This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2909908, IEEE/ACM Transactions on Computational Biology and Bioinformatics



# SL$^2$MF: Predicting Synthetic Lethality in Human Cancers via Logistic Matrix Factorization

Yong Liu, Min Wu*, Chenghao Liu, Xiao-Li Li, and Jie Zheng*

**Abstract**—Synthetic lethality (SL) is a promising concept for novel discovery of anti-cancer drug targets. However, wet-lab experiments for detecting SLs are faced with various challenges, such as high cost, low consistency across platforms or cell lines. Therefore, computational prediction methods are needed to address these issues. This paper proposes a novel SL prediction method, named SL$^2$MF, which employs logistic matrix factorization to learn latent representations of genes from the observed SL data. The probability that two genes are likely to form SL is modeled by the linear combination of gene latent vectors. As known SL pairs are more trustworthy than unknown pairs, we design importance weighting schemes to assign higher importance weights for known SL pairs and lower importance weights for unknown pairs in SL$^2$MF. Moreover, we also incorporate biological knowledge about genes from protein-protein interaction (PPI) data and Gene Ontology (GO). In particular, we calculate the similarity between genes based on their GO annotations and topological properties in the PPI network. Extensive experiments on the SL interaction data from SynLethDB database have been conducted to demonstrate the effectiveness of SL$^2$MF.

**Index Terms**—Synthetic lethality, machine learning, logistic matrix factorization, importance weighting, human cancers.

## 1 Introduction

A complex disease like cancer is unlikely caused by the defect of only one gene. The understanding of genetic interactions, therefore, is becoming more important in cancer biology and medicine [1]. A prominent type of genetic interaction, called synthetic lethality, has drawn much attention in the field of cancer therapeutics [2], [3]. A pair of genes is called synthetic lethality (SL) if the defect of a single gene will not affect the cell viability, whereas the defects of both genes will cause cell death or significant impairment of cell fitness. Therefore, targeting a nonessential SL partner gene of a cancer-specific mutated gene would selectively kill (or prohibit the proliferation of) the cancer cells but spare normal cells. With the availability of "omics" technologies and high-throughput cancer genomics data, the SLs in the human genome promise to be a gold mine for novel discovery of anti-cancer drug targets. Indeed, both wet-lab screening and computational data mining of SLs in the genomes of human as well as model animal species (*e.g.* yeast) are under intensive research.


- *Yong Liu is currently with Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), and Alibaba-NTU Singapore Joint Research Institute, Nanyang Technological University, Singapore. Email: stephenliu@ntu.edu.sg.*
- *Jie Zheng is currently with School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China. Part of this work was performed while he was an Assistant Professor at School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798. Email: zhengjie@shanghaitech.edu.cn.*
- *Min Wu and Xiaoli Li are currently with the Institute for Infocomm Research (I$^2$R), A*Star, Singapore 138632. Email: {wumin, xlli}@i2r.a-star.edu.sg.*
- *Chenghao Liu is currently with School of Information Systems, Singapore Management University, Singapore. Email: twinsken@gmail.com.*
- ** Corresponding authors*


*Manuscript received xxx, 2018; revised xxx, 2018.*

High-throughput wet-lab screenings have been conducted to search for SLs genome-wide, using the following technologies. First, chemical libraries are used to identify inhibitors of gene or metabolic activities that can kill cancer cells selectively [4]. Secondly, pooled RNAi screens (using siRNA or shRNA libraries), which target the gene expression at the mRNA level, have been widely adopted for SL detection [5], [6]. Considering the complementary strengths of the chemical and RNAi screening technologies, the two types of technologies are sometimes integrated into the more comprehensive approach of chemical-genetic screening [7]. Thirdly, the emerging CRISPR-based genome editing technology has been recently employed to screen for essential genes and SLs [8], [9]. It is expected that this new technology can increase the accuracy and power of screening than the aforementioned technologies. However, the wet-lab screenings for finding SLs are still faced with different challenges, *e.g.*, high cost, off-target effects, lack of consistency across different platforms or cell lines, and unclear mechanisms. Therefore, computational methods for predicting SLs would be useful complements to the wet-lab screenings.

In the literature, various computational methods have been proposed for SL prediction in recent decade [10], [11]. These methods can be classified into the following three categories. The methods in the first category are *in silico* knockouts in metabolic networks. As genome-wide metabolic networks for human and model species are available, single- and double-knockout of genes can be simulated in these networks. By running flux-balance analysis (FBA), the phenotypic cellular effects of these knockouts can be estimated. Based on such metabolic modeling, researchers have carried out predictions for essential and SL genes in yeast [12], *C.elegans* [13], human cancer cell lines [14], as well as other species [15]. The second category refers to knowledge-based methods [16],

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2909908, IEEE/ACM Transactions on Computational Biology and Bioinformatics



[17], [18], which predict SLs based on the knowledge or hypotheses about SL interactions, *e.g.*, SL genes tend to be co-expressed, and SL genes are likely to have similar topological properties in biological networks. For example, network properties, *e.g.*, graph centrality [19], network flow [20], and connectivity homology [21], are widely explored for SL prediction. Based on the hypotheses that SL genes are often co-expressed and seldom co-mutated, DAISY [16] applied three independent procedures for predicting SLs from SCNA (somatic copy number alternation), shRNA and gene expression profiles. Similar to DAISY, MiSL [18] also analysed mutation, copy number and gene expression for SL prediction. The third category includes supervised machine learning methods that build up classification models based on existing SL data to predict novel SL pairs. As a large amount of SL data are available in yeast, various classification models, *e.g.*, decision tree [22], MLE [23], and ensemble classifiers [24], [25], have been studied for yeast SL prediction.

However, existing computational methods for SL prediction have the following limitations when applied to human cancer genome. First, existing supervised learning methods [22], [23], [24], [25] were proposed to work on known SL data in yeast instead of human. Without sufficient human SL data for training the models, an alternative way to predict SL for human is to apply the comparative genomics approaches [11], [25], [26]. However, human and yeast are evolutionarily distant from each other. Therefore, the comparative analysis between human and yeast for SL prediction may not be reliable. Secondly, knowledge-based methods [16], [18], [19], [21] leverage on the knowledge about the biological networks and other genomic data (*e.g.*, mutation and gene expression profiles). However, they do not employ much information about underlying mechanisms of known SL data in human.

Recently, a comprehensive database called SynLethDB for human SLs becomes publicly available [27]. As such, supervised learning methods can be applied on SynLethDB for predicting human SLs. On the other hand, matrix factorization techniques, which were developed for recommendation problems [28], [29], have been successfully applied for various bioinformatics tasks, *e.g.*, PPI prediction [30], drug-target interaction prediction [31], and drug response prediction [32]. Motivated by these approaches, in this paper, we have proposed a novel matrix factorization model, named SL$^2$MF, for SL prediction. Differing from existing methods for human SL prediction, SL$^2$MF employs logistic matrix factorization (LMF) [33] to model the probability that a pair of genes are likely to have SL interaction. Specifically, a latent vector is assigned to each gene to describe its properties learnt from the data. The probability of two genes to be SL is defined as a logistic function of their latent vectors. To further enhance the prediction accuracy, SL$^2$MF has been extended to integrate the knowledge from PPI networks and Gene Ontology (GO) annotations. This is achieved by exploiting neighborhood regularization to constrain that the learnt latent vectors of genes with similar GO and/or PPI topological properties should be similar in the latent space. We have conducted extensive experiments to evaluate the performance of SL$^2$MF. The experimental results demonstrate the effectiveness of the proposed method and show that using the biological knowledge derived from PPI network and GO can help improve the prediction accuracy. In addition, the empirical comparison between SL$^2$MF and DAISY also indicates that SL$^2$MF is a useful complement to existing knowledge-based SL prediction methods.

## 2 Synthetical Lethality Prediction Model

This section first introduces the notations and problem formulation, and then describes details of the proposed SL$^2$MF model.

### 2.1 Preliminary

In this paper, we denote the set of genes by $\mathcal{U} = \{u_i\}_{i=1}^m$, where $m$ is the number of genes. Moreover, we use a binary matrix $\boldsymbol{Y} \in \mathbb{R}^{m \times m}$, where each element is denoted by $y_{ij} \in \{0, 1\}$, to describe the SL interaction data. If there exists an observed SL interaction between two genes $u_i$ and $u_j$, we set $y_{ij}$ to 1; otherwise, we set $y_{ij}$ to 0. Note that $\boldsymbol{Y}$ is symmetric, *i.e.*, $y_{ij} = y_{ji}$. Thus, we treat the gene pairs $(u_i, u_j)$ and $(u_j, u_i)$ as the same pair. Then, we denote the set of all gene pairs by $\mathcal{O} = \{(u_i, u_j) | 1 \leq i < m, i+1 \leq j \leq m\}$. In other words, we define $\mathcal{O}$ only considering the upper half of $\boldsymbol{Y}$. In addition, we denote the set of observed SL pairs by $\mathcal{O}^+ = \{(u_i, u_j) | y_{ij} = 1, 1 \leq i < m, i+1 \leq j \leq m\}$. The remaining gene pairs in the upper half of $\boldsymbol{Y}$ are denoted by $\mathcal{O}^- = \mathcal{O} \setminus \mathcal{O}^+$. We call the pairs in $\mathcal{O}^-$ as "unknown pairs", because there does not exist evidence demonstrating whether these gene pairs are SLs or not.

Moreover, we denote the GO semantic similarities between genes by $\boldsymbol{S}^G \in \mathbb{R}^{m \times m}$, where the $(i, j)$ element $s_{ij}^G$ denotes the GO semantic similarity between $u_i$ and $u_j$. The PPI topological similarities between genes [1] are denoted by $\boldsymbol{S}^P \in \mathbb{R}^{m \times m}$, where the $(i, j)$ element $s_{ij}^P$ is the PPI topological similarity between $u_i$ and $u_j$. The computation of GO semantic similarities and PPI topological similarities are introduced in Section 3.1. For each gene $u_i$, we use $N^G(u_i)$ to denote the set of $k_1$ genes that are most similar with $u_i$, measured by the GO semantic similarities $\boldsymbol{S}^G$, where $N^G(u_i) \subset \mathcal{U} \setminus u_i$. Similarly, we use $N^P(u_i)$ to denote the set of $k_2$ genes that are most similar with $u_i$, measured by the PPI topological similarities $\boldsymbol{S}^P$, where $N^P(u_i) \subset \mathcal{U} \setminus u_i$. In this work, $N^G(u_i)$ and $N^P(u_i)$ are called the GO nearest neighbors and PPI nearest neighbors of $u_i$, respectively.

The problem studied in this work can be defined as follows: *given a set of observed gene pairs $\mathcal{O}^+$ that are known to be SL pairs, how to predict a set of gene pairs that are most likely to form SL interactions from the "unknown pairs" $\mathcal{O}^-$.* We tackle this problem by first predicting the probability that two genes form SL relationship, and then ranking the candidate gene pairs based on the predicted probabilities in descending orders, such that the top-ranked gene pairs are most likely to be SL pairs. Specifically, the method of logistic matrix factorization [31], [33] is utilized to model the SL interaction probabilities between genes. Moreover, neighborhood regularization [31] is used to incorporate both

1. We use genes instead of gene products (*i.e.*, proteins) when we mention their PPI topological similarity.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2909908, IEEE/ACM Transactions on Computational Biology and Bioinformatics



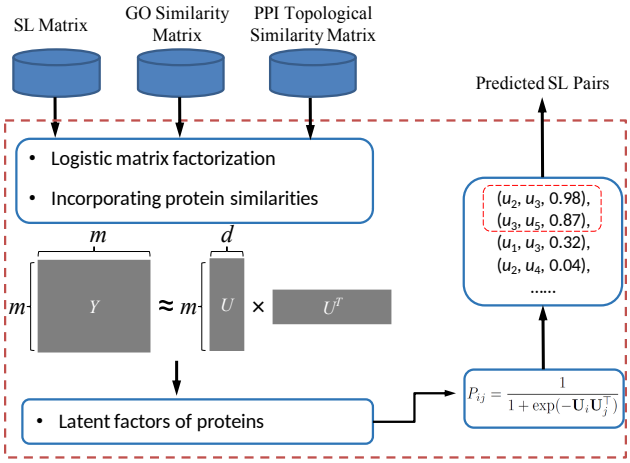



Fig. 1: The overall framework of SL$^2$MF.

GO semantic similarities and PPI topological similarities between genes to enhance the prediction accuracy. Figure 1 shows the overall framework of the proposed SL$^2$MF model.

### 2.2 Logistic Matrix Factorization

The proposed method SL$^2$MF is developed based on logistic matrix factorization, which has been successfully applied in personalized music recommendation [33] and drug-target interaction prediction [31]. The objective of logistic matrix factorization is to map genes into a shared low dimensional latent space and model the probabilities of SL interactions between genes using their latent representations. For each gene $u_i$, a latent vector $\boldsymbol{U}_i \in \mathbb{R}^{1\times d}$ is used to describe its properties, where $d$ denotes the dimensionality of the latent space, and $d \ll m$. We combine the latent vectors of all genes into a matrix $\boldsymbol{U} \in \mathbb{R}^{m\times d}$, where $\boldsymbol{U}_i$ is the $i^{th}$ row in $\boldsymbol{U}$. Then, the SL interaction probability between two genes $u_i$ and $u_j$ is defined by the following logistic function [33]:

$$p_{ij} = \frac{1}{1+\exp(-\boldsymbol{U}_i\boldsymbol{U}_j^\top)}. \tag{1}$$

To model the SL data, we use the gene pairs in $O^+$ that have observed SL interactions as positive training examples, and use the unknown pairs in $\mathcal{O}^-$ as negative training examples. Following previous study [31], we assign higher importance weights to gene pairs that are observed SL interactions. In other words, a known SL pair $(u_i, u_j) \in \mathcal{O}^+$ is treated as $c_{ij}$ $(c_{ij} > 1)$ positive training samples, an unknown gene pair is used as only *one* negative training sample.[2] By assuming all the training samples are independent with each other, we can define the likelihood of the observed SL data as follows:

$$p(\mathcal{O}|\boldsymbol{U}) = \prod_{(u_i,u_j)\in\mathcal{O}^+} p_{ij}^{c_{ij}y_{ij}}(1-p_{ij})^{c_{ij}(1-y_{ij})} \prod_{(u_i,u_j)\in\mathcal{O}^-} p_{ij}^{y_{ij}}(1-p_{ij})^{(1-y_{ij})}. \tag{2}$$

2. This work is not focusing on developing strategies to weight gene pairs. Thus, we empirically set the weights based on the confidence scores assigned to SLs in SynLethDB (see Section 3.3 for details). Moreover, the proposed SL$^2$MF method is a general framework which can also integrate more sophisticated weighting methods, for example the re-weighted probabilistic models (RPM) proposed in [34].

Note that $c_{ij}(1-y_{ij}) = 1-y_{ij}$ when $y_{ij} = 1$, and $y_{ij} = c_{ij}y_{ij}$ when $y_{ij} = 0$. Thus, we can rewrite the likelihood in Eq.(2) as follows:

$$p(\mathcal{O}|\boldsymbol{U}) = \prod_{i=1}^{m}\prod_{j=i+1}^{m} p_{ij}^{c_{ij}y_{ij}}(1-p_{ij})^{1-y_{ij}}. \tag{3}$$

Zero-mean spherical Gaussian priors are placed on the gene latent vectors as follows [33]:

$$p(\boldsymbol{U}|\sigma^2) = \prod_{i=1}^{m}\mathcal{N}(\boldsymbol{U}_i|0,\sigma^2\boldsymbol{I}), \tag{4}$$

where $\boldsymbol{I}$ is the identity matrix, and $\sigma^2$ is the parameter used to control the variances of Gaussian distributions. Through Bayesian inference, we have

$$p(\boldsymbol{U}|\mathcal{O},\sigma^2) \propto p(\mathcal{O}|\boldsymbol{U})p(\boldsymbol{U}|\sigma^2). \tag{5}$$

The logarithm of the posterior distribution is as follows:

$$\log p(\boldsymbol{U}|\mathcal{O},\sigma^2) \propto \sum_{i=1}^{m}\sum_{j=i+1}^{m}\Big[-(1+c_{ij}y_{ij}-y_{ij})\log\left(1+\exp(\boldsymbol{U}_i\boldsymbol{U}_j^\top)\right) + c_{ij}y_{ij}\boldsymbol{U}_i\boldsymbol{U}_j^\top\Big] - \frac{\lambda}{2}\|\boldsymbol{U}\|_F^2, \tag{6}$$

where $\lambda = \frac{1}{\sigma^2}$, and $\|\cdot\|_F$ is the Frobenius norm of a matrix. Then, the gene latent vectors can be learned by maximizing the posterior distribution. It is equivalent with minimizing the following loss function:

$$L_{\mathcal{O}} = \sum_{i=1}^{m}\sum_{j=i+1}^{m}\left[(1+c_{ij}y_{ij}-y_{ij})\log\left(1+\exp(\boldsymbol{U}_i\boldsymbol{U}_j^\top)\right) - c_{ij}y_{ij}\boldsymbol{U}_i\boldsymbol{U}_j^\top\right] + \frac{\lambda}{2}\|\boldsymbol{U}\|_F^2. \tag{7}$$

### 2.3 Incorporating Gene Similarities

As mentioned earlier, two types of gene similarities, *i.e.*, GO semantic similarities and PPI topological similarities, are considered to improve the prediction accuracy. Specifically, we assume that genes with similar functional and/or network properties should have similar representations in the latent space. Then, we propose to minimize the following loss function to exploit the GO semantic similarities between genes for SL prediction:

$$L_{\mathcal{G}} = \frac{1}{2}\sum_{i=1}^{m}\sum_{u_j\in N^G(u_i)} s_{ij}^G\|\boldsymbol{U}_i-\boldsymbol{U}_j\|_2^2. \tag{8}$$

Similarly, the PPI similarities between genes can also be incorporated by minimizing the following loss function:

$$L_{\mathcal{P}} = \frac{1}{2}\sum_{i=1}^{m}\sum_{u_j\in N^P(u_i)} s_{ij}^P\|\boldsymbol{U}_i-\boldsymbol{U}_j\|_2^2. \tag{9}$$

Note that the proposed method only considers $k_1$ GO nearest neighbors and $k_2$ PPI nearest neighbors of each gene $u_i$, instead of all its neighbors (i.e., $\mathcal{U}\setminus u_i$), to improve the prediction accuracy. Because using all the neighbors may potentially introduce noisy information and thus reduce the model accuracy. In addition, the experimental results in Section 3.4 also demonstrate that better performance can be achieved by considering only the nearest neighbors of a gene than considering all of its neighbors.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2909908, IEEE/ACM Transactions on Computational Biology and Bioinformatics



**Algorithm 1:** The **SL$^2$MF Algorithm**

**Input :** $\boldsymbol{Y}$, $\boldsymbol{W}$, $\boldsymbol{S}^G$, $\boldsymbol{S}^P$, $d$, $\lambda$, $\alpha$, $\beta$, $\gamma$, $k_1$, $k_2$
**Output:** $\boldsymbol{U}$

Initialize gene latent vectors $\boldsymbol{U}$ using the Gaussian distribution $\mathcal{N}(0, 1/\sqrt{d})$;
Compute the adjacency matrix $\boldsymbol{A}$ and the Laplacian matrix $\boldsymbol{L}^G$ according to Eq. (11);
Compute the adjacency matrix $\boldsymbol{B}$ and the Laplacian matrix $\boldsymbol{L}^P$ according to Eq. (13);
Set $\varphi_{ik} = 0, \forall 1 \leq i \leq m$ and $1 \leq k \leq d$;
**for** $iter = 1, 2, \cdots, max\_iter$ **do**
  Compute the interaction probability matrix $\boldsymbol{P}$ according to Eq. (1);
  `// compute the gradient w.r.t. U`
  $\boldsymbol{Z} \leftarrow \big[\boldsymbol{W} \odot (\boldsymbol{P} - \boldsymbol{Y}) + \lambda \boldsymbol{I} + \alpha \boldsymbol{L}^G + \beta \boldsymbol{L}^P\big]\boldsymbol{U}$;
  **for** $i = 1, \ldots, m$ **do**
    **for** $k = 1, \ldots, d$ **do**
      `//` $Z_{ik}$ `is the` $(i,k)$ `element in` $\boldsymbol{Z}$
      $\varphi_{ik} \leftarrow \varphi_{ik} + Z_{ik} \cdot Z_{ik}$;
      `//` $U_{ik}$ `is the` $(i,k)$ `element in` $\boldsymbol{U}$
      $U_{ik} \leftarrow U_{ik} - \gamma \frac{Z_{ik}}{\sqrt{\varphi_{ik}}}$;

### 2.4 The Unified SL$^2$MF Model

The final prediction model can be formulated by considering the SL interaction data as well as both GO semantic similarities and PPI topological similarities between genes. By substituting Eq. (8) and Eq. (9) into Eq. (7), the unified SL$^2$MF model is obtained as follows:

$$\min_{\boldsymbol{U}} L_{\mathcal{O}} + \alpha L_{\mathcal{G}} + \beta L_{\mathcal{P}}, \tag{10}$$

where $\alpha$ and $\beta$ are regularization parameters controlling the influences from GO nearest neighbors and PPI nearest neighbors, respectively.

For simplicity, we define two adjacency matrices $\boldsymbol{A} \in \mathbb{R}^{m \times m}$ and $\boldsymbol{B} \in \mathbb{R}^{m \times m}$ to describe the nearest neighbors of genes. The $(i,j)$ element of $\boldsymbol{A}$ is defined as follows:

$$a_{ij} = \begin{cases} s_{ij}^G & \text{if } u_j \in N^G(u_i), \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Then, the loss function $L_{\mathcal{G}}$ can be rewritten as follows:

$$L_{\mathcal{G}} = \frac{1}{2}\text{tr}(\boldsymbol{U}^\top \boldsymbol{L}^G \boldsymbol{U}), \tag{12}$$

where $\text{tr}(\cdot)$ denotes the trace of a matrix, $\boldsymbol{L}^G = \boldsymbol{D}^G - (\boldsymbol{A} + \boldsymbol{A}^\top)$, $\boldsymbol{D}_G \in \mathbb{R}^{m \times m}$ is a diagonal matrix, in which the diagonal elements are defined as $d_{ii}^G = \sum_{j=1}^m (a_{ij} + a_{ji})$. Similarly, the $(i,j)$ element of $\boldsymbol{B}$ is defined as follows:

$$b_{ij} = \begin{cases} s_{ij}^P & \text{if } u_j \in N^P(u_i), \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

We can rewrite the loss function $L_{\mathcal{P}}$ as follows:

$$L_{\mathcal{P}} = \frac{1}{2}\text{tr}(\boldsymbol{U}^\top \boldsymbol{L}^P \boldsymbol{U}), \tag{14}$$

where $\boldsymbol{L}^P = \boldsymbol{D}^P - (\boldsymbol{B} + \boldsymbol{B}^\top)$. $\boldsymbol{D}^P \in \mathbb{R}^{m \times m}$ is a diagonal matrix, in which the diagonal elements are defined as $d_{ii}^P = \sum_{j=1}^m (b_{ij} + b_{ji})$. Then, the unified model in Eq. (10) can be rewritten as follows:

$$\min_{\boldsymbol{U}} \frac{1}{2}\sum_{i=1}^m \sum_{j=1}^m w_{ij}\big[\log\big(1 + \exp(\boldsymbol{U}_i \boldsymbol{U}_j^\top)\big) - y_{ij}\boldsymbol{U}_i \boldsymbol{U}_j^\top\big] + \frac{\lambda}{2}\|\boldsymbol{U}\|_F^2 + \frac{\alpha}{2}\text{tr}(\boldsymbol{U}^\top \boldsymbol{L}^G \boldsymbol{U}) + \frac{\beta}{2}\text{tr}(\boldsymbol{U}^\top \boldsymbol{L}^P \boldsymbol{U}), \tag{15}$$

where $w_{ij}$ is defined as follows:

$$w_{ij} = \begin{cases} c_{ij} & \text{if } y_{ij} = 1 \text{ and } i \neq j, \\ 1 & \text{if } y_{ij} = 0 \text{ and } i \neq j, \\ 0 & \text{if } i = j. \end{cases} \tag{16}$$

Note that $y_{ij} = y_{ji}$ and $c_{ij} = c_{ji}$, thus $w_{ij} = w_{ji}$.

The problem in Eq. (15), denoted by $L$, can be solved by the gradient descent optimization procedure. The gradient of Eq. (15) with respect to $\boldsymbol{U}$ is as follows:

$$\frac{\partial L}{\partial \boldsymbol{U}} = \big[\boldsymbol{W} \odot (\boldsymbol{P} - \boldsymbol{Y}) + \lambda \boldsymbol{I} + \alpha \boldsymbol{L}^G + \beta \boldsymbol{L}^P\big]\boldsymbol{U}, \tag{17}$$

where $\odot$ is the Hadamard product of two matrices. $\boldsymbol{W} \in \mathbb{R}^{m \times m}$ is the weighting matrix, where the $(i,j)$ element is $w_{ij}$ (refer to Eq. (16)). $\boldsymbol{P} \in \mathbb{R}^{m \times m}$ is the interaction probability matrix, in which the $(i,j)$ element is $p_{ij}$ (refer to Eq. (1)). Following [33], we adopt the AdaGrad algorithm [35] to accelerate the convergence of the gradient descent procedure. The details of the optimization algorithm developed for SL$^2$MF are summarized in Algorithm 1.

Once having learned the model parameters $\boldsymbol{U}$, Eq (1) is used to compute the probability that a candidate pair of genes are SL. Then, all candidate gene pairs $\mathcal{O}^-$ are ranked based on the predicted probabilities in descending orders, and the top-ranked gene pairs will be chosen as the predictions.

### 2.5 Discussions

As shown in Eq. (10) and Eq. (15), the unified model is a linear combination of the optimization problem in Eq. (7) and the neighborhood regularization constraints Eq. (8) and Eq. (9). Therefore, the construction of the unified model does not fully follow the spirit of probabilistic models. However, inspired by previous research work about social recommendation [36], we can note that the usage of the regularization terms $\frac{\lambda}{2}\|\boldsymbol{U}\|_F^2 + \frac{\alpha}{2}\text{tr}(\boldsymbol{U}^\top \boldsymbol{L}^G \boldsymbol{U}) + \frac{\beta}{2}\text{tr}(\boldsymbol{U}^\top \boldsymbol{L}^P \boldsymbol{U})$ in Eq. (15) is equivalent to assigning the following priors on the gene latent vectors $\boldsymbol{U}$:

$$p(\boldsymbol{U}|\boldsymbol{L}^G, \boldsymbol{L}^P, \lambda, \alpha, \beta) \propto \mathcal{MN}_{m \times d}\Big(\boldsymbol{0}, (\lambda \boldsymbol{I} + \alpha \boldsymbol{L}^G + \beta \boldsymbol{L}^P)^{-1}, \boldsymbol{I}\Big), \tag{18}$$

where $\mathcal{MN}_{a \times b}(\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Theta})$ is a matrix variate normal (MVN) distribution[3] [37] with the mean $\boldsymbol{M} \in \mathbb{R}^{a \times b}$, row covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{a \times a}$, and column covariance $\boldsymbol{\Theta} \in \mathbb{R}^{b \times b}$. In Eq. (18), the relations between different rows of $\boldsymbol{U}$ are described by the row precision matrix: $\boldsymbol{\Omega} = \lambda \boldsymbol{I} + \alpha \boldsymbol{L}^G + \beta \boldsymbol{L}^P$. Note that each

3. The density function for a random matrix $\boldsymbol{X}$ following the MVN distribution $\mathcal{MN}_{a \times b}(\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Theta})$ is as follows:

$$p(\boldsymbol{X}) = \frac{\exp(-\frac{1}{2}\text{tr}[\boldsymbol{\Theta}^{-1}(\boldsymbol{X} - \boldsymbol{M})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{M})])}{(2\pi)^{ab/2}|\boldsymbol{\Theta}|^{a/2}|\boldsymbol{\Sigma}|^{b/2}}.$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2909908, IEEE/ACM Transactions on Computational Biology and Bioinformatics



row of $\boldsymbol{U}$ denotes the latent vector of an individual gene. Therefore, in other words, the relations between different genes are modeled by the row precision matrix $\boldsymbol{\Omega}$. Through Bayesian inference, we have

$$p(\boldsymbol{U}|\mathcal{O}, \boldsymbol{L}^G, \boldsymbol{L}^P, \lambda, \alpha, \beta) \propto p(\mathcal{O}|\boldsymbol{U})p(\boldsymbol{U}|\boldsymbol{L}^G, \boldsymbol{L}^P, \lambda, \alpha, \beta). \tag{19}$$

The gene latent vectors $\boldsymbol{U}$ can be learned by maximizing the above posterior distribution, which is equivalent to solving the optimization problem in Eq. (15).

## 3 Results

In this section, we first introduce the data used in the experiments and the experimental settings. Then, we show the performances of SL$^2$MF under different settings.

### 3.1 Data and Experimental Settings

The proposed SL$^2$MF method works on the SL interaction matrix and various gene similarity matrices, including PPI topological similarity matrix and GO semantic similarity matrix. The SL dataset was downloaded from the database of SynLethDB [27]. SynLethDB integrates SL interaction data from four different sources: (1) SL pairs manually curated [38], (2) SL pairs extracted by text mining [27], (3) genetic interactions detected by GenomeRNAi [39] and shRNA (i.e., the DECIPHER project[4]), and (4) SL pairs computationally predicted by DAISY [16]. Overall, there are 19,667 SL pairs involving 6,375 genes in the SL matrix. The sparsity of the SL interaction matrix is $99.90\%$, thus the SL prediction is a very challenging prediction task. In SynLethDB, a confidence score is assigned to each SL pair, considering the accuracies of different identification methods. These confidence scores have been exploited to define the weights assigned to SL pairs in SL$^2$MF (refer to Table 2 for details). Moreover, we downloaded the HPRD database [40] to construct the PPI topological similarity matrix. FSWeight [41], which is a topological similarity between genes based on the number of common neighbors in the PPI data, was then calculated. Among various methods for computing the GO term similarity, we used the method proposed in [42] and then calculated the GO semantic similarity between genes. It is known that GO has three sub-ontologies, namely biological process (BP), molecular function (MF) and cellular component (CC). In GO (version: June 2017), there are 29,660 terms in BP, 11,120 terms in MF, and 4,115 terms in CC, respectively. BP has many more terms and are more enriched than MF and CC. Therefore, we computed the semantic similarity between genes only using their BP terms.

In our experiments, we adopted 5-fold cross-validation to evaluate the proposed method. The known SL pairs $\mathcal{P}$ in SynLethDB were equally divided into 5 non-overlapping subsets (*i.e.*, $\mathcal{P}_1$, $\mathcal{P}_2$, $\mathcal{P}_3$, $\mathcal{P}_4$, $\mathcal{P}_5$). In each round, a subset $\mathcal{P}_i(1 \leq i \leq 5)$ of SL interactions were chosen for model testing, and the other four subsets were used as positive examples for model training (i.e., $\mathcal{O}^+$). The objective of SL$^2$MF is to predict the potential SL pairs in $\mathcal{O}^-$ (*i.e.*, $\mathcal{P}_i$) via ranking the gene pairs in $\mathcal{P}_i$ before other gene pairs in $\mathcal{O}^-$. Therefore, we use AUC (*i.e.*, area under the ROC curve) and AUPR (*i.e.* area under the precision-recall curve) as evaluation metrics. In particular, the AUC and AUPR scores were calculated based on the predictions of all candidate gene-gene pairs.

4. http://www.decipherproject.net/shRNA-libraries/bi-specific/

TABLE 1: The 5-fold cross-validation AUC and AUPR scores of BLM-NII and SL$^2$MF with respect to different settings.

| Method | AUC | AUPR |
|---|---|---|
| BLM-NII | 0.7228±0.0293 | 0.0011±0.0010 |
| SL Matrix | 0.8051±0.0060 | 0.2081±0.0072 |
| SL+PPI | 0.8330±0.0029 | 0.1326±0.0046 |
| SL+GO | 0.8437±0.0049 | 0.2414±0.0062 |
| SL+PPI+GO | 0.8480±0.0048 | 0.2388±0.0057 |

The proposed SL prediction model is built based on the known SL interaction data in human, the GO semantic similarities and PPI topological similarities between genes. Therefore, existing supervised learning methods [22], [23], [24], [25] for yeast SL prediction and other knowledge-based methods [16], [18], [19], [21] for human SL prediction cannot be directly used as baselines in this study. As SL prediction shares similar spirits with drug-target interaction prediction [31], we adopt the similarity-based drug-target interaction prediction method BLM-NII [43] as the baseline method. In BLM-NII, the linear combination weight $\alpha$ was chosen from $\{0, 0.1, 0.2, \cdots, 1.0\}$, and the *max* function was used to integrate the interaction scores predicted independently by using the GO semantic similarities and the PPI topological similarities. For SL$^2$MF, we empirically set the dimensionality of latent space $d$ to 50. We assigned uniform weights to SL pairs as shown in Table 2, and the parameter $c$ was set to 50. The regularization parameters $\lambda$, $\alpha$, and $\beta$ were set to $0.01$, $1.0$, and $10$, respectively. The learning rate $\gamma$ of the optimization algorithm was set to $2^{-5}$. The number of GO semantic nearest neighbors and PPI topological nearest neighbors $k_1$ and $k_2$ were set to $100$. In Section 3.4, we will discuss the impacts of these parameters.

### 3.2 5-fold Cross Validation Results

Table 1 shows the AUC and AUPR scores of BLM-NII and SL$^2$MF with respect to different inputs. In Table 1, "SL Matrix" refers to SL$^2$MF without using any similarity matrices, "SL+PPI" refers to SL$^2$MF incorporating only the PPI topological similarity matrix via setting the parameter $\alpha$ to 0 in Eq. (15), and "SL+GO" refers to SL$^2$MF incorporating only the GO semantic similarity matrix by setting $\beta$ to 0. Similarly, "SL+PPI+GO" indicates that SL$^2$MF works on SL matrix together with both the PPI topological similarity matrix and the GO semantic similarity matrix. Based on the results in Table 1, we can draw the following two conclusions.

First, both PPI topological similarities and GO similarities can help improve the performance for predicting SL interactions, in terms of AUC. For example, the AUC scores are improved by 3.47% and 4.79% after incorporating the PPI topological similarities and GO similarities, respectively. Moreover, the GO similarities can also help improve AUPR scores. However, by incorporating the PPI similarities into the model, AUPR scores are dropped off. This is to be expected, since the strategy

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2909908, IEEE/ACM Transactions on Computational Biology and Bioinformatics



TABLE 2: Different definitions of the importance weights assigned to SL pairs. Here $\varepsilon_{ij}$ is the confidence score of each SL pair in SynLethDB [27].

| | Definitions |
|---|---|
| Uniform Weights | $c_{ij} = c$ |
| Linear Weights | $c_{ij} = 1 + c\varepsilon_{ij}$ |
| Loglinear Weights | $c_{ij} = 1 + c\log(1 + \varepsilon_{ij})$ |

used to improve AUC scores is not guaranteed to improve AUPR scores, and also because AUPR punishes highly ranked false positives much more than AUC [44]. One potential explanation for this observation is as follows. When incorporating PPIs, less non-SL gene pairs are ranked before SL gene pairs, thus increasing AUC. However, these falsely ranked non-SL gene pairs may be ranked at higher positions, thus decreasing AUPR.

Secondly, the GO semantic similarity matrix performs better than the PPI topological similarity matrix in improving the prediction accuracy. Compared with the PPI topological similarity matrix, the GO semantic similarity matrix is able to further improve the AUC and AUPR by 1.28% and 82.05%, respectively. One reason could be that the PPI topological similarity matrix is much sparser than the GO semantic similarity matrix. In particular, the sparsity of the PPI topological similarity matrix $\boldsymbol{S}^P$ and GO semantic similarity matrix $\boldsymbol{S}^G$ are 98.65% and 17.04%, respectively. The sparsity is defined as the percentage of zero elements in the matrix.

Thirdly, as shown in Table 1, we can notice that the proposed SL$^2$MF method significantly outperforms BLM-NII, in terms of AUC and AUPR, and the AUPR score achieved by BLM-NII is around 0.001. The potential reason is that both similarity matrices $\boldsymbol{S}^G$ and $\boldsymbol{S}^P$ are sparse. Thus, BLM-NII cannot exploit enough gene similarities for accurate SL prediction. In addition, this result also demonstrates the effectiveness of SL$^2$MF in handling sparse information for SL prediction.

### 3.3 Benefits of Importance Weights

The SL pairs collected in SynLethDB are usually supported by different types of evidences. They are more trustworthy and important than the unknown pairs. Hence, we design a parameter $c_{ij}$ in Eq. (3) to control the importance levels for known SL pairs. In particular, each SL pair is treated as $c_{ij}$ positive training instances while each unknown pair is treated as ***a single*** negative instance. In this paper, we have studied three different definitions of the importance weights $c_{ij}$ assigned to SL pairs based on the confidence scores defined in SynLethDB [27], as shown in Table 2.

Figure 2 shows the performances of the proposed SL$^2$MF method with respect to different settings of $c_{ij}$. As shown in Figure 2, better performances can usually be achieved by assigning uniform weights to SL pairs. This observation is to be expected. Because the calculation of AUC and AUPR scores does not consider the quality of the SL pairs. Uniform weights may achieve better performance than linear weights and log-linear weights, in terms of AUC and AUPR. Moreover, the confidence scores provided in SynLethDB are empirically set by the authors of SynLethDB. These scores may not be optimal for SL prediction tasks. For

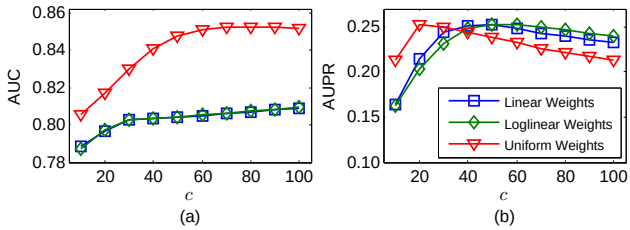



Fig. 2: The AUC and AUPR scores of SL$^2$MF with different definitions of the weight $c_{ij}$ for "SL+GO+PPI".

uniform weights, when $c$ is set to 1 (*i.e.*, known SL pairs are equally important as unknown pairs), the performance of SL$^2$MF is poor, *i.e.*, the AUC is only 0.5119. When we increase the value of $c$, the performance of SL$^2$MF is significantly improved. However, when $c$ is large enough (*e.g.*, $c > 50$), the performance of SL$^2$MF becomes stable. As such, we recommend that 50 is a proper setting for $c$ when predicting SL interactions by incorporating both similarities.

### 3.4 Parameter Sensitivity Analysis

This sections focuses on sensitivity analysis for parameters $d$, $\alpha$, $\beta$, $\lambda$, $k_1$, and $k_2$. Note that we use SL matrix and both similarity matrices (*i.e.*, "SL+GO+PPI") to show the effects of parameters $d$, $\alpha$, and $\beta$. For parameters $k_1$ and $k_2$, we show their settings for "SL+PPI" and "SL+GO", respectively.

Parameter $d$ is the dimensionality of the learned latent vectors of genes. As shown in Figure 3 (a), AUC becomes stable when $d \geq 30$. Moreover, AUPR generally improves with the increase of $d$. However, larger $d$ causes more computation time used to learn gene latent vectors. Considering both the efficiency and accuracy, we empirically set $d$ to 50 in this study. Parameters $\alpha$ and $\beta$ are coefficients controlling contributions of GO semantic similarity matrix and PPI topological similarity matrix, respectively. Figures 3 (c) and 3 (d) show AUC and AUPR scores of proposed method with respect to different settings of $\alpha$, by fixing $\beta$ to 10. We can observe that the optimal value for $\alpha$ is 1. In Figures 3 (e) and (f), we fix $\alpha$ to 1 and then vary the values of $\beta$. We can observe that both AUC and AUPR are consistently high when $\beta \in [0.0001, 10]$, but they are decreased when $\beta$ is further increased to 100 and 1000. Gradually decreasing $\beta$ from 10 to 0.0001 will not affect the performance, indicating that the PPI topological similarity matrix has less impact on the performance than the GO semantic similarity matrix. This result is also consistent with Table 1, where "SL+GO" performs better than "SL+PPI". Moreover, Figures (g) and (f) summarize the performances with respect to different settings of $\lambda$. We can notice that both AUC and AUPR are very stable when $\lambda \in [0.0001, 10]$. However, when $\lambda$ is further increased, AUC is dramatically decreased and AUPR is slightly increased.

Parameters $k_1$ and $k_2$ are the numbers of nearest neighbors used in regularization constraints Eq. (8) and Eq. (9). For "SL+GO", it is clear that the optimal value of $k_1$ is 150 as shown in Figures 4 (a) and (b). For "SL+PPI", the number of nearest neighbors utilized by regularization is supposed to be small, due to the high sparsity of the PPI topological similarity matrix. Figure 4 (c), where the optimal value of $k_2$ is 10, also confirms this supposition. Overall, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2909908, IEEE/ACM Transactions on Computational Biology and Bioinformatics



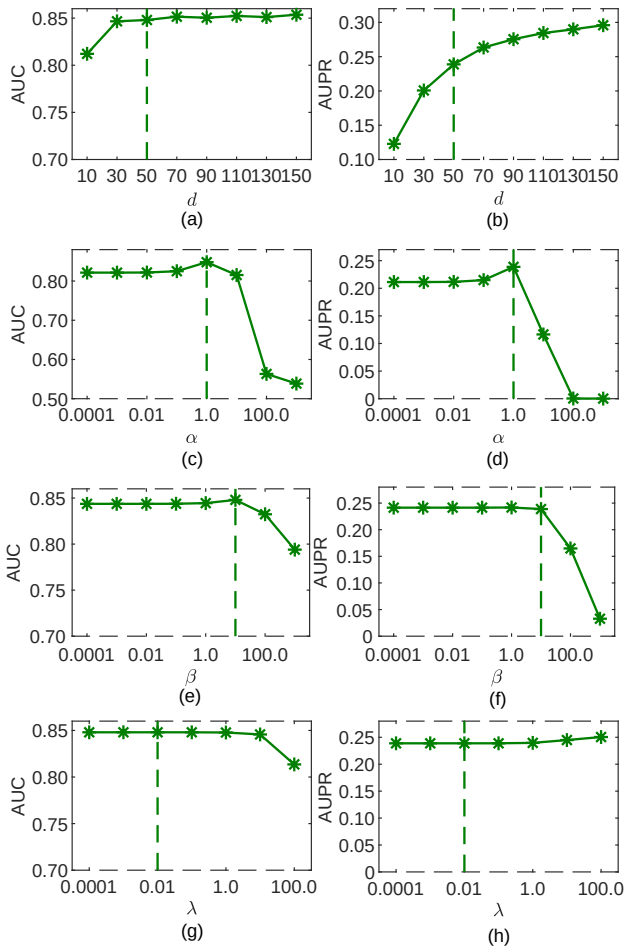



Fig. 3: Impacts of $d$, $\alpha$, $\beta$, and $\lambda$ for "SL+GO+PPI".

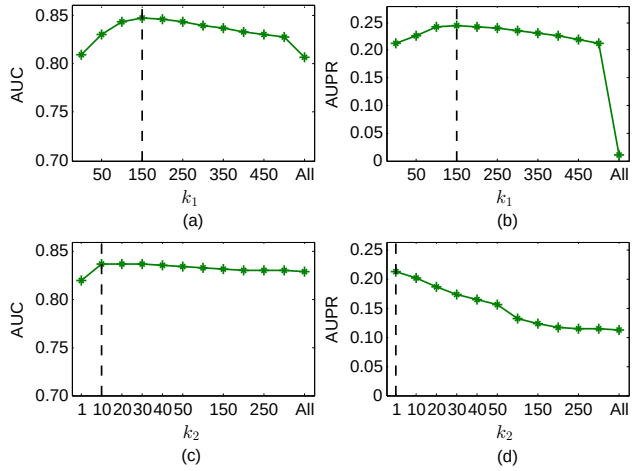



Fig. 4: Impacts of $k_1$ for "SL+GO" and $k_2$ for "SL+PPI".

recommended settings of parameters $k_1$ and $k_2$ are 150 and 10 for "SL+GO" and "SL+PPI", respectively. Moreover, the rightmost points in Figure 4 refer to AUC and AUPR scores using "*All*" the neighbors (*i.e.*, 6,374 neighbors). Therefore, better performances can be achieved by exploiting only a few nearest neighbors instead of all the neighbors.

### 3.5 Comparison with DAISY

As mentioned earlier, SL$^2$MF is a supervised matrix factorization model, whereas DAISY [16] is a well-known knowledge-based method for SL prediction which is based on three hypotheses about SL gene pairs. More importantly, we have conducted experiments and obtained all the results above using the whole SynLethDB dataset comprising 19,667 SL pairs, including 5,740 SL pairs predicted by DAISY. To fairly compare the predictions of SL$^2$MF with those of DAISY, we modify the data setting for SL$^2$MF as follows.

We first remove the SL pairs predicted by DAISY from SynLethDB. Then, we also reserve a validation set consisting of SL pairs with high reliability (e.g., high confidence scores in SynLethDB). After excluding these two parts, the remaining SL pairs in SynLethDB will be used as positive training samples to train SL$^2$MF. In particular, we have worked on two scenarios for comparing SL$^2$MF and DAISY as shown in Figure 5. In scenario 1, the manually curated SL pairs (*i.e.*, the SynLethality database [38] which is also the part of SynLethDB with the highest confidence scores) are used as validation set. In scenario 2, the validation set is expanded by further including those SL pairs extracted by text mining [27].

Once completing the model training of SL$^2$MF, the candidate SLs are ranked based on the predicted probabilities. The validation AUC scores of SL$^2$MF are 0.7330 in scenario 1 and 0.6701 in scenario 2. The AUC score in scenario 2 is relatively lower, probably because less SL pairs are used to train SL$^2$MF in the second scenario, which might affect the performance of the trained SL$^2$MF model. However, in both scenarios, the validation AUPR scores of SL$^2$MF are less than 0.005. One potential reason is that AUPR punishes highly ranked false positives much more than AUC [44], and the validation SLs do not rank high based on the interaction probabilities predicted by SL$^2$MF.

Moreover, we have also studied the top 5,740 SL pairs (i.e., the same size of DAISY) predicted by SL$^2$MF. In the first scenario, 14 out of 5,740 pairs predicted by SL$^2$MF are also in the SynLethality database, while DAISY has no overlap with SynLethality database. In the second scenario, 27 out of 5,740 pairs predicted SL$^2$MF are verified by the validation set (*i.e.*, SynLethality + Text Mining), while only 3 SL pairs in DAISY are verified by text mining. In addition, we also observed that SL$^2$MF and DAISY have limited overlap in their predictions. In scenario 1, they have only 5 predicted SL pairs in common: (POLR2A, WRAP53), (PARP1, PRKDC), (EGFR, IGFBP3), (POLD1, POLR2A), (CHEK1, MRE11); in scenario 2, only 3 SL pairs are commonly predicted by both methods: (POLR2A, WRAP53), (EGFR, IGFBP3), (POLD1, POLR2A). Such a small overlap actually makes sense as SL$^2$MF and DAISY have different prediction mechanisms and use different data sources for prediction. For example, in scenario 1, the 5,740 SL pairs predicted by SL$^2$MF correspond to 3,555 genes. Meanwhile, the 5,740 SL pairs predicted by DAISY involve 3,796 genes. However, there are only 998 genes covered by both DAISY and SL$^2$MF predictions. Given SL$^2$MF's good performance evaluated by cross validation, we believe that SL$^2$MF is a useful complement to the existing methods, such as DAISY.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2909908, IEEE/ACM Transactions on Computational Biology and Bioinformatics



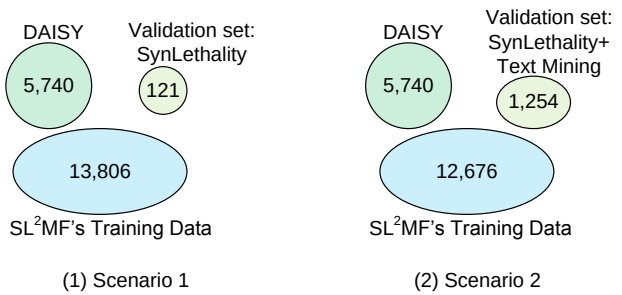



Fig. 5: Two different scenarios to run SL$^2$MF for comparison with DAISY. The set of SLs predicted by DAISY and the validation SL set have no overlap in scenario 1, and these two sets have 3 common SL pairs in scenario 2. Both scenarios have 19,667 SL pairs.

## 4 Conclusion and Future Work

In this paper, we have proposed a novel method for predicting the genetic interactions of synthetic lethality, named SL$^2$MF, which exploits logistic matrix factorization to model the probability that two genes are likely to form synthetic lethality. Because the observed SL pairs are supported by different types of evidence, we proposed to assign higher importance weights to known SL pairs, and assigned lower importance weights to other gene pairs. To further enhance the prediction accuracy of SL$^2$MF, we integrated both GO semantic similarities and PPI topological similarities between genes into learning gene latent vectors. Extensive experiments have been performed to evaluate the prediction accuracies of SL$^2$MF, under different settings. Experimental results have shown that SL$^2$MF is able to achieve promising performance for SL prediction. We believe that it can serve as a good baseline for future studies on this topic.

The future work will focus on the following directions. First, we plan to integrate more sophisticated weighting method, *e.g.*, the re-weighted probabilistic model (RPM) proposed in [34], to improve the prediction accuracy of SL$^2$MF. Moreover, the lack of a nice independent validation dataset, which covers all required information of various methods, to compare SL$^2$MF with existing SL prediction methods is a weakness of this work. To address this issue, it will be a highly desirable future work to expand the current version of SynLethDB database to incorporate SLs identified by siRNA/CRISPR based knockdown screens or other existing methods. Furthermore, we also would like to integrate more types of data (*e.g.*, protein domains, TCGA data) to further improve SL$^2$MF's performance. As more data are processed, we can also investigate feature-based classification models (*e.g.*, Gradient Boosting Machines [45]) for SL prediction. In addition, more effective strategies would also be studied to integrate different types of gene similarities to improve the prediction performances. Last but not least, verifying the effectiveness of SL$^2$MF through wet-lab experiments is also a potential research direction for future work.

## 5 Acknowledgement

This research is supported, in part, by the National Research Foundation, Prime Minister's Office, Singapore under its IDM Futures Funding Initiative, and the Alibaba-NTU Singapore Joint Research Institute. This research is also supported, in part, by the Start-up grant of ShanghaiTech University.

## References

[1] A. Ashworth, C. J. Lord, and J. S. Reis-Filho, "Genetic interactions in cancer progression and treatment," *Cell*, vol. 145, no. 1, pp. 30–38, 2011.

[2] L. H. Hartwell, P. Szankasi, C. J. Roberts, A. W. Murray, and S. H. Friend, "Integrating genetic approaches into the discovery of anticancer drugs," *Science*, vol. 278, no. 5340, pp. 1064–1068, 1997.

[3] D. P. McLornan, A. List, and G. J. Mufti, "Applying synthetic lethality for the selective targeting of cancer," *New England Journal of Medicine*, vol. 371, no. 18, pp. 1725–1735, 2014.

[4] A. Simons, N. Dafni, I. Dotan, Y. Oron, and D. Canaani, "Establishment of a chemical synthetic lethality screen in cultured human cells," *Genome research*, vol. 11, no. 2, pp. 266–273, 2001.

[5] N. C. Turner, C. J. Lord, E. Iorns, R. Brough, S. Swift, R. Elliott, S. Rayter, A. N. Tutt, and A. Ashworth, "A synthetic lethal sirna screen identifying genes mediating sensitivity to a parp inhibitor," *The EMBO journal*, vol. 27, no. 9, pp. 1368–1377, 2008.

[6] J. Luo, M. J. Emanuele, D. Li, C. J. Creighton, M. R. Schlabach, T. F. Westbrook, K.-K. Wong, and S. J. Elledge, "A genome-wide rnai screen identifies multiple synthetic lethal interactions with the ras oncogene," *Cell*, vol. 137, no. 5, pp. 835–848, 2009.

[7] M. M. Martins, A. Y. Zhou, A. Corella, D. Horiuchi, C. Yau, T. Rakshandehroo, J. D. Gordan, R. S. Levin, J. Johnson, J. Jascur *et al.*, "Linking tumor mutations to drug responses via a quantitative chemical–genetic interaction map," *Cancer discovery*, vol. 5, no. 2, pp. 154–167, 2015.

[8] D. Du, A. Roguev, D. E. Gordon, M. Chen, S.-H. Chen, M. Shales, J. P. Shen, T. Ideker, P. Mali, L. S. Qi *et al.*, "Genetic interaction mapping in mammalian cells using crispr interference," *Nature Methods*, 2017.

[9] K. Han, E. E. Jeng, G. T. Hess, D. W. Morgens, A. Li, and M. C. Bassik, "Synergistic drug combinations for cancer identified in a crispr screen for pairwise genetic interactions," *Nature Biotechnology*, 2017.

[10] B. Boucher and S. Jenna, "Genetic interaction networks: better understand to better predict," *Frontiers in genetics*, vol. 4, 2013.

[11] T. Zhan and M. Boutros, "Towards a compendium of essential genes–from model organisms to synthetic lethality in cancer cells," *Critical reviews in biochemistry and molecular biology*, vol. 51, no. 2, pp. 74–85, 2016.

[12] D. Deutscher, I. Meilijson, S. Schuster, and E. Ruppin, "Can single knockouts accurately single out gene functions?" *BMC Systems Biology*, vol. 2, no. 1, p. 50, 2008.

[13] P. F. Suthers, A. Zomorrodi, and C. D. Maranas, "Genome-scale gene/reaction essentiality and synthetic lethality analysis," *Molecular systems biology*, vol. 5, no. 1, p. 301, 2009.

[14] O. Folger, L. Jerby, C. Frezza, E. Gottlieb, E. Ruppin, and T. Shlomi, "Predicting selective drug targets in cancer through metabolic networks," *Molecular systems biology*, vol. 7, no. 1, p. 501, 2011.

[15] A. Pratapa, S. Balachandran, and K. Raman, "Fast-sl: an efficient algorithm to identify synthetic lethal sets in metabolic networks," *Bioinformatics*, p. btv352, 2015.

[16] L. Jerby-Arnon, N. Pfetzer, Y. Y. Waldman, L. McGarry, D. James, E. Shanks, B. Seashore-Ludlow, A. Weinstock, T. Geiger, P. A. Clemons *et al.*, "Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality," *Cell*, vol. 158, no. 5, pp. 1199–1209, 2014.

[17] S. Srihari, J. Singla, L. Wong, and M. A. Ragan, "Inferring synthetic lethal interactions from mutual exclusivity of genetic events in cancer," *Biology direct*, vol. 10, no. 1, p. 57, 2015.

[18] S. Sinha, D. Thomas, S. Chan, Y. Gao, D. Brunen, D. Torabi, A. Reinisch, D. Hernandez, A. Chan, E. B. Rankin *et al.*, "Systematic discovery of mutation-specific synthetic lethals by mining pan-cancer human primary tumor data," *Nature Communications*, vol. 8, 2017.

[19] T. Kranthi, S. Rao, and P. Manimaran, "Identification of synthetic lethal pairs in biological systems through network information centrality," *Molecular BioSystems*, vol. 9, no. 8, pp. 2163–2167, 2013.

[20] F. Zhang, M. Wu, X.-J. Li, X.-L. Li, C. K. Kwoh, and J. Zheng, "Predicting essential genes and synthetic lethality via influence propagation in signaling pathways of cancer cell fates," *Journal of bioinformatics and computational biology*, vol. 13, no. 03, p. 1541002, 2015.

[21] A. Jacunski, S. J. Dixon, and N. P. Tatonetti, "Connectivity homology enables inter-species network models of synthetic lethality," *PLoS Comput Biol*, vol. 11, no. 10, p. e1004506, 2015.

[22] S. L. Wong, L. V. Zhang, A. H. Tong, Z. Li, D. S. Goldberg, O. D. King, G. Lesage, M. Vidal, B. Andrews, H. Bussey *et al.*, "Combining biological networks to predict genetic interactions," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 44, pp. 15 682–15 687, 2004.

[23] B. Li, W. Cao, J. Zhou, and F. Luo, "Understanding and predicting synthetic lethal genetic interactions in saccharomyces cerevisiae using domain genetic interactions," *BMC systems biology*, vol. 5, no. 1, p. 73, 2011.

[24] G. Pandey, B. Zhang, A. N. Chang, C. L. Myers, J. Zhu, V. Kumar, and E. E. Schadt, "An integrative multi-network and multi-classifier approach to predict genetic interactions," *PLoS Comput Biol*, vol. 6, no. 9, p. e1000928, 2010.

[25] M. Wu, X. Li, F. Zhang, X. Li, C.-K. Kwoh, and J. Zheng, "In silico prediction of synthetic lethality by meta-analysis of genetic interactions, functions, and pathways in yeast and human cancer," *Cancer informatics*, vol. Suppl. 3, p. 71, 2014.

[26] R. Deshpande, M. K. Asiedu, M. Klebig, S. Sutor, E. Kuzmin, J. Nelson, J. Piotrowski, S. H. Shin, M. Yoshida, M. Costanzo *et al.*, "A comparative genomic approach for identifying synthetic lethal interactions in human cancer," *Cancer research*, vol. 73, no. 20, pp. 6128–6136, 2013.

[27] J. Guo, H. Liu, and J. Zheng, "Synlethdb: synthetic lethality database toward discovery of selective and sensitive anticancer drug targets," *Nucleic acids research*, vol. 44, no. D1, pp. D1011–D1017, 2016.

[28] Y. Shi, M. Larson, and A. Hanjalic, "Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges," *ACM Computing Surveys (CSUR)*, vol. 47, no. 1, p. 3, 2014.

[29] A. Ramlatchan, M. Yang, Q. Liu, M. Li, J. Wang, and Y. Li, "A survey of matrix completion methods for recommendation systems," *Big Data Mining and Analytics*, vol. 1, no. 4, pp. 308–323, 2018.

[30] H. Wang, H. Huang, C. Ding, and F. Nie, "Predicting protein–protein interactions from multimodal biological data sources via nonnegative matrix tri-factorization," *Journal of Computational Biology*, vol. 20, no. 4, pp. 344–358, 2013.

[31] Y. Liu, M. Wu, C. Miao, P. Zhao, and X.-L. Li, "Neighborhood regularized logistic matrix factorization for drug-target interaction prediction," *PLoS Comput Biol*, vol. 12, no. 2, p. e1004760, 2016.

[32] 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, p. 513, 2017.

[33] C. C. Johnson, "Logistic matrix factorization for implicit feedback data," *NIPS 2014 Workshop on Distributed Machine Learning and Matrix Computations*, 2014.

[34] Y. Wang, A. Kucukelbir, and D. M. Blei, "Robust probabilistic modeling with bayesian data reweighting," in *International Conference on Machine Learning*, 2017, pp. 3646–3655.

[35] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J Mach Learn Res*, vol. 12, pp. 2121–2159, 2011.

[36] Y. Liu, P. Zhao, X. Liu, M. Wu, L. Duan, and X.-L. Li, "Learning user dependencies for recommendation," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 2379–2385.

[37] A. K. Gupta and D. K. Nagar, *Matrix variate distributions*. CRC Press, 1999, vol. 104.

[38] X.-j. Li, S. K. Mishra, M. Wu, F. Zhang, and J. Zheng, "Syn-lethality: an integrative knowledge base of synthetic lethality towards discovery of selective anticancer therapies," *BioMed research international*, vol. 2014, 2014.

[39] E. E. Schmidt, O. Pelz, S. Buhlmann, G. Kerr, T. Horn, and M. Boutros, "Genomernai: a database for cell-based and in vivo rnai phenotypes, 2013 update," *Nucleic acids research*, vol. 41, no. D1, pp. D1021–D1026, 2013.

[40] T. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal *et al.*, "Human protein reference database—2009 update," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D767–D772, 2009.

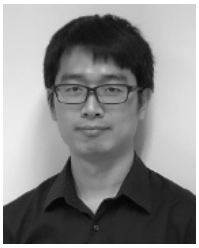
[41] H. N. Chua, W.-K. Sung, and L. Wong, "Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions," *Bioinformatics*, vol. 22, no. 13, pp. 1623–1630, 2006.

[42] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, "A new method to measure the semantic similarity of go terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, 2007.

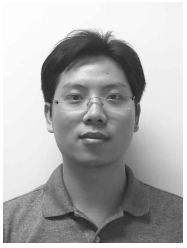
[43] J.-P. Mei, C.-K. Kwoh, P. Yang, X.-L. Li, and J. Zheng, "Drug–target interaction prediction by learning from local information and neighbors," *Bioinformatics*, vol. 29, no. 2, pp. 238–245, 2012.

[44] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 233–240.

[45] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 785–794.



**Yong Liu** is currently a Research Scientist in the Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), and Alibaba-NTU Singapore Joint Research Institute, Nanyang Technological University, Singapore. He received his B.S. from University of Science and Technology of China in 2008 and Ph.D. from Nanyang Technological University in 2016. He was a Data Scientist in NTUC Enterprise, Singapore from November 2017 to July 2018, and a Research Scientist in the Data Analytics Department at the Institute for Infocomm Research (I2R), A*STAR, Singapore from November 2015 to October 2017. His research areas include recommender systems, social media mining, and bioinformatics. His research papers appear in leading international conferences and journals. He has been invited as a PC member of major conferences such as KDD, IJCAI, AAAI, CIKM, ICDM, and reviewer for IEEE/ACM transactions. He is a member of ACM.

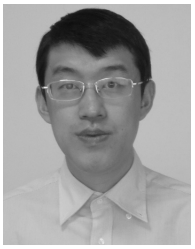


**Min Wu** is currently a Research Scientist in the Data Analytics Department at the Institute for Infocomm Research (I2R) under the Agency for Science, Technology and Research (A*STAR), Singapore. He received the B.Eng. from the University of Science and Technology of China (USTC), China in 2006 and his Ph.D. degree from Nanyang Technological University, Singapore in 2011. He received the best paper awards in the 15th International Conference on Bioinformatics (InCoB 2016) and the 20th International Conference on Database Systems for Advanced Applications (DASFAA 2015). He also won the IJCAI contest 2015 on repeated buyers prediction after sales promotion. His current research interests include machine learning, data mining and bioinformatics.

**Chenghao Liu** is a Postdoctoral Research Fellow in the School of Information Systems (SIS), Singapore Management University (SMU), Singapore. He received his Bachelor degree and Ph.D degrees from the Zhejiang Uniersity. His research interests include large-scale machine learning (online learning and deep learning) with application to tackle big data analytics challenges across a wide range of real-world applications.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2019.2909908, IEEE/ACM Transactions on Computational Biology and Bioinformatics





**Xiaoli Li** is currently head of Data Analytics Department at the Institute for Infocomm Research, A*STAR, Singapore. He also holds adjunct associate professor positions at the National University of Singapore and Nanyang Technological University. His research interests include data mining, machine learning and bioinformatics. He has served as a PC member/workshop chair/session chair in leading data mining related conferences (including KDD, ICDM, SDM, PKDD/ECML, PAKDD, WWW, AAAI, and CIKM) and as an editor of bioinformatics-related books. In 2005, he received the Best Paper Award in the 16th International Conference on Genome Informatics (GIW 2005). In 2011, he received the Best Paper Runner-Up Award in the 16th International Conference on Database Systems for Advanced Applications (DASFAA 2011). Xiaoli has published more than 170 papers, including top tier conferences such as KDD, ICDM, SDM, PKDD/ECML, ICML, IJCAI, AAAI, ACL, EMNLP, SIGIR, CIKM, UbiCom, etc. as well as some top tier journals such as IEEE Transactions TKDE, Bioinformatics and IEEE Transactions on Reliability.

**Jie Zheng** is an Associate Professor at the School of Information Science and Technology, ShanghaiTech University, Shanghai, China. He received his B. Eng (honors) in 2000 from Zhejiang University in China, and his Ph.D. in 2006 from the University of California, Riverside in USA, both in Computer Science. From 2006 to 2011, he was a Postdoctoral Visiting Fellow and Research Associate at the National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), USA. From Feb. 2011 to July 2018, he was an Assistant Professor at the School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore. Dr. Zheng's research interests include bioinformatics, computational genomics and systems biology, and biomedical data science. He develops novel computational methods (e.g. machine learning and data mining algorithms, artificial intelligence techniques, dynamical and data-driven models) to help answer biomedical questions. While trained as a Computer Scientist, Dr. Zheng maintains active and long-standing collaborations with Life Scientists.