# DIVERSE: Bayesian Data IntegratiVE learning for precise drug ResponSE prediction

**Betül Güvenç Paltun** [1] **Samuel Kaski** [1 2] **Hiroshi Mamitsuka** [1 3]

## 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 which are made difficult by 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[1], a Bayesian importance-weighted matrix factorization framework to predict drug responses from a newly combined data, which we assembled from five independent data sources. DIVERSE integrates the data sources systematically, in a step-wise manner to decide which information is useful to be incorporated and how significant the information is for the prediction task. The approach enables DIVERSE to predict values for entirely unseen drug response vectors to given cell lines by leveraging from previous experiments performed on similar drugs and cell lines; thereby, DIVERSE achieves highly accurate predictions despite heterogeneous data sources. The implementation of our method[2] is available online.

## 1. Introduction

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 (Güvenç Paltun et al., 2021). 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 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 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 data set one 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.

## 2. Methods

DIVERSE allows predicting drug responses of cancer cell lines by incorporating information from heterogeneous data sets. DIVERSE consists of two key elements: 1) Bayesian

[1]Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Espoo, Finland [2]Department of Computer Science, The University of Manchester, United Kingdom [3]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Japan. Correspondence to: Betül Güvenç Paltun <betul.guvenc@aalto.fi>.

The 2021 ICML Workshop on Computational Biology. Copyright 2021 by the author(s).

[1] http://dx.doi.org/10.1109/TCBB.2021.3065535

[2] https://github.com/bguvenc/DIVERSE

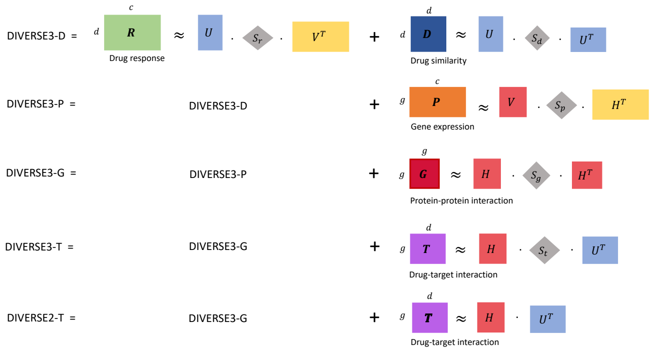

*Figure 1.* Overview of our systematic framework, DIVERSE, of integrating multiple data sets: importance weight tri-(or bi-)matrix factorization. We start with adding $\mathbf{D}$ to $\mathbf{R}$ (first row: DIVERSE3-D). We then add $\mathbf{P}$ to DIVERSE3-D (second row: DIVERSE3-P). Similarly we add $\mathbf{G}$ to DIVERSE3-P (third row: DIVERSE3-G) and $\mathbf{T}$ 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

*Table 1.* 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 | **0.0183 ± 0.0033** | **0.6772 ± 0.0193** |

Note: Std. Dev. stands for standard deviation. cls-mean (cell-line specific mean), all-mean (overall mean).

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) Hybrid matrix factorization (HMF) model (Brouwer & Lió, 2017) to simultaneously integrate heterogeneous data sets. This combination of methods is new for predicting drug sensitivity.

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. Additional given inputs are drug similarity, gene expression, protein-protein interaction, drug-target interaction and cell line-drug interaction data sources.

Given the observed measurements of cell lines, drugs, and the 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.

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 or matrix bi-factorization 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. Figure 1 shows a schematic picture of our framework, which for five data auxiliary data matrices produces five prediction methods having progressively more auxiliary data.

## 3. Results and Conclusion

We empirically validated the performance of DIVERSE, comparing with five other methods, including three state-of-the-art methods, under 5x5-fold cross-validation. Table 1 shows the MSE and Sc of the ten compared methods, where the lowest MSE and largest Sc are highlighted in bold. The five methods of DIVERSE 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.

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. however, 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.

## References

Brouwer, T. and Lió, P. Bayesian hybrid matrix factorisation for data integration. In *Artificial Intelligence and Statistics*, pp. 557–566. PMLR, 2017.

Güvenç Paltun, B., Mamitsuka, H., and Kaski, S. Improving drug response prediction by integrating multiple data sources: matrix factorization, kernel and network-based approaches. *Briefings in bioinformatics*, 22(1):346–359, 2021.