a computational model over the diverse omics data, including genomic, molecular, and environmental information. However, choosing informative and reliable data sources from among the different types of data is a challenging problem. We propose DIVERSE, a framework of Bayesian importance-weighted tri- and bi-matrix factorization (DIVERSE3 or DIVERSE2) to predict drug responses from data of cell lines, drugs, and gene interactions. DIVERSE integrates the data sources systematically, in a step-wise manner, examining the importance of each added data set in turn. More specifically, we sequentially integrate five different data sets, which have not all been combined in earlier bioinformatic methods for predicting drug responses. Empirical experiments show that DIVERSE clearly outperformed five other methods including three state-of-the-art approaches, under cross-validation, particularly in out-of-matrix prediction, which is closer to the setting of real use cases and more challenging than simpler in-matrix prediction. Additionally, case studies for discovering new drugs further confirmed the performance advantage of DIVERSE.

Index Terms—Personalized medicine, drug response prediction, machine learning, Bayesian methods, data integration

1 Introduction

Identification of predictive biomarkers for drug sensitivity plays a significant role for assigning the most effective treatments to patients with complex diseases such as cancer [1]. However, it is impracticable to clinically assess each patient's response to disease due to the large population. Patients of the same cancer type may differ in their responses to a specific medical therapy because of the large genetic diversity of cancer [2]. Personalized medicine provides an understanding of cancer cell lines at the molecular level and recommends individualized therapies to patients that allow high efficacy in different cancer types by measuring drug responses [3].

The research is most often done with cell lines which, even though much simpler than real patients, are already complex enough and require multiple data sets to characterize sufficiently for prediction. Since cancer cell lines show distinct characteristics caused by a multitude of factors, including genetic mutations, molecular interactions, and environmental sources, complicates the discovery of predictive biomarkers. Fortunately, recent high-throughput technologies have generated a considerable amount of biological data from different viewpoints. This diverse data could allow precise computational prediction of drug sensitivity of cancer cell lines based on molecular interactions, genomic features, and chemical structures. However, although large-scale data have been generated for drug response prediction, many machine learning methods have failed to achieve good performance for multiple heterogeneous data sources, because these methods have been designed for only a single type of data. Thus a challenging task is to build precise prediction models on diverse data, coming from different sources, which are difficult to compare. In fact, data integration has to overcome several obvious problems, such as different data sizes, complexity, and noisiness. However, more importantly, data-integrative machine learning methods need to decide which information is useful to be incorporated and how significant the information is for the prediction task. This is the most critical problem to be addressed for machine learning models with diverse multi-omics data. For this problem, we propose DIVERSE, a framework to efficiently integrate scientifically diverse data, i.e., genomic, chemical and molecular interaction information, to predict missing drug responses of cancer cell lines. The three key points of DIVERSE are:

i) DIVERSE integrates five biologically different data sets: drug similarity, gene expression, protein-protein interaction, drug-target interaction and cell line-drug interaction. To the best of our knowledge, this is the largest number of heterogeneous data sources

- *Betül Güvenç Paltun is with the Department of Computer Science, Aalto University, 02150 Espoo, Finland. E-mail: betul.guvenc@aalto.fi.*
- *Samuel Kaski is with the Department of Computer Science, Aalto University, 02150 Espoo, Finland, and also with the Department of Computer Science, The University of Manchester, M13 9PL Manchester, U.K. E-mail: samuel.kaski@aalto.fi.*
- *Hiroshi Mamitsuka is with Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji 611-0011, Japan, and also with the Department of Computer Science, Aalto University, 02150 Espoo, Finland. E-mail: mami@kuicr.kyoto-u.ac.jp.*

Manuscript received 16 October 2020; revised 16 February 2021; accepted 4 March 2021. Date of publication 11 March 2021; date of current version 8 August 2022.
(Corresponding author: Betül Güvenç Paltun.)
Digital Object Identifier no. 10.1109/TCBB.2021.3065535

combined for drug response prediction so far. No competing bioinformatics methods can integrate the same types of data sets.

ii) It does not allow any of the five different data sets dominate the prediction. DIVERSE adds one data set by one in a systematic and step-wise manner.

iii) It is methodologically flexible. Most existing studies ignore uncertainty, and hence cannot accept missing values. Second, in general, integrating different data sets makes it harder to obtain the correct rank of given data or matrices. DIVERSE solves these two practically important problems by using a Bayesian setting.

We empirically validated the performance of DIVERSE, comparing with five other methods, including three state-of-the-art methods, under 5x5-fold cross-validation. Experimental results indicate that DIVERSE significantly outperformed all compared methods in both mean-squared error (MSE) and Spearman correlation coefficient (Sc), particularly for out-of-matrix prediction, which is a real-world setting and much harder than in-matrix prediction. Results clearly show the performance advantage of DIVERSE over the current methods for predicting drug responses. Also, the results indicate that the MSE and Sc of DIVERSE were smoothly improved by the step-wise addition of each data set. These advantages of DIVERSE were confirmed by several case studies.

2 Related Work

The promise of personalized medicine has been a theme in clinical discussion for some time, and researchers have developed a variety of computational methods. Some state-of-the-art algorithms focusing on drug response prediction include elastic net [4], support vector machines [5], kernel ridge regression (KRR) [6], random forest and neural networks [7]. However, most of the traditional approaches underestimate the complexity of cancer caused by a number of environmental factors, genetic mutations, and somatic alterations in drug response prediction. A variety of studies have been taking into account the complex relationships between cell lines, chemical structures, and genomic alterations to identify predictive biomarkers [3].

The advantage of incorporating heterogeneous information for drug response prediction analysis has been highlighted in recent studies [8], [9]. Ammad-ud din *et al.* proposed in their cwKBMF that utilizing genomic data increases predictive performance, and incorporating prior biological knowledge enhances it even further [10]. SRMF [11] was proposed as a matrix factorization method to simultaneously incorporate drug and cell line similarity information for drug response prediction. Multiple non-negative matrix factorization (MultiNMF) models have been designed for integrating data sets by sharing one of the factor matrices, details can be found in the study by [12]. DrugCellNet [13] assumes that the response of a known drug in a new cell line is a weighted combination of the responses of the neighboring cell lines. [14] focused on solving the "small n, large p" problem when the number of genes is larger than samples, through integrating cell line and drug information with a PPI network by utilizing functional links. HNMDRP was proposed by [15] as a classification problem of whether that the drug is whether sensitive or resistant, based on data on gene expression profiles from cell lines, drug chemical structure features, drug-target interactions, and PPIs. Even though HNMDRP utilizes enough data, it cannot predict unseen (new) drugs and cell lines. Recently, [16] developed a time- and memory-efficient learning method with multiple pairwise kernels that can integrate various types of biological data sources, which, however, requires that the data comes in the form of observations for pairs of entities, such that data only includes drug and cell line information. Another network-based drug response prediction method MOLI [17], was recently proposed and built on deep neural networks for feature selection. MOLI learns the features of each data item, and then concatenates them for predicting drug response values. Detailed comparisons of recent machine learning models in data-integrative drug response prediction can be found in a recent review by [3].

DIVERSE has three unique aspects to address the limitations of existing models and to develop more robust models. That is, DIVERSE can 1) integrate several entities such as drugs, cell lines, and genes to predict missing entries and unseen drugs. For example, many methods have to use cell line similarity information transformed from gene expression data since they can not integrate gene-related information; 2) use a computationally feasible method to integrate data sets because when a new data set is combined, DIVERSE can use the same entity-specific factors without additional computation; 3) use non-negativity constraints in a Bayesian setting to reduce overfitting on noisy data, sustaining better interpretation after factorization.

3 Methods

DIVERSE allows predicting drug responses of cancer cell lines by incorporating information from heterogeneous data sets. DIVERSE consists of two key elements: 1) Section 3.1.2–4: Bayesian non-negative matrix factorization that is used to determine latent factors of data sets, including data describing relations between drugs, cell lines, and genes. 2) Section 3.1.5–6: hybrid matrix factorization model to simultaneously integrate heterogeneous data sets. This combination of methods is new for predicting drug sensitivity.

3.1 Prediction Model

3.1.1 Prediction Problem

The goal of this work is to predict missing entries of a drug response matrix given the other matrices. This problem consists of two different tasks. First, we predict an unknown value of a pair of a drug and a cell line, for a drug for which other values are already given (observed). Second, we predict all responses of an unseen (new) drug which has no observed values in the matrix yet. Drug response data consists of IC50 values that give the effectiveness of drugs on different cell lines. Additional given inputs are drug similarity, gene expression, protein-protein interaction, drug-target interaction and cell line-drug interaction data sources. Details of the data will be described in Section 4.1.

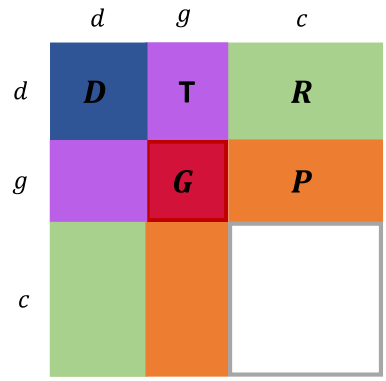

Fig. 1. Conceptual integration configuration of the multiple data from three types of entities; d, g and c denote drugs, genes and cell lines respectively.

3.1.2 Notation

The main input is the drug response matrix $\mathbf{R} \in \mathbb{R}^{N \times M}$, in which rows correspond to drugs and columns to cell lines. Each entry in $\mathbf{R}$ is the response value of a single drug in a certain cell line. In order to represent associations between cell lines and genes, we use $\mathbf{P} \in \mathbb{R}^{M \times L}$. The chemical similarity matrix of drugs is encoded as $\mathbf{D} \in \mathbb{R}^{N \times N}$. $\mathbf{T} \in \mathbb{R}^{L \times N}$ is the drug-target interaction matrix, where $\mathbf{T}_{ij}$ 1 if there is interaction between drug i and gene j. Similarly, protein-protein interaction (PPI) matrix $\mathbf{G} \in \mathbb{R}^{L \times L}$ represents the functional relations between proteins. Here matrices are denoted by capital letters. The detailed interaction between matrices can be seen in Fig. 1. Table 1 shows the list of notations used throughout this paper.

3.1.3 Non-Negative Matrix Tri-Factorization

The drug response matrix $\mathbf{R}$ can be mapped to a non-negative low-dimensional latent factor space and regarded as the product of three matrices as follows:

$$\mathbf{R} \approx \mathbf{U}\mathbf{S_r}\mathbf{V}^{\mathsf{T}}. \quad \text{Here} \quad \mathbf{U}, \mathbf{V}, \mathbf{S_r} \in \mathbb{R}^{+}, \text{and}, \tag{1}$$

$\mathbf{U} \in \mathbb{R}^{N \times K_d}$, $\mathbf{V} \in \mathbb{R}^{M \times K_c}$ describe the relationship of the latent factors to drugs and cell lines, respectively. The $\mathbf{S_r} \in \mathbb{R}^{K_d \times K_c}$ defines the latent relation between drugs and cell lines.

We use a probabilistic approach to formulate the factorization, which allows us to handle missing values efficiently. We assume a priori that each relation is drawn from Gaussian distribution with precision τ. The likelihood function for the observed data is:

$$p(\mathbf{R}|\mathbf{U}, \mathbf{S_r}, \mathbf{V}, \tau_r) = \prod_{i,j} \mathcal{N}(\mathbf{R}_{ij}; \mathbf{U}_i \cdot \mathbf{S_r} \cdot \mathbf{V}_j^{\mathsf{T}}, \tau_r^{-1}). \tag{2}$$

We choose priors that allow us to constrain latent matrices to be non-negative and permit an efficient inference procedure. Thus, the priors for the latent matrices are chosen to be exponentially distributed with scales λ_{k_d}, and λ_{k_c},

$$\mathbf{U}_{ik_d} \sim \text{Exp}(\lambda_{k_d}), \quad \mathbf{V}_{ik_c} \sim \text{Exp}(\lambda_{k_c}), \quad \mathbf{S_r} \sim \text{Exp}(\lambda_{s_r}) \tag{3}$$

where $k_d \in \{1, \ldots, K_d\}$ and $k_c \in \{1, \ldots, K_c\}$. The model is formulated with conjugate priors where noise variance is chosen as gamma distribution with shape α_r and scale β_r,

$$\tau_r \sim \mathcal{G}(\tau_r; \alpha_r, \beta_r). \tag{4}$$

TABLE 1
The List of Symbols and Notations Used in This Paper

Symbol	Description
$\mathbf{R}$	Drug response matrix (main input), $\mathbf{R} \in \mathbb{R}^{N \times M}$
$\mathbf{D}$	Drug similarity matrix, $\mathbf{D} \in \mathbb{R}^{N \times N}$
$\mathbf{P}$	Gene expression matrix, $\mathbf{P} \in \mathbb{R}^{M \times L}$
$\mathbf{G}$	Protein-protein interaction matrix, $\mathbf{G} \in \mathbb{R}^{L \times L}$
$\mathbf{T}$	Drug-target interaction matrix, $\mathbf{T} \in \mathbb{R}^{L \times N}$
$\mathbf{U}$	Low-rank representation of drugs, $\mathbf{U} \in \mathbb{R}^{N \times K_d}$
$\mathbf{V}$	Low-rank representation of cell lines, $\mathbf{V} \in \mathbb{R}^{M \times K_c}$
$\mathbf{H}$	Low-rank representation of genes, $\mathbf{H} \in \mathbb{R}^{L \times K_g}$
$\mathbf{S_r}$	Low-rank relation matrix of drugs and cell lines, $\mathbf{S_r} \in \mathbb{R}^{K_d \times K_c}$
$\mathbf{S_p}$	Low-rank relation matrix of cell lines and genes, $\mathbf{S_p} \in \mathbb{R}^{K_c \times K_g}$
$\mathbf{S_d}$	Low-rank similarity matrix of drugs, $\mathbf{S_d} \in \mathbb{R}^{K_d \times K_d}$
$\mathbf{S_g}$	Low-rank similarity matrix of genes, $\mathbf{S_g} \in \mathbb{R}^{K_g \times K_g}$
$\mathbf{S_t}$	Low-rank relation matrix of drugs and genes, $\mathbf{S_t} \in \mathbb{R}^{K_g \times K_d}$
$\lambda_\mathbf{k}$	Set of prior parameters of latent factors, $\lambda_k = \{\lambda_{k_d}, \lambda_{k_c}, \lambda_{k_g}\}$
w	Set of importance weights; $w = \{w_r, w_p, w_d, w_t, w_g\}$
τ	Set of noise parameters; $\tau = \{\tau_r, \tau_p, \tau_d, \tau_t, \tau_g\}$ for all data
M	Main block includes main prediction data
F	Feature block includes feature matrices
S	Similarity block includes similarity matrices

3.1.4 Selecting the Rank

Bayesian settings help to seek the exact rank automatically in contrast to traditional matrix factorization methods by having a prior, and when integrating out parameters, the model ends up realizing that some components have zero contribution to the result. Instead of performing model selection to find the number of ranks for latent matrices, we define an upper bound, and the model determines the correct number of components. We define hyperpriors over prior parameters which are shared by columns of matrices to perform automatic model selection;

$$\lambda_{k_d} \sim \mathcal{G}(\lambda_{k_d}; \alpha_{k_d}, \beta_{k_d}), \quad \lambda_{k_c} \sim \mathcal{G}(\lambda_{k_c}; \alpha_{k_c}, \beta_{k_c}), \tag{5}$$

as used by [18]. If prior has a low value, the entire column will be activated or eliminated if prior has a high value.

3.1.5 Inference

Given the observed measurements of cell lines, drugs, and genomic features, the posterior distribution of the model parameters is computed via the Bayes theorem. Since the model has been formulated with conjugate priors, Gibbs sampling can be conveniently used to sample new values for each parameter from their conditional distribution of given data and the current values of the other parameters. Derivation of conditional distributions from posterior distribution is straightforward due to using conjugate priors. The detailed iterative process of Gibbs sampling is summarized in Algorithm 1.

Algorithm 1. Gibbs Sampling Algorithm for Drug Response Prediction

Input: Drug response matrix $\mathbf{R}$
Output: Approximated drug response matrix

Initialize model parameters: $\mathbf{U}_0, \mathbf{S_r}_0, \mathbf{V}_0, \lambda_{k_d^0}, \lambda_{s_r^0}, \lambda_{k_c^0}$
for each iteration: i = 1, ..., T **do**
  Sample model hyperparameters:
  $\lambda_{k_d}^t \sim p(\lambda_{k_d}|\mathbf{U}^t, \lambda_{k_d^0})$
  $\lambda_{s_r}^t \sim p(\lambda_{s_r}|\mathbf{S_r}^t, \lambda_{s_r^0}))$
  $\lambda_{k_c}^t \sim p(\lambda_{k_c}|\mathbf{V}^t, \lambda_{k_c^0}))$
  for each drug, i = 1,...,N **do**
    $\mathbf{U}_i^{t+1} \sim p(\mathbf{U}_i|\mathbf{R}, \mathbf{S}^t, \mathbf{V}^t, \lambda_{k_d}^t)$
  end for
  for each relation, $k = 1, \ldots, K_d$ **do**
    $\mathbf{S_r}_k^{t+1} \sim p(\mathbf{S_r}_k|\mathbf{R}, \mathbf{U}^t, \mathbf{V}^t, \lambda_{s_r}^t)$
  end for
  for each cell lines, j = 1,...,M **do**
    $\mathbf{V}_j^{t+1} \sim p(\mathbf{V}_i|\mathbf{R}, \mathbf{S}^t, \mathbf{U}^t, \lambda_{k_c}^t)$
  end for
end for

3.1.6 Hybrid Matrix Factorization

The purpose of our model is to predict the missing responses of drugs to given cell lines and unseen drugs for given multiple cancer cell lines by incorporating prior information. In order to infer drug responses of cancer cell lines and improve the accuracy, we apply Bayesian hybrid matrix factorization (HMF) to integrate several data sets concurrently as side information [19].

HMF considers heterogeneous integration over three types of data blocks, each being a set of matrices: 1) main block $\mathcal{M}$ has the main matrices to be considered, 2) similarity block $\mathcal{S}$ has similarity matrices, and 3) feature block $\mathcal{F}$ has matrices, each relating two entities, while these matrices are not in $\mathcal{M}$. HMF is suitable for our problem setting of integrating multiple data matrices, at the same time by sharing factors between data sets. Additionally, sharing latent matrices can be effectively used for data integration and improve the factorization [20].

Let us show an example of HMF, with main block $\mathcal{M} = \{\mathbf{R}, \mathbf{P}\}$ and similarity block $\mathcal{S} = \{\mathbf{D}, \mathbf{G}\}$. We then simultaneously factorize these four matrices, where each matrix is factorized into a product of three non-negative low-dimensional matrices. Factorization of $\mathbf{R}$ is already given in (1), and then the rest three matrices, $\mathbf{P}$, $\mathbf{D}$, and $\mathbf{G}$, can be factorized as follows:

$$\mathbf{P} \approx \mathbf{V}\mathbf{S_p}\mathbf{H}^\mathsf{T}, \quad \mathbf{D} \approx \mathbf{U}\mathbf{S_d}\mathbf{U}^\mathsf{T}, \quad \mathbf{G} \approx \mathbf{H}\mathbf{S_g}\mathbf{H}^\mathsf{T}, \tag{6}$$

with the additional constraint that all latent factors are non-negative. Thus you can easily see that four input matrices can be factorized into only three latent factors, $\mathbf{V}$, $\mathbf{H}$, and $\mathbf{U}$ except $\mathbf{S}_*$.

3.1.7 Importance Weights

We learn the importance of each data to investigate the contribution to the prediction task. In this way, we can ensure that no single side data source will dominate the prediction task, and the method will find a solution that better fits all data sets. The importances are learned by modifying the likelihood functions of HMF to include a set of importance weights w_r, w_d, w_p, w_g and learning them from data. After adding the weights, the likelihoods are,

$$\begin{aligned} \mathbf{R} &\sim \mathcal{N}(\mathbf{R}; \mathbf{U} \cdot \mathbf{S_r} \cdot \mathbf{V}^\mathsf{T})^{w_r}, \quad \mathbf{D} \sim \mathcal{N}(\mathbf{D}; \mathbf{U} \cdot \mathbf{S_d} \cdot \mathbf{U}^\mathsf{T})^{w_d} \\ \mathbf{P} &\sim \mathcal{N}(\mathbf{P}; \mathbf{V} \cdot \mathbf{S_p} \cdot \mathbf{H}^\mathsf{T})^{w_p}, \quad \mathbf{G} \sim \mathcal{N}(\mathbf{G}; \mathbf{H} \cdot \mathbf{S_g} \cdot \mathbf{H}^\mathsf{T})^{w_g}. \end{aligned} \tag{7}$$

3.2 Integrating Side Information

We propose a data integration framework to improve the efficiency of the prediction of anticancer drug responses in cell lines by incorporating heterogeneous data about observed relationships among cell lines, drugs, and genes. Our framework incorporates those multiple relationships as matrices into the data blocks of HMF. The model assumes that the drug responses and all side information sources are conditionally independent given the parameters so that the likelihood can be written as the product over these sources. In particular, to examine the importance of each data matrix, we integrate multiple matrices in a step-wise manner which, starting with $\mathbf{R}$, adds one data matrix one by one, in the order of $\mathbf{D}$, $\mathbf{P}$, $\mathbf{G}$ and $\mathbf{T}$. At each step, we examine the importance of the added data matrix by checking the predictive performance. The detailed relation between the datasets and their parameters, priors and hyper-priors can be seen in Fig. 2 as a graphical illustration for a better understanding.

3.2.1 Data Integration: Step-Wise Methods

Our framework is a step-wise workflow of gradually integrating multiple data matrices, where at each step we explore the importance of each added matrix through the importance weight, and each step is based on matrix tri-factorization (MTF) or matrix bi-factorization (MF) of HMF. We call the method DIVERSE (for *Bayesian Data IntegratiVE learning for drug ResponSE prediction*), particularly DIVERSE3 (*Importance Weight matrix-Tri-Factorization*) or DIVERSE2 (*Importance Weight matrix-Bi-Factorization*), depending on the manner of factorization. Fig. 3 shows a schematic picture of our framework, which for five data auxiliary data matrices produces five prediction methods having progressively more auxiliary data.

(1) *Incorporating drug similarity data (DIVERSE3-D)*
Drug similarity is one of the most commonly used side information sources to improve drug response prediction. We start by adding $\mathbf{D}$ to $\mathbf{R}$ to demonstrate the method yields comparable results to earlier methods using the same two data sources. For given likelihood functions of drug response and drug similarity matrices (7), we integrate the data matrices in the total likelihood function as follows,

$$p(\theta|\mathbf{R}, \mathbf{D}) \propto p(\theta)p(\mathbf{R}|\mathbf{U}, \mathbf{S_r}, \mathbf{V}, \tau_r)^{w_r} p(\mathbf{D}|\mathbf{U}, \mathbf{S_d}, \tau_d)^{w_d}. \tag{8}$$

where θ denotes all parameters, $p(\theta)$ is the prior, and the two last terms the likelihoods for the two data sources, weighted by data set specific weights w_r and w_d.

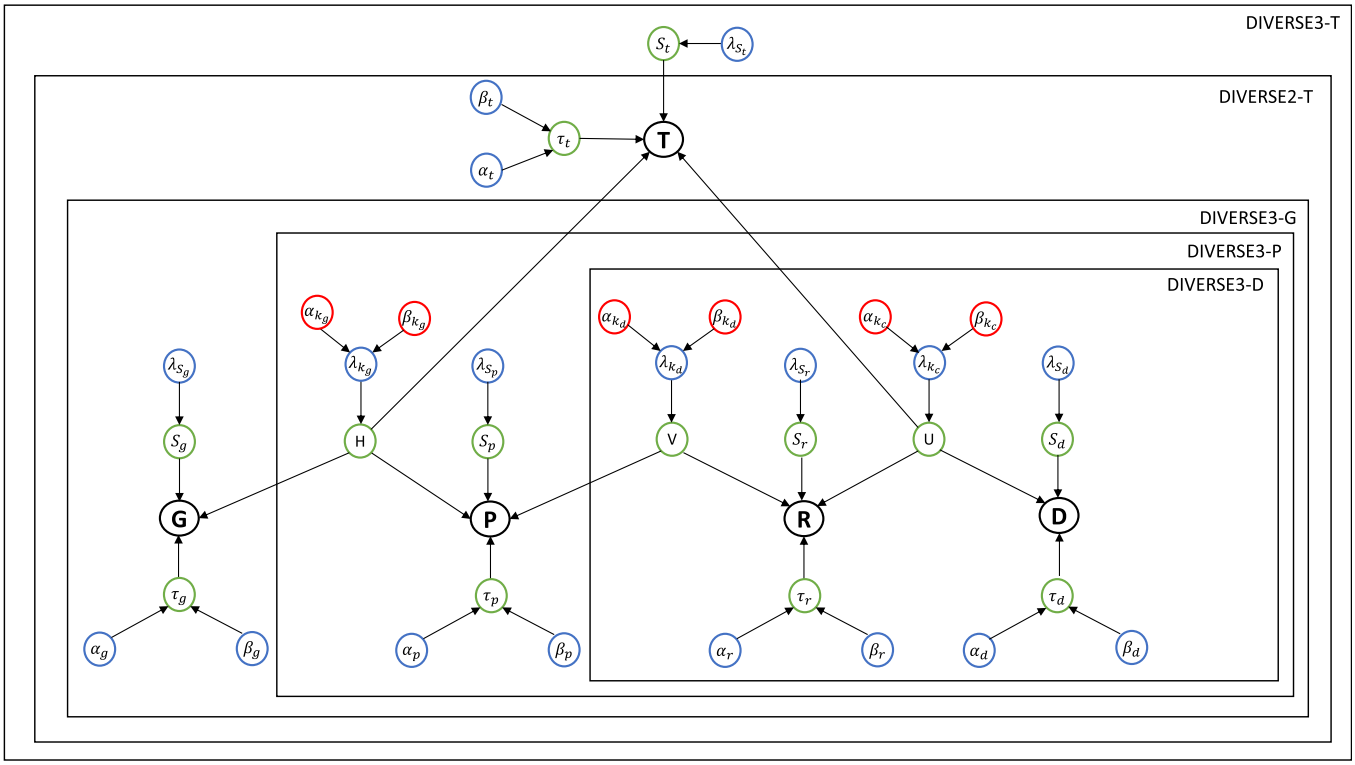

Fig. 2. Graphical illustration of the DIVERSE method used for prediction of drug responses. The figure demonstrates given matrices, shared latent variables with their priors any hyper-priors. Each plate shows a different version of DIVERSE that depends on the manner of factorization and the integrated side information. In particular, black colored nodes denote the matrices, green nodes represent the parameters of the matrices, blue ones are prior to the projection matrices, and the red nodes represent the hyper-priors of the model. See text for more details.

(2) *Incorporating gene expression data (DIVERSE3-P)*

Gene expression has also been utilized for the prediction of drug responses since a considerable amount of gene expression data has become publicly available. When combining gene expression data **P** with DIVERSE3-D, we can write the posterior probability as

$$p(\theta|\mathbf{R},\mathbf{D},\mathbf{P}) \propto p(\theta)p(\mathbf{R}|\mathbf{U},\mathbf{S_r},\mathbf{V},\tau_r)^{w_r}p(\mathbf{D}|\mathbf{U},\mathbf{S_d},\tau_d)^{w_d} \\ p(\mathbf{P}|\mathbf{V},\mathbf{S_p},\mathbf{H},\tau_p)^{w_p}. \tag{9}$$

(3) *Incorporating protein-protein interaction data (DIVERSE3-G)*

Protein-protein interaction is another significant source that researchers have recently started to incorporate to predicting cell line-drug associations. We integrate this information into DIVERSE3-P. The posterior distribution of the four data sets is

$$p(\theta|\mathbf{R},\mathbf{D},\mathbf{P},\mathbf{G}) \propto p(\theta)p(\mathbf{R}|\mathbf{U},\mathbf{S_r},\mathbf{V},\tau_r)^{w_r}p(\mathbf{D}|\mathbf{U},\mathbf{S_d},\tau_d)^{w_d} \\ p(\mathbf{P}|\mathbf{V},\mathbf{S_p},\mathbf{H},\tau_p)^{w_p}p(\mathbf{G}|\mathbf{H},\mathbf{S_g},\tau_g)^{w_g}. \tag{10}$$

(4) *Incorporating drug-target interaction data (DIVERSE3/2-T)*

In the last step, for a given drug-target interaction data set, we have two different scenarios that decompose **T** in different ways

$$\mathbf{T} \approx \mathbf{H}\mathbf{S_t}\mathbf{U}^\mathsf{T} \quad \text{or} \quad \mathbf{T} \approx \mathbf{H}\mathbf{U}^\mathsf{T}, \quad \text{such that} \quad \mathbf{U},\mathbf{H},\mathbf{S_t} \in \mathbb{R}^+. \tag{11}$$

The idea behind these two distinct ways is that so far we have repeatedly used MTF, and then now we can try two cases: we 1) keep using MTF, or 2) switch to MF, where latent factors **H** and **U** can be more regularized by **T** than MTF, which might be useful for prediction. This different experimental setup will reveal the flexibility of the model. In other words, by switching to MF, which has fewer parameters, the decomposition can be more regularized.

(1) *DIVERSE3-T*

We decompose **T** into three matrices so that this factorization will have an advantage of using interactions between the two latent vector spaces. Especially because this data is binary, simultaneous factorization would be preferable. In this scenario, blocks are given as $\mathcal{M} = \{\mathbf{R},\mathbf{P},\mathbf{T}\}$ and $\mathcal{S} = \{\mathbf{D},\mathbf{G}\}$. We modify the posterior such as,

$$p(\theta|\mathbf{R},\mathbf{D},\mathbf{P},\mathbf{G}) \propto p(\theta)p(\mathbf{R}|\mathbf{U},\mathbf{S_r},\mathbf{V},\tau_r)^{w_r}p(\mathbf{D}|\mathbf{U},\mathbf{S_d},\tau_d)^{w_d} \\ p(\mathbf{P}|\mathbf{V},\mathbf{S_p},\mathbf{H},\tau_p)^{w_p}p(\mathbf{G}|\mathbf{H},\mathbf{S_g},\tau_g)^{w_g} \\ *p(\mathbf{T}|\mathbf{H},\mathbf{S_t},\mathbf{U},\tau_t)^{w_t}. \tag{12}$$

(2) *DIVERSE2-T*

In the second scenario, we use MF which requires fewer parameters for T, and uses latent factors of drug and gene entities obtained from the main block

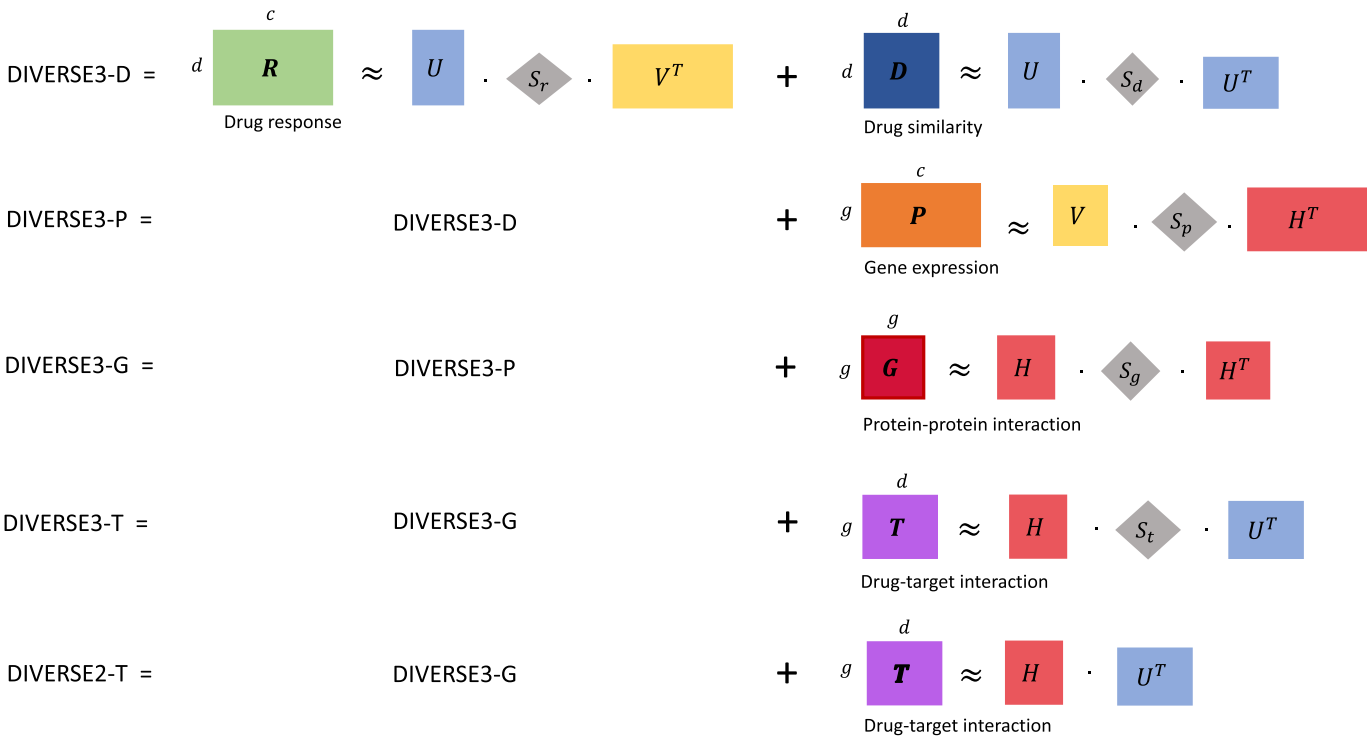

Fig. 3. Overview of our systematic framework, DIVERSE, of integrating multiple data sets: importance weight tri-(or bi-)matrix factorization. We start with adding **D** to **R** (first row: DIVERSE3-D). We then add **P** to DIVERSE3-D (second row: DIVERSE3-P). Similarly we add **G** to DIVERSE3-P (third row: DIVERSE3-G) and **T** to DIVERSE3-G (fourth row: DIVERSE3-T). Another option of the last addition is bi-matrix factorization, and this is the last row: DIVERSE2-T.

$$p(\theta|\mathbf{R},\mathbf{D},\mathbf{P},\mathbf{G}) \propto p(\theta)p(\mathbf{R}|\mathbf{U},\mathbf{S_r},\mathbf{V},\tau_r)^{w_r}p(\mathbf{D}|\mathbf{U},\mathbf{S_d},\tau_d)^{w_d} p(\mathbf{P}|\mathbf{V},\mathbf{S_p},\mathbf{H},\tau_p)^{w_p}p(\mathbf{G}|\mathbf{H},\mathbf{S_g},\tau_g)^{w_g} *p(\mathbf{T}|\mathbf{H},\mathbf{U},\tau_g)^{w_t}. \quad (13)$$

4 Experimental Evaluation

4.1 Data

We used five publicly available data sources, which consist of measurements on three types of entities: drugs, cell lines, and genes, for predicting the response of cancer cell lines. Fig. 1 shows a conceptual scheme on the relation between data sets. All data sets vary in different ranges, such as drug-target interaction is binary data set, while drug similarity data ranges between [0,100]. For the consistency between integrated data sources, we scaled all data to the range between [0,1]. The statistics of the five data sets we used in our experiments are summarized in Table 2.

TABLE 2
Statistics on Five Data Sets in Our Experiments

	#drugs	#cell lines	#genes	observed	sources
R	255	956	-	0.82	GDSC
D	255	-	-	0.92	Pubchem
P	-	956	232	1	GDSC
G	-	-	232	0.5	STRING
T	255	-	232	0.009	GDSC + Chembl

Note: **R**: *drug response*, **D**: *drug similarity*, **P**: *gene expression*, **G**: *protein-protein interaction*, **T**: *drug-target interaction*.

4.1.1 Drug Response (R)

We obtained drug response data (**R**) from Genomics of Drug Sensitivity in Cancer (GDSC) [21], consisting of IC50 values that measuring drug activity concentration required for 50 percent inhibition (a lower value of IC50 indicates a better sensitivity of a cell line to a given drug). For 265 drugs and 992 cell lines, the data had been log-transformed. After the pre-processing, we obtained 255 drugs and 956 cell lines.

4.1.2 Drug Similarity (D)

Drug similarity (**D**), based on the chemical structural similarity between compounds, is usually used to identify compounds sharing similar biological or chemical activity. We used the PubChem Score Matrix Service [22] for extracting 2D similarity scores of the 255 drug compounds in **R**.

4.1.3 Gene Expression (P)

We used gene expression data (**P**) provided by the GDSC project. The data had been measured with Affymetrix Human Genome U219 Arrays, and normalized by using RMA. We used the genes found both in drug-target interactions and gene expression, resulting in 232 genes with their interactions with 956 cell lines.

4.1.4 Drug-Target Interaction (T)

Drug–target interaction data (**T**) were collected from GDSC and Chembl [23] databases. We extracted drug-target interactions for 255 drugs, which also exist in the drug response matrix and 232 genes, which are found in the gene expression data.

4.1.5 Protein-Protein Interaction (G)

PPI (**G**) is very noisy but might be helpful to understand the behavior of drug responses, since drug effects can be affected by proteins through various networks, such as metabolic pathways. We retrieved protein interactions from STRING [24], which includes physical and functional associations.

4.2 Experiment settings

4.2.1 In-Matrix and Out-of-Matrix Prediction

We empirically evaluated the predictive performance of the five methods presented in Section 3.2.1 by associations between cell lines and drugs. For comparison of these methods, we considered two tasks:

(i) *in-matrix prediction*: we predict missing values in **R**.
(ii) *out-of-matrix prediction*: we predict values for entirely unseen drug response vectors to given cell lines.

4.2.2 Determining Importance Weight

An important and hard problem is to find the best set of importance values, particularly for the method with a larger number of data sets. For this problem, we examined various number of importance values for both in-matrix and out-of-matrix predictions. We used the values in the range of [0,1] with a fixed interval, resulting in that all values are [0.2, 0.4, 0.6, 0.8, 1]. Our way of exploring a set of optimal importance values is a greedy manner. That is, we repeated the following three steps: 1) at one method in Section 3.2.1, we added one data set and then tested each of all five possible importance values, meaning that we conducted five experiments for each method. 2) we adopted the value, which provides the best performance among the five possibilities. 3) we then moved to the next method to add another data set.

4.2.3 Cross-Validation

We conducted five times 5-fold (5x5-fold) cross-validation with different random cross-validation folds where we held a subset of drugs as a test set and trained the model on the rest of the drugs. We predicted the response values for drugs in the test set, by using the trained model.

4.2.4 Performance Evaluation Measures

We evaluated the predictive performance of all methods by using the Spearman correlation coefficient (Sc) and MSE between the observed (true) and the predicted IC50s. We focused on drug-averaged Spearman correlation scores across test drugs since the correlation over all drug responses between true and predicted drug sensitivity scores might overestimate the predictive performance.

4.3 Comparison Methods

We note that our framework allows to integrate five different data sets, while so far to the best of our knowledge, there are no bioinformatics methods, which can incorporate the five different data sets (three entity types) for prediction of drug response values. Thus we were unable to find any competing methods, which can use same data sets as our framework.

We first used two baselines, by following [25], i.e., the mean of training drug response values as a prediction for the unobserved drug responses, where we considered 1) cell-line specific mean (cls-mean) and 2) overall mean (all-mean). We then used two state-of-the-art machine learning approaches: MultiNMF [12] and KRR [6]. Lastly, we used DrugCellNet, which is a straightforward but efficient network interpolation method for drug response prediction. We chose DrugCellNet since DrugCellNet already outperformed standard machine learning methods such as Elastic-Net, random forest and support vector regression in [13]. We note that these three methods (MultiNMF, KRR and DrugCellNet) cannot use the entire five data sets in our experiment, though DIVERSE can handle all these five. Instead, these three methods used only drug response (**R**) and drug similarity (**D**) data sets.

We tuned relevant hyper-parameters of MultiNMF and KRR by using grid search on each training set. We selected the size of dimensionality K of MultiNMF as 10. We followed the the same experimental procedure as described above for the compared methods, i.e., 5x5-fold cross-validation.

4.4 Performance Results

We entirely compared the following ten methods: cls-mean (cell-line specific mean), all-mean (overall mean); MultiNMF, KRR, DrugCellNet (drug response and similarity), DIVERSE3-D, DIVERSE3-P, DIVERSE3-G and finally DIVERSE2-T and DIVERSE3-T considering two cases: i) in-matrix prediction and ii) out-of-matrix prediction. Note that, our primary motivation is to focus on out-of matrix prediction, since this problem is more challenging, and also a more realistic setting, in which unknown drugs are given.

4.4.1 In-Matrix Prediction Performance

Table 3 shows the MSE and Sc of the ten compared methods under 5x5-fold cross-validation. The five methods of DIVERSE achieved significantly smaller MSE than the two baselines and three state-of-the-art methods. The performance of MultiNMF was the second (after DIVERSE) in both MSE and Sc, which might be reasonable, because MultiNMF is also based on NMF, though using drug responses and drug similarity data only. The performance of KRR was worse compared to MultiNMF and DrugCellNet, probably because regression may have negative output values, although all true outputs are here known to be

TABLE 3
MSE and Sc (Average Scores of 5x5 Cross-Validation) of Ten Compared Methods in In-Matrix Prediction

	MSE $\pm$ Std. Dev.	Sc $\pm$ Std. Dev.
cls-mean	0.5226 ± 0.0036	–
all-mean	0.4181 ± 0.0726	–
MultiNMF	0.0209 ± 0.0020	0.4717 ± 0.0041
KRR	0.0625 ± 0.0046	0.0111 ± 0.0250
DrugCellNet	0.0532 ± 0.0002	0.4481 ± 0.0056
DIVERSE3-D	0.0047 ± 0.0001	0.4852 ± 0.0023
DIVERSE3-P	0.0048 ± 0.0008	0.4833 ± 0.0040
DIVERSE3-G	0.0047 ± 0.0001	0.4840 ± 0.0039
DIVERSE2-T	0.0047 ± 0.0001	0.4841 ± 0.0037
DIVERSE3-T	0.0048 ± 0.0001	0.4844 ± 0.0032

Note: Std. Dev. stands for standard deviation.

TABLE 4
MSE and Sc (Average Scores of 5x5 Cross-Validation) of Ten Compared Methods in Out-of-matrix Prediction

	MSE ± Std. Dev.	Sc ± Std. Dev.
cls-mean	0.5227 ± 0.0027	–
all-mean	0.4181 ± 0.0726	–
MultiNMF	0.1581 ± 0.0721	0.1457 ± 0.0180
KRR	0.0764 ± 0.0125	0.2976 ± 0.0361
DrugCellNet	0.0455 ± 0.0044	0.3423 ± 0.0259
DIVERSE3-D	0.0194 ± 0.0049	0.6750 ± 0.0186
DIVERSE3-P	0.0189 ± 0.0049	0.6770 ± 0.0188
DIVERSE3-G	0.0186 ± 0.0035	0.6762 ± 0.0179
DIVERSE2-T	0.0185 ± 0.0040	0.6765 ± 0.0187
DIVERSE3-T	$\mathbf{0.0183 \pm 0.0033}$	$\mathbf{0.6772 \pm 0.0193}$

Note: Std. Dev. stands for standard deviation.

non-negative. Overall, DIVERSE outperformed all compared methods, indicating that DIVERSE would be most robust to predict missing data. On the other hand, the differences between the five methods of DIVERSE were rather unclear. This result implies that adding multiple data sets might not necessarily be so effective for in-matrix prediction.

4.4.2 Out-of-Matrix Prediction Performance

Table 4 shows the MSE and Sc of the ten compared methods under 5x5-fold cross-validation, where the lowest MSE and largest Sc are highlighted in bold. The five methods of DIVERSE again achieved significantly smaller MSE and higher Sc scores than the other five methods. MultiNMF was worst among the existing methods, implying that NMF is ineffective for out-of-matrix prediction though being useful for filling missing values. DrugCellNet was next to DIVERSE in both MSE and Sc. Among the five methods of DIVERSE, starting with DIVERSE3-D, the MSE was decreasing like DIVERSE3-P, DIVERSE3-G, finally resulting in DIVERSE3-T, the smallest value among all ten compared methods. This result indicates that step-wise data set addition of DIVERSE worked well for integrating heterogeneous data sets. Also, this result was confirmed by Sc, where Sc was basically increased by adding more data sets, finally reaching 0.6772, which was again the highest among all ten compared methods.

We can compare our two ways of adding **T**, i.e., DIVERSE3-T and DIVERSE2-T, from a performance perspective. Table 4 shows the MSE of DIVERSE3-T was 0.0183, which was smaller than the MSE of DIVERSE2-T which was 0.185. This difference sounds very slight, and we checked box plots of MSE and Sc, which are shown in Fig. 4, where the left two figures ((a) MSE and (b) Sc) show the case of DIVERSE3-T (DIVERSE3-D → DIVERSE3-P → DIVERSE3-G → DIVERSE3-T) and the right two figures ((c) MSE and (d) Sc) show the case of DIVERSE2-T (DIVERSE3-D → DIVERSE3-P → DIVERSE3-G → DIVERSE2-T). In these figures, the thick black line in each box shows the median, and the mean value of each case is shown by black cross. We observe that the mean value of (a) decreased clearly (particularly at the last DIVERSE3-T), while the decrease of the mean of (c) is rather mild, particularly at the last DIVERSE2-T. Thus we can see DIVERSE3-T achieved a better performance than DIVERSE2-T, implying that matrix tri-factorization is better than matrix bi-factorization here. Eventually, strong regularization by **T** (see the bottom of Fig. 3) might not be so useful.

Finally, we examine the importance weights, which were computed when a data set is newly added in the greedy procedure of DIVERSE. Table 5 shows the importance weights obtained by the best performance case when we added a newly data set (for example **D** for DIVERESE3-D)

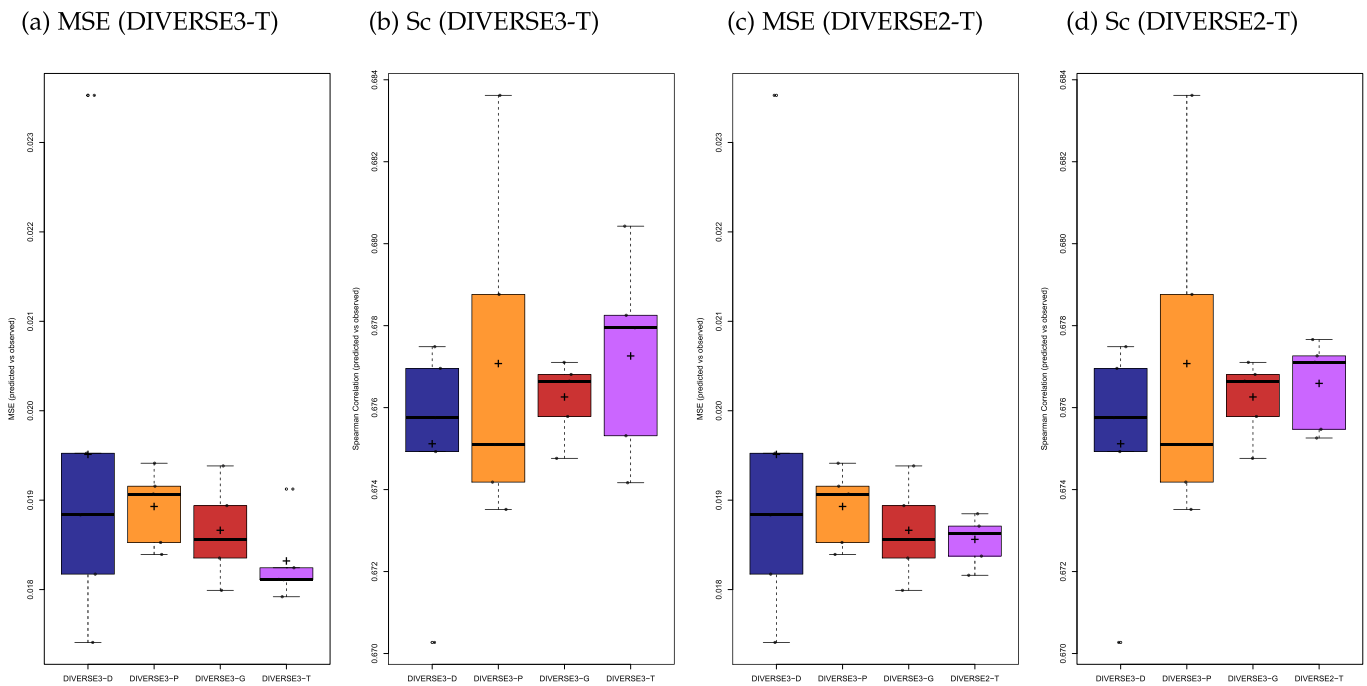

Fig. 4. Performance results of out-of-matrix prediction under 5x5-fold cross validation. (a) MSE and (b) Sc of the case of DIVERSE3-T when **T** is added, and (c) MSE and (d) Sc of the case of DIVERSE2-T when **T** is added. The thick black horizontal line in each box is the median, while black cross of each box is the mean. Thus please use each black cross to examine the results.

TABLE 5
Importance Weights of the Best Performance Case of Each Time in 5x5-Fold Cross-Validation

	1st	2nd	3rd	4th	5th	Average
DIVERSE3-D	0.2	0.2	0.2	0.2	0.2	0.2
DIVERSE3-P	0.8	0.8	0.2	0.8	1	0.68
DIVERSE3-G	1	0.2	0.4	0.6	0.8	0.6
DIVERSE2-T	0.4	0.4	0.6	0.4	1	0.56
DIVERSE3-T	1	1	0.8	0.4	1	0.84

in DIVERSE, for each of the five times 5-fold cross-validation, and also the average over the five times. The highest average importance weight was obtained by DIVERSE3-T, indicating that the importance weight was large when **T**, i.e., drug-target interactions, was added. However, other data sets also had rather large average importance weights, like 0.56 to 0.68, except **D**, i.e., drug similarity, with always 0.2. Interestingly, this result implies that drug similarity might not have been so significant.

4.4.3 Case Study

To find potential drugs for our cancer cell lines from a different perspective, we checked how well data integration worked for prediction improvement for individual cases. The idea here is if we have drugs for which prediction was improved by integrating more data, we can predict whether the drug is useful for a given cell line. From this assessment, we could raise three sample drugs: CUDC-101, Gemcitabine, and SN-38 (known as also Irinotecan), for which primary targets are EGFR/ERBB2, pyrimidine antimetabolite, and TOP1 respectively. Table 6 shows how each version of DIVERSE improved the correlation score between the observed and predicted values of each of the three drugs. These results indicate the existence of highly predictable drugs, and also, as a methodology, our framework of data integration worked on predicting the IC50 values. Furthermore, these results show that integrating biological side information is useful to predict unseen drugs from existing drug screening values and improve efficiency.

To understand the results obtained by DIVERSE3-T more, we trained DIVERSE3-T by using all available data and predicted responses of unseen chemical compounds. The idea is that we may find a new drug, if a positive response is predicted by DIVERSE3-T, even if the observed

TABLE 6
Average MSE and Spearman Correlation Scores Over 5x5-Fold Cross-Validation for Multiple Cancer Cell Lines

	CUDC101		Gemcitabine		SN-38	
	MSE	Sc	MSE	Sc	MSE	Sc
DIVERSE3-D	0.00096	0.916	0.00182	0.930	0.00146	0.861
DIVERSE3-P	0.00099	0.910	0.00203	0.923	0.00122	0.884
DIVERSE3-G	0.00092	0.915	0.00211	0.919	0.00165	0.841
DIVERSE2-T	0.00087	0.922	0.00189	0.929	**0.00119**	**0.894**
DIVERSE3-T	**0.00081**	**0.926**	**0.00138**	**0.948**	0.00121	0.887

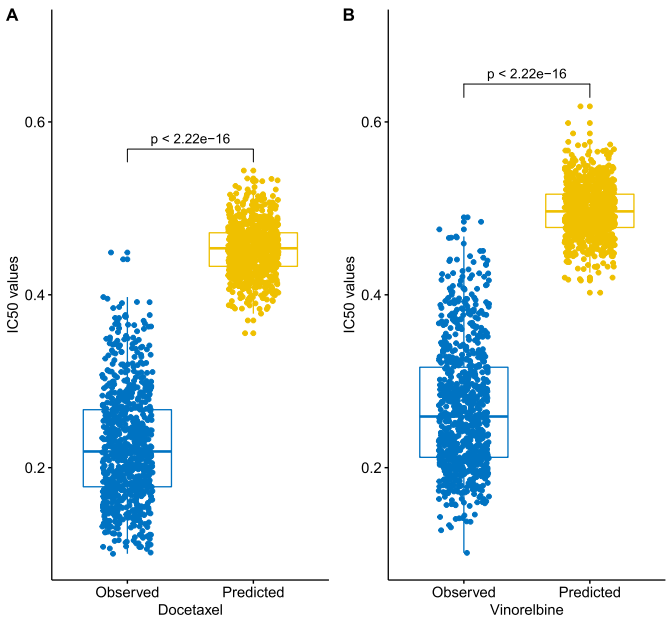

Fig. 5. Box plots of two test drugs from GDSC dataset: A. Docetaxel and B. Vinorelbine. The t-test was used to measure the statistical difference in the mean between the predicted and observed response values of cell lines for each drug.

(true) data are not positives, i.e., negatives. Thus we first chose compounds with different values between the predicted (by DIVERSE3-T) and observed values and then ran a one-sided paired t-test to confirm the significance of the difference in the mean. We then obtained two potential drugs for our cancer cell lines: 1) Docetaxel and 2) Vinorelbine. Fig. 5 shows the distributions of predicted and observed values for these two compounds. For these two drugs, the values predicted by DIVERSE3-T were significantly higher than the observed values, with p-values $< 2.22e-16$. We further checked the literature and some more details of these two drugs can be given as follows: 1) Docetaxel is one of the powerful known drugs which has many interactions with other drugs and also known interactions with mitosis pathway such as TUBB1 and MAP2/MAP4 proteins; where 13 percent of response values are missing. 2) Vinorelbine is a similar drug with Docetaxel where approximately 10 percent of the entries are missing and targeting microtubule destabilizer in the mitosis pathway. Thus these two might be used as cancer drugs.

5 Discussion and Conclusion

Cancer is a complex disease affected by genotypes and associated with other factors including phenotypes, environmental exposures, drugs, and chemical molecules. No single data source can explain the underlying factors and capture complexity. Machine learning methods that combine heterogeneous data from multiple sources have thus emerged as critical, statistical and computational approaches. Although various methods have been developed for anticancer drug response prediction, challenges remain in many aspects, such as choosing the informative data sources suitable for training and testing models, computational approaches that can incorporate many sources efficiently, and deciding how such models are evaluated and validated.

We have proposed DIVERSE, Bayesian matrix factorization with importance weights, a new framework to infer drug-cell line associations incorporating side information collected from different sources. To overcome the challenges which we mentioned above, we focused on integrating as much data as possible, which reaches five data sets in our experiments. In DIVERSE, the data can be systematically integrated in a cascade manner, examining the importance of each incorporated data set. We empirically validated the performance of DIVERSE, comparing with five other methods, including three state-of-the-art methods, under 5x5-fold cross-validation. Experimental results indicate that DIVERSE would clearly be useful for out-of-matrix prediction, which is a real-world setting and much harder than in-matrix prediction. In particular, the results indicate that the performance (MSE) of DIVERSE was smoothly improved by the step-wise addition of more data sets. These advantages of DIVERSE were confirmed by several case studies. Even though our proposed framework has achieved encouraging results, it cannot avoid the following disadvantages. We prefer using Gibbs sampling since it is one of the most robust optimization methods which gives the ability to estimate the full Bayesian posterior, especially for sparse data sets. However, it converges slowly and requires additional samples to estimate the posterior compared to other optimization techniques such as variational Bayes. Second, incorporating more meaningful data effectively could improve the predictive performance such as the drug similarity data which is based on the 2D chemical structural similarity between compounds in our work. Even though 2D features give sufficient features to represent a drug, 3D structure features might also play a crucial role. Similar cases also can be considered for other integrated sources. For example; better identification of gene-drug associations provides a comprehensive understanding of effective treatments for patients [26]. We benefit from one-gene-to-one-drug associations for drug-gene interaction data, however, if different drugs interact with each other and targets "many-genes-to-many drugs", the drug response therapy may be further improved [27], [28].

Predictive performance might be further improved if more informative data sources can be incorporated into our methods, and exploring new data sets would be direct future work. Another direction might be to predict the response of drug combinations by integrating more data since drug combination therapy could provide an effective strategy to overcome drug resistance [29]. There might be another challenge rising here because of the big data problems that require carefully chosen feature selection methods. Also in machine learning, various techniques, including those in deep learning, are continuously being developed. Incorporating such new techniques into our method for better prediction or interpretability would be interesting future work.

Acknowledgments

This work has been supported in part by the Business Finland (1718/31/2014); Academy of Finland (315896); the Finnish Center for Artificial Intelligence FCAI (320181); JST ACCEL (JPMJAC1503); MEXT KAKENHI (16H02868 and 19H04169). The authors would like to thank Aalto University Science-IT for providing computational resources.

References

[1] N. B. La Thangue and D. J. Kerr, "Predictive biomarkers: A paradigm shift towards personalized cancer medicine," *Nat. Rev. Clin. Oncol.*, vol. 8, no. 10, 2011, Art. no. 587.

[2] L. Zhang, X. Chen, N.-N. Guan, H. Liu, and J.-Q. Li, "A hybrid interpolation weighted collaborative filtering method for anti-cancer drug response prediction," *Front. Pharmacol.*, vol. 9, 2018, Art. no. 1017.

[3] B. Güvenç Paltun, H. Mamitsuka, and S. Kaski, "Improving drug response prediction by integrating multiple data sources: matrix factorization, kernel and network-based approaches," *Briefings Bioinf.*, vol. 22, pp. 346–359, 2019.

[4] J. Barretina *et al.*, "The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–607, 2012.

[5] M. P. Menden *et al.* "Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties," *PLoS One*, vol. 8, no. 4, 2013, Art. no. e61318.

[6] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.

[7] J. C. Costello *et al.*, "A community effort to assess and improve drug sensitivity prediction algorithms," *Nat. Biotechnol.*, vol. 32, no. 12, 2014, Art. no. 1202.

[8] H. Liu, Y. Zhao, L. Zhang, and X. Chen, "Anti-cancer drug response prediction using neighbor-based collaborative filtering with global effect removal," *Mol. Ther.-Nucleic Acids*, vol. 13, pp. 303–311, 2018.

[9] N.-N. Guan, Y. Zhao, C.-C. Wang, J.-Q. Li, X. Chen, and X. Piao, "Anticancer drug response prediction in cell lines using weighted graph regularized matrix factorization," *Mol. Ther.-Nucleic Acids*, vol. 17, pp. 164–174, 2019.

[10] M. Ammad-ud din *et al.*, "Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization," *Bioinformatics*, vol. 32, no. 17, pp. i455–i463, 2016.

[11] L. Wang, X. Li, L. Zhang, and Q. Gao, "Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization," *BMC Cancer*, vol. 17, no. 1, 2017, Art. no. 513.

[12] N. Fujita, S. Mizuarai, K. Murakami, and K. Nakai, "Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses," *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, 2018.

[13] N. Zhang, H. Wang, Y. Fang, J. Wang, X. Zheng, and X. S. Liu, "Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model," *PLoS Comput. Biol.*, vol. 11, no. 9, 2015, Art. no. e1004498.

[14] Z. Stanfield, M. Coşkun, and M. Koyutürk, "Drug response prediction as a link prediction problem," *Sci. Rep.*, vol. 7, 2017, Art. no. 40321.

[15] F. Zhang, M. Wang, J. Xi, J. Yang, and A. Li, "A novel heterogeneous network-based method for drug response prediction in cancer cell lines," *Sci. Rep.*, vol. 8, no. 1, pp. 1–9, 2018.

[16] A. Cichonska *et al.*, "Learning with multiple pairwise kernels for drug bioactivity prediction," *Bioinformatics*, vol. 34, no. 13, pp. i509–i518, 2018.

[17] H. Sharifi-Noghabi, O. Zolotareva, C. C. Collins, and M. Ester, "Moli: Multi-omics late integration with deep neural networks for drug response prediction," *Bioinformatics*, vol. 35, no. 14, pp. i501–i509, 2019.

[18] V. Y. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization with the/spl beta/-divergence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1592–1605, Jul. 2013.

[19] T. Brouwer and P. Lio, "Bayesian hybrid matrix factorisation for data integration," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, 2017, pp. 557–566.

[20] D. Zhang, S. Chen, and Z.-H. Zhou, "Two-dimensional non-negative matrix factorization for face representation and recognition," in *Proc. Int. Workshop Anal. Model. Faces Gestures*, 2005, pp. 350–363.

[21] W. Yang *et al.*, "Genomics of drug sensitivity in cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D955–D961, 2012.

[22] S. Kim *et al.*, "Pubchem substance and compound databases," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1202–D1213, 2015.

[23] A. Gaulton *et al.*, "The chembl database in 2017," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D945–D954, 2017.

[24] D. Szklarczyk *et al.*, "The string database in 2011: Functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Res.*, vol. 39, no. suppl_1, pp. D561–D568, 2010.
[25] M. Ammad-Ud-Din *et al.*, "Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization," *J. Chem. Inf. Model.*, vol. 54, no. 8, pp. 2347–2359, 2014.
[26] J. Chen, H. Peng, G. Han, H. Cai, and J. Cai, "HOGMMNC: A higher order graph matching with multiple network constraints model for gene–drug regulatory modules identification," *Bioinformatics*, vol. 35, no. 4, pp. 602–610, 2019.
[27] J. Huang, J. Chen, B. Zhang, L. Zhu, and H. Cai, "Evaluation of gene–drug common module identification methods using pharmacogenomics data," *Briefings Bioinf.*, Jun. 2020, doi: 10.1093/bib/bbaa087.
[28] J. Cai, H. Cai, J. Chen, and X. Yang, "Identifying "many-to-many" relationships between gene-expression data and drug-response data via sparse binary matching," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 17, no. 1, pp. 165–176, Jan./Feb. 2020.
[29] X. Chen, B. Ren, M. Chen, Q. Wang, L. Zhang, and G. Yan, "NLLSS: predicting synergistic drug combinations based on semi-supervised learning," *PLoS Comput. Biol.*, vol. 12, no. 7, 2016, Art. no. e1004975.

Betül Güvenç Paltun (Senior Member, IEEE) received the MSc degree in computational science and engineering from Bogazici University, Istanbul, Turkey, in 2016. She is currently working toward the PhD degree at the Probabilistic Machine Learning Group, Aalto University, Espoo, Finland. Her current research interests include the application of machine learning techniques and bioinformatics.

Samuel Kaski (Senior Member, IEEE) received the DSc (PhD) degree in computer science from the Helsinki University of Technology, Espoo, Finland, in 1997. He is currently a professor with Aalto University, Espoo, Finland and the University of Manchester, UK, and the director of the Finnish Center for Artificial Intelligence FCAI and the ELLIS Unit Helsinki. His research interests include probabilistic machine learning and its applications in AI, computational biology, digital health, user interaction, and brain signal analysis.

Hiroshi Mamitsuka received the BS degree in biophysics and biochemistry, the ME degree in information engineering, and the PhD degree in information sciences from the University of Tokyo, Tokyo, Japan, in 1988, 1991, and 1999, respectively. He is currently a joint professor with the Bioinformatics Center, Institute for Chemical Research, Kyoto University and Aalto University. His research interests include machine learning, data mining and their applications in bioinformatics and chemoinformatics.