A New Weighted Imputed Neighborhood-Regularized Tri-Factorization One-Class Collaborative Filtering Algorithm: Application to Target Gene Prediction of Transcription Factors

Hansaim Lim and Lei Xie

Abstract—Identifying target genes of transcription factors (TFs) is crucial to understand transcriptional regulation. However, our understanding of genome-wide TF targeting profile is limited due to the cost of large-scale experiments and intrinsic complexity of gene regulation. Thus, computational prediction methods are useful to predict unobserved TF-gene associations. Here, we develop a new Weighted Imputed Neighborhood-regularized Tri-Factorization one-class collaborative filtering algorithm, WINTF. It predicts unobserved target genes for TFs using known but noisy, incomplete, and biased TF-gene associations and protein-protein interaction networks. Our benchmark study shows that WINTF significantly outperforms its counterpart matrix factorization-based algorithms and tri-factorization methods that do not include weight, imputation, and neighbor-regularization, for TF-gene association prediction. When evaluated by independent datasets, accuracy is 37.8 percent on the top 495 predicted associations, an enrichment factor of 4.19 compared with random guess. Furthermore, many predicted novel associations are supported by literature evidence. Although we only use canonical TF-gene interaction data, WINTF can directly be applied to tissue-specific data when available. Thus, WINTF provides a potentially useful framework to integrate multiple omics data for further improvement of TF-gene prediction and applications to other sparse and noisy biological data. The benchmark dataset and source code are freely available at https://github.com/XieResearchGroup/WINTF.

Index Terms—Collaborative filtering, recommender system, tri-factorization, transcription factor, gene regulatory network

1 Introduction

Transcription factors (TFs) regulate gene expression via complex interactions with the target genes, and the regulations are crucial for cellular organizations and development. TFs can activate or deactivate the target genes by binding to the recognition DNA sequences, also known as motifs. TFs can interact with each other or recruit other protein components to form a protein complex to start transcription [1]. Such complex regulations explain the relative complexity of higher metazoans compared to lower organisms, such as unicellular eukaryotes or prokaryotes. The number of distinct genes itself cannot explain the complexity of organisms. It is known that the human genome contains only twice as many genes as Drosophila, and the difference is mainly from the duplication of the same gene rather than new ones [2]. Thus, the incredibly high complexity of humans cannot be understood without knowing the fact that human genome contains approximately one TF per every ten genes [3]. The complicated gene regulation by TFs seems to play an important role in development. In Drosophila, for example, deletion of one TF gene (Antennapedia) is known to cause a serious phenotypic defect – legs are on the head where antennae should be [4]. Therefore, understanding the associations between TFs and target genes is an important research topic in the biological and biomedical sciences.

Recent advancement of sequencing and molecular biology technology has led to laboratory techniques to identify TF-gene associations on a large scale, and the experimental data have been utilized for computational studies to integrate results from different experiments [5]. Chromatin immunoprecipitation (ChIP)-based methods include ChIP-chip [6], ChIP-seq [7], and ChIP-PET [8]. The sequences from ChIP methods are enriched around the binding sites for the TFs. Therefore, the target genes for TFs can be Inferred by mapping the sequence read peaks to the genome. Several studies have focused on identifying the true associations from the ChIP data by statistically comparing the sequence peaks to the background signal [9], [10], [11]. DamID is an alternative to ChIP techniques to identify TF-DNA interactions [12]. ChIP Enrichment Analysis (ChEA) [13] is a freely available tool that combines TF-DNA associations manually

- *H. Lim is with the PhD Program in Biochemistry, the City University of New York, 365 5th Avenue, New York, NY 10016 USA. E-mail: hlim1@gradcenter.cuny.edu.*
- *L. Xie is with the Department of Computer Science, Hunter College, and the Graduate Center, the City University of New York, 695 Park Ave, New York, NY 10065 USA and also with the Feil Family Brain & Mind Research Institute, Weill Cornell Medical College, Cornell University, 413 E 69th St, New York, NY 10021 USA. E-mail: lei.xie@hunter.cuny.edu.*

Manuscript received 27 Nov. 2018; revised 30 Aug. 2019; accepted 17 Jan. 2020. Date of publication 27 Jan. 2020; date of current version 3 Feb. 2021. (Corresponding author: Lei Xie.)
Digital Object Identifier no. 10.1109/TCBB.2020.2968442

curated and automatically collected from 115 publications for the ChIP-X experiments, which are the three ChIP techniques and DamID. ChEA takes a set of genes (whose expression levels are significantly changed) and finds the potential TFs that are likely to interact with most of the genes [13]. ChEA represents a reliable but incomplete resource for known TF-target gene associations; thus, it can be used as a benchmark for algorithm development.

Although the laboratory techniques mentioned above are essential for studying TF-DNA associations, they are not complete. Sequencing data from experiments contain noisy reads that are not necessarily indicating the TF-DNA interactions. In addition, it is well known that the quality of the antibody used in ChIP protocols are crucial for successful experiments [14]. The antibody specificity may be insufficient, or it could block successive interactions between TFs, making it difficult to observe indirect interactions. DamID is largely limited by its resolution as the GATC motifs are required, although it does not require the use of antibodies and therefore has advantages over ChIP protocols [15]. Thus, the currently known TF-gene associations are incomplete, biased, and noisy due to the limitations of experimental techniques. Computational tools to infer missing TF-gene associations are needed to gain comprehensive understanding of the gene regulations.

Collaborative filtering methods are a group of computational algorithms that are widely used in many areas to infer unobserved associations based on the observed ones with or without additional information [16]. The early generations of collaborative filtering methods are based on probabilistic models and aimed for business concerns, such as recommending products for users in Amazon.com and Barnes and Noble [17], [18]. First proposed by Paatero and Tapper in 1994, nonnegative matrix factorization (MF) [19] has been a popular choice for recommendation problems, especially after the development of fast multiplicative update rules by Lee and Seung [20], [21]. One of the most successful collaborative filtering applications is the popular Netflix challenge, where the user-video preferences are predicted using the past activity of the users [22]. The early collaborative filtering methods heavily rely on the availability of the information about past activity, and it is difficult to make predictions for users without a history of their choices. To overcome the drawback, later generation collaborative filtering methods attempt to utilize additional information, including user-user or item-item similarities [16]. Recently, Yao *et al.* developed an one-class collaborative filtering algorithm, wiZAN-dual, that utilizes both user-user and item-item similarity information as well as regularization and imputation parameters to improve prediction accuracy [23]. FASCINATE is an extension of wiZAN-dual on a multilayered network [24]. REMAP is an application of wiZAN-dual for biomedical problems [25]. REMAP predicts off-targets of drugs based on the drug-drug similarity and target-target similarity as well as the information about the known targets. In the comprehensive benchmark studies, REMAP outperforms other state-of-the-art methods. Thus, we will only use REMAP as a baseline for the performance evaluation.

As shown in REMAP, biomedical and biochemical association predictions can be modeled as collaborative filtering problems by replacing users with drugs and items with targets. Similarly, the unobserved TF-DNA associations can be predicted using REMAP. However, the drawback of matrix factorization (MF)-based collaborative filtering is that the factorized low-rank matrices for both users and items must have the same rank. That is, the user-side and item-side latent features must be in the same latent space, which is unrealistic, particularly if the number of users and items are very different. Moreover, the relationship between the user and the item is modeled by the inner product between two latent features. The inner product could be too simple to capture complex nonlinear relationships between two biological entities. In this study, we present a new weighted imputed neighborhood-regularized tri-factorization algorithm (WINTF), an extension of REMAP, which allows us to set different feature sizes for user and items as well as increase the power of modeling complex relationships among them. We apply WINTF to the target gene prediction of TF, in which the latent features of TFs and genes are set into different ranks. In the benchmark studies, WINTF achieves better prediction accuracy for TF-gene association prediction, compared with REMAP and vanilla tri-factorization method that do not use weight, imputation, and neighborhood-regularization. Many of our predicted novel associations are supported by evidences from the literature. Further application to tissue-specific TF-gene association prediction will significantly improve our understanding in transcriptional regulation.

2 Related Works and Contributions

This section is a review of the existing methods for target gene identification tools and relevant databases, followed by methods in the similar mathematical framework. Current TF-related studies mainly focus on prioritizing the TF-DNA binding peaks to collect the putative TF-gene associations from ChIP-X experiments and the databases for the collected TF-gene associations. To the best of our knowledge, there are few machine learning-based TF-gene prediction tools that take known TF-gene associations as input to predict unknown ones. Thus, the method proposed in this paper is the first machine learning algorithm for the target gene prediction of transcription factors.

Target identification from profile (TIP) is a probabilistic model that ranks target genes for TFs based on the relative binding signal strength from ChIP experiments, with an assumption that the binding signal is normally distributed [26]. Identifying target genes (iTAR) is an online server, which is designed to overcome the limitation from the normality assumption in TIP by applying Gaussian mixture model for p-value estimation [27]. Covariance based extraction of regulatory targets using multiple time series (CERMT) predicts TF target genes under an assumption that the true target genes for TFs will show similar response pattern to the TFs [28]. Targetfinder is a tool to predict target genes based on the assumption that the genes with similar expression profiles are likely to be regulated by the same TFs [29]. These methods either take ChIP experimental data as input or utilize gene expression data to compare the input genes. A recent study combining these ideas predicts functional TF-gene associations by correlating ChIP data and gene expression profiles [30].

TRANSFAC, a database for TF-gene interactions from experimental data, has been managed and updated to adopt

new data across different organisms as well as tissue-specific regulations [31], [32]. In addition to the information about TFs, their binding sites, and target genes, TRANSFAC database now contains information about the control of gene expressions, the source cell line for TFs, and binding sites for different experimental conditions, if available [33]. JASPAR is another TF-gene database for matrix-based TF binding sites from published experimental results [34]. JASPAR was recently updated to include multiple species in six taxonomic groups [35]. The encyclopedia of DNA elements (ENCODE) project, initiated in 2004, aimed to identify all functional elements in the human genome, which includes TF-gene associations [36]. ChEA provides a large collection of TF-gene association data manually curated and computationally extracted from over 100 publications for ChIP-X experiments [13]. TRRUST is a more recently developed database for human TF-gene associations from text-mining a massive amount of literature abstracts [37]. TRRUST version 2 contains TF-gene associations in mice as well as more associations for humans [38]. TRANSFAC, JASPAR, ENCODE project and ChEA databases are listed in Harmonizome, an integrated knowledgebase about genes and proteins, developed to facilitate access to and learning from a large amount of biomedical data [39]. Human transcriptional regulation interactions database (HTRIdb) is claimed to be a freely available database containing experimentally verified human TF-gene associations [40].

As reviewed in the introduction, MF-based models have been applied to infer unknown associations such as unobserved drug-target binding. SymNMF is an MF-based method to integrate and infer missing similarity information between drugs and targets from multiple sources [41]. MTF differs from MF in that the input matrix is factorized into three smaller matrices (e.g., matrices F, G, and S in Table 1), instead of two (e.g., matrices F and G where $r = s$). Unlike aforementioned MF or MTF-based ranking methods, which heuristically optimizes the feature sizes (e.g., r and s in Table 1), MTF-based supervised clustering fixes the feature sizes and regularizes the network by prior knowledge. Hwang et al. developed an MTF-based clustering method (R-NMTF) for disease phenotypes and genes regularized by phenotype similarity and protein-protein interaction data [42]. Park *et al.* developed NTriPath to cluster cancer types and genes regularized by protein-protein interaction data [43]. While the output clusters may be used for certain ranking tasks, these methods require prior knowledge in the number of clusters and correct cluster labels in addition to the inputs for M(T)F- based ranking methods. The matrix tri-factorization has been applied to gene function prediction, patient stratification, and disease module detection [44], [45].

Compared with existing work, our contributions in this paper include:

- We develop a new algorithm WINTF, which for the first time incorporates sample weight, imputation, and side information into the existing tri-factorization frameworks [44], [45], making it better handle noisy and sparse data.
- We develop an efficient optimization algorithm based on the multiplicative update rule.

TABLE 1
Symbol Definitions and Descriptions

Symbols	Definition
$m,\ n$	Number of unique genes and TFs.
$r,\ s$	Feature sizes for genes and transcription factors, respectively. $r < m$, and $s < n$.
w	Scalar reliability weight. $w \in [0,\ 1]$
p	Scalar imputation score. $p \in [0,\ 1]$
$\mathbf{\Theta},\ \mathbf{\Theta}^c$	Set of observed and unobserved associations.
$R_{(i,j)}$	Element at i^{th} row and j^{th} column of matrix R.
R	Known association matrix. $R_{(i,j)} = 1\ if\ (i,j) \in \mathbf{\Theta},\ 0\ otherwise.\ R \in \mathbb{R}^{m \times n}$
F	Low-rank feature matrix for genes. $F \in \mathbb{R}^{m \times r}$
G	Low-rank feature matrix for TFs. $G \in \mathbb{R}^{n \times s}$
S	Low-rank feature interaction matrix. $S \in \mathbb{R}^{r \times s}$
M	Gene-gene similarity score matrix. It is a symmetric, positive matrix. $M \in \mathbb{R}^{m \times m}$
N	TF similarity score matrix, defined similarly to M. $N \in \mathbb{R}^{n \times n}$
$D_M,\ D_N$	Degree matrices for M and N, respectively. D_M and D_N are diagonal, positive matrices.
W	Weight matrix. $W_{(i,j)} = 1\ if\ (i,j) \in \mathbf{\Theta},\ w\ otherwise.\ W \in \mathbb{R}^{m \times n}$
P	Imputation matrix. $P_{(i,j)} = 0\ if\ (i,j) \in \mathbf{\Theta},\ p\ otherwise.\ P \in \mathbb{R}^{m \times n}$
$1_{m \times n}$	Indicator matrix containing 1 at every position. $1_{m \times n} \in \mathbb{R}^{m \times n}$
λ_r	Regularization parameter. $\lambda_r \in [0,\ 1]$
$\lambda_F,\ \lambda_G$	Importance weights for genes and TFs.
$tr(M)$	Trace of matrix M.
$\|M\|$	Frobenius norm of matrix M.

Matrices are capitalized and italicized, and scalars are in lower cases.

- For the first time, we develop an accurate machine learning algorithm for the target gene prediction of transcription factors.

3 METHODS

3.1 Prediction Method Description

In this section, we first present a mathematical formulation of the one-class collaborative filtering problem. The optimization function for our prediction method, weighted and imputed neighborhood-regularized tri-factorization (WINTF) in Eq. (1) with the symbols described in Table 1. Then, we explain how WINTF differs from REMAP, a single-ranked version of WINTF. We also present the update rules for WINTF, based on the multiplicative update rule by Lee and Seung [20].

$$\begin{aligned} J = & \sum_{(u,i)} W_{(u,i)} \left(R_{(u,i)} + P_{(u,i)} - (FSG^T)_{(u,i)} \right)^2 \\ & + \lambda_r (\|F\|^2 + \|S\|^2 + \|G\|^2) \\ & + \lambda_F tr(F^T (D_M - M) F) \\ & + \lambda_G tr(G^T (D_N - N) G). \end{aligned} \tag{1}$$

The problem WINTF solves is to find the nonnegative low-rank matrices F, S, and G that minimizes the optimization function in Eq. (1). The optimization function above consists of four terms. Although the formula is slightly different from that for REMAP, most ideas in the function are the same. The shared ideas are explained in the following paragraph.

WINTF is an extension of REMAP. REMAP [25] was applied to predict off-target drug-gene associations based on the wiZAN-dual algorithm [23]. REMAP and WINTF share several ideas. They take the known user-item (drug-target in REMAP application) associations with user-user similarity scores and item-item similarity scores. The inputs are therefore three matrices: user-item association, user-user similarity, and item-item similarity matrices. The core MF algorithm tries to find the low-rank matrices containing the feature vector representations of users and items, such that the inner product of the matrices reconstructs the known association matrix. WINTF and REMAP also commonly take a penalty weight, an imputation value, a regularization parameter, and importance weights for user-user and item-item similarity information as user-defined parameters. The penalty weight indicates the reliability of the known associations, and the imputation value indicates the probability of unknown associations to be positive. They can be either obtained from a priori knowledge, such as the false positive rate of high-throughput experiments or tuned as hyperparameters. The last two terms in Eq. (1) accounts for the homophily effect (i.e., similar users prefer similar items). The two importance weights λ_F and λ_G control how much the corresponding similarity scores affect the optimization. In both WINTF and REMAP the homophily effect is an important idea. The similarity scores, which can be measured by external methods (e.g., chemical structural similarity for different drugs), are used to reflect the homophily effect by updating the low-rank matrices so that the feature vectors for two similar users or items are close in their euclidean distance. More details about the design of the optimization function are available in references [25], [46].

The key difference between WINTF and REMAP is that WINTF finds three low-rank matrices to approximate the known association matrix, while REMAP and other traditional MF methods find only two. The optimization function for REMAP can be obtained by removing the matrix S in the Eq. (1). Without the matrix S, however, one can easily see that the matrix inner product FG^T must be in the same dimension as the known association matrix R. Thus, the matrices F and G must have the same rank, meaning the feature size for both users and items are the same. The single-rank constraint is undesirable unless the actual feature sizes are coincidentally identical. The feature interaction matrix S makes it possible to set the rank of F different from that of G. By introducing the matrix S into the traditional MF methods, better predictive performances are expected due to more flexible choices for feature sizes. An additional feature interaction matrix S necessarily increase the running time for the algorithm. In a later section we show that the increased computational cost is affordable in most modern computers.

The optimization algorithm for WINTF is based on the multiplicative update rule [20], similar to the algorithm for REMAP [25]. As in other MF problems, the optimization problem in Eq. (1) is not convex due to the coupling of F, S, and G. Therefore, the multiplicative update rule finds a fixed-point solution for a local optimum of the problem with the nonnegativity constraint. The update rules for the three low-rank matrices are the following.

$$F_{(u,r)} \leftarrow F_{(u,r)} \sqrt{\frac{\left[(1-wp)RGS^T + wp1_{\boldsymbol{m}\times\boldsymbol{n}}GS^T + \lambda_F MF\right]_{(u,r)}}{\left[(1-w)\hat{R}_{\Theta}GS^T + wF(SG^TGS^T) + \lambda_r F + \lambda_F D_M F\right]_{(u,r)}}} \tag{2}$$

$$G_{(i,s)} \leftarrow G_{(i,s)} \sqrt{\frac{\left[(1-wp)R^TFS + wp1_{m\times n}FS + \lambda_G NG\right]_{(i,s)}}{\left[(1-w)\hat{R}_{\Theta}FS + wG(S^TF^TFS) + \lambda_r G + \lambda_G D_N G\right]_{(i,s)}}} \tag{3}$$

$$S_{(r,s)} \leftarrow S_{(r,s)} \sqrt{\frac{\left[(1-wp)F^TRG + wp(F^T1_{m\times n}G)\right]_{(r,s)}}{\left[(1-w)F^T\hat{R}_{\Theta}G + wF^T(FSG^T)G + \lambda_r S\right]_{(r,s)}}}. \tag{4}$$

The predicted score matrix for known associations, $\hat{R}_{\Theta}$, is defined as follows.

$$\hat{R}_{\Theta\ (u,i)} = \begin{cases} \left(FSG^T\right)_{(u,i)} & if\ (u,i) \in \boldsymbol{\Theta} \\ 0 & otherwise \end{cases}. \tag{5}$$

Note that we use a global scalar weight w and a global scalar imputation p, instead of position-specific weight matrix W and imputation matrix P. The update rules above are derived by considering the partial derivatives with regard to each low-rank matrix, while considering the other two low-rank matrices constant. Therefore, we update the low-rank matrices one at a time, while not changing the other two. The update process under the multiplicative update rules can be described as a gradient descent method with specially designed learning rates. The pseudocode for our method is below in Algorithm 1.

Algorithm 1. WINTF

1. **Input matrices:** M, N, and R
   Input scalars: $\Omega = \{r, s, w, p, \lambda_r, \lambda_F, \lambda_G\}$ and max_iter
2. **Define:** f_2, f_3, f_4, f_5 from Eqs. (2), (3), (4), and (5) above
3. **Initialize:** $F_0, S_0, G_0 \leftarrow RandomUniformMatrix \in [0,1]$
4. **for** $i = 0$ to (max_iter -1)
5. $\quad \widehat{R}_{\Theta,i} \leftarrow f_5(R, F_i, S_i, G_i)$
6. $\quad F_{i+1} \leftarrow f_2(F_i, S_i, G_i, \widehat{R}_{\Theta,i}, \Omega)$
7. $\quad G_{i+1} \leftarrow f_3(F_i, S_i, G_i, \widehat{R}_{\Theta,i}, \Omega)$
8. $\quad S_{i+1} \leftarrow f_4(F_i, S_i, G_i, \widehat{R}_{\Theta,i}, \Omega)$
9. $\quad$ replace NaNs in $F_{i+1}, G_{i+1}, S_{i+1}$ with 0
10. **Return:** real-valued matrices F, S, G

Once the updates are complete, $\hat{R}$, the prediction score matrix for all TF-gene associations can be calculated by the inner product of the three low-rank matrices. The prediction score matrix for unknown associations, $\hat{R}_{\Theta^c}$, can be obtained by subtracting $\hat{R}_{\Theta}$ from $\hat{R}$, which contains prediction scores for both known and unknown associations.

$$\hat{R}_{\Theta^c} = \hat{R} - \hat{R}_{\Theta},\ where\ \ \hat{R} = FSG^T.$$

Fig. 1 summarizes the process of predicting an unknown gene-TF association (Gene_3 and TF_1 in the figure) by WINTF. The derivation and proof of the update rules with

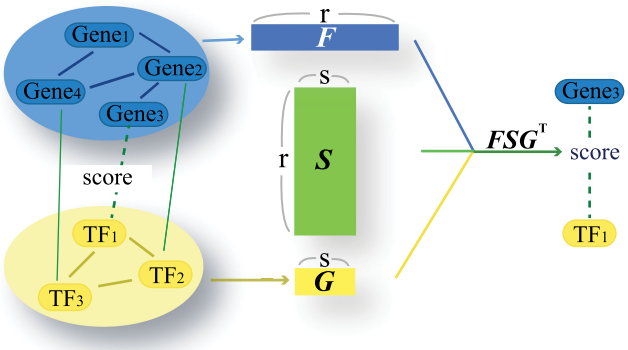


Fig. 1. WINTF concept figure. WINTF takes gene-TF, gene-gene, and TF-TF networks as input, and updates the three low-rank feature matrices, F, S, and G. Once WINTF update process is complete (Algorithm 1), the prediction score for an unknown gene-TF association $(\mathrm{Gene}_3 - \mathrm{TF}_1)$ is calculated by a dot product of the corresponding low-rank features. Blue and yellow rectangles represent the feature vectors for the gene and TF, respectively. In other words, blue rectangle is $F_{(3,:)}$ and yellow rectangle is $G_{(1,:)}$.

the justification of using scalar weight and imputation values are in the following section.

3.2 Theorems and Proofs

In the Section 3.1, we proposed the update rules for the three low-rank matrices that find a local minimum of the cost function J defined in Eq. (1). The cost function J is non-convex. Thus, we update one low-rank matrix at a time, while considering the others as constant. When S and G are fixed, the cost function J can be simplified to J_F.

$$J_F = \left\|W \odot ((R+P) - FSG^T)\right\|^2 + \lambda_r \|F\|^2 + \lambda_F tr(F^T(D_M - M)F), \qquad s.t.\ F \geq 0. \tag{6}$$

The partial derivative of the J_F with regards to F is the following.

$$\frac{1}{2}\nabla_F J = \frac{1}{2}\frac{\partial J_F}{\partial F} = -W \odot W \odot (R+P)GS^T - \lambda_F MF + W \odot W \odot FSG^T GS^T + \lambda_F F + \lambda_F D_M F. \tag{7}$$

Based on the multiplicative update rule proposed by Lee and Seung [20], we obtain the update rule for F as follows.

$$F_{(u,r)} \leftarrow F_{(u,r)} \sqrt{\frac{[W \odot W \odot (R+P)GS^T + \lambda_F MF]_{(u,r)}}{[W \odot W \odot (FSG^T)GS^T + \lambda_r F + \lambda_F D_M F]_{(u,r)}}} \tag{8}$$

In the method section, we proposed simplified update rules that reduce the computational complexity from the large dimension of the weight and imputation matrices. The simplified update rule for Eq. (8) is the Eq. (2).

In the remainder of this section, we first show that the fixed-point solution of Eq. (8) satisfies the KKT condition, and that Eqs. (2) and (8) are mathematically equivalent. Then, we show that the cost function in Eq. (6) decreases monotonically under the update rule in Eq. (8).

Theorem 1. *The fixed-point solution of Eq. (8) satisfies the KKT condition.*

Proof. The Lagrangian of Eq. (6) is the following (Λ is the Lagrange multiplier).

$$L_{J_F} = \left\|W \odot ((R+P) - FSG^T)\right\|^2 + \lambda_r \|F\|^2 + \lambda_F tr((F^T D_M F) - (F^T MF)) - tr(\Lambda F). \tag{9}$$

□

Let $\frac{\partial}{\partial F} L_{J_F} = 0$, we obtain the following.

$$2(-W \odot W \odot (R+P)GS^T + W \odot W \odot FSG^T GS^T + \lambda_r F + \lambda_F D_M F - \lambda_F MF) = \Lambda \tag{10}$$

From the KKT complementary slackness condition, we obtain the following.

$$[-W \odot W \odot (R+P)GS^T + W \odot W \odot FSG^T GS^T + \lambda_r F + \lambda_F D_M F - \lambda_F MF]_{(u,r)} \odot F_{(u,r)} = 0. \tag{11}$$

Eq. 12 is the fixed-point solution of Eq. (8), which satisfies Eq. (11).

$$[W \odot W \odot (R+P)GS^T + \lambda_F MF]_{(u,r)} = [W \odot W \odot (FSG^T)GS^T + \lambda_r F + \lambda_F D_M F]_{(u,r)}. \tag{12}$$

Next, we show that Eq. (2) is equivalent to Eq. (8). We use $1^{\mathrm{A}}, 1^{\Theta}$, and 1^{Θ^c} as the indicator matrices for full, observed, and unobserved data, respectively, so that $1_{m\times n} = 1^{\mathrm{A}} = 1^{\Theta} + 1^{\Theta^c}$, and $1^{\Theta} = \mathrm{R}$. Based on that, the weight and imputation values are for unobserved associations only, the equations below turn the weight matrix W and imputation matrix P into scalar weight w and scalar imputation value p, respectively. Note that the weight matrix W contains the square root of the global weight w on unobserved positions and zero on observed ones.

$$\begin{aligned} &W \odot W \odot (R+P)GS^T \\ &= (\mathbf{1}^{\Theta} \odot R + wp \cdot \mathbf{1}^{\Theta^c})GS^T = (R + wp \cdot \mathbf{1}^{A} - wp \cdot \mathbf{1}^{\Theta})GS^T \\ &= (1 - \mathrm{wp}) \cdot RGS^T + wp \cdot \mathbf{1}_{m\times n}\mathrm{GS}^T, \end{aligned}$$

and

$$\begin{aligned} &(W \odot W \odot FSG^T)GS^T \\ &= W \odot W \odot (\widehat{R}_{\Theta} + \widehat{R}_{\Theta^C})GS^T = (\widehat{R}_{\Theta} + w \cdot 1^{\Theta^c} \odot \widehat{R}_{\Theta^C})GS^T \\ &= (1-w)\widehat{R}_{\Theta}GS^T + w \cdot FSG^T GS^T. \end{aligned}$$

Substituting the two equations above into Eq. (8) proves that Eq. (2) is equivalent to Eq. (8).

Theorem 2. *The cost function in Eq. (6) decreases monotonically under the update rule in Eq. (8).*

Proof. To prove Theorem 2, we start from the cost function Eq. (6). □

According to the auxiliary function strategy [47], $H(\mathrm{F}, \tilde{\mathrm{F}})$ is an auxiliary function of $J(\mathrm{F})$ if it satisfies the following conditions.

$$H(F,F) = J(F), \quad and \quad H(F,\widetilde{F}) \geq J(F). \tag{13}$$

Defining $\mathrm{F}^{(t+1)} = arg\min_F H(\mathrm{F}, \mathrm{F}^{(t)})$ proves that $J(\mathrm{F}^{(t)})$ monotonically decreases since the following condition is met by the design of the auxiliary function.

$$J\left(\mathrm{F}^{(t)}\right) = H\left(\mathrm{F}^{(t)}, \mathrm{F}^{(t)}\right) \geq H\left(\mathrm{F}^{(t+1)}, \mathrm{F}^{(t)}\right) \geq J\left(\mathrm{F}^{(t+1)}\right). \tag{14}$$

We first find an auxiliary function satisfying the conditions in Eq. (13), and then solve for the auxiliary function, which is the global minimum of the auxiliary function.

$$\begin{aligned}
&H(F, \widetilde{F}) \\
&= -2\sum_{u=1}^{m}\sum_{k=1}^{r}[(W \odot W \odot (R+P))GS^T]_{(u,k)}\widetilde{F}_{(u,k)}\left(1+\log\left(\frac{F_{(u,k)}}{\widetilde{F}_{(u,k)}}\right)\right) \\
&\quad -\sum_{u=1}^{m}\sum_{v=1}^{m}\sum_{k=1}^{r}\lambda_F M_{(u,v)}\widetilde{F}_{(v,k)}\widetilde{F}_{(u,k)}\left(1+\log\left(\frac{F_{(v,k)}F_{(u,k)}}{\widetilde{F}_{(v,k)}\widetilde{F}_{(u,k)}}\right)\right) \\
&\quad +\sum_{u=1}^{m}\sum_{k=1}^{r}\lambda_r F_{(u,k)}^2 + \sum_{u=1}^{m}\sum_{k=1}^{r}\frac{[(W \odot W \odot \widetilde{F}SG^T)GS^T]_{(u,k)}F_{(u,k)}^2}{\widetilde{F}_{(u,k)}} \\
&\quad +\sum_{u=1}^{m}\sum_{k=1}^{r}\frac{[\lambda_F D_M \widetilde{F}]_{(u,k)}F_{(u,k)}^2}{\widetilde{F}_{(u,k)}}.
\end{aligned} \tag{15}$$

It is trivial to show $H(F,F) = J(F)$. To show $H(F,\widetilde{F}) \geq J(F)$, we name the five terms in Eq. (15) as $H1, H2, H3, H4$, and $H5$, respectively. Then, using the inequality $x \geq 1 + \log(x)$, the H1 becomes the following.

$$\begin{aligned}
H1 &\geq -2\sum_{u=1}^{m}\sum_{k=1}^{r}[(W \odot W \odot (R+P))GS^T]_{(u,k)}F_{(u,k)} \\
&= -2tr[(W \odot W \odot (R+P))GS^T F]
\end{aligned} \tag{16}$$

$$H2 \geq -\sum_{u=1}^{m}\sum_{v=1}^{m}\sum_{k=1}^{r}\lambda_F M_{(u,v)}F_{(v,k)}F_{(u,k)} = -\lambda_F tr(F^T MF). \tag{17}$$

Then, for $H3$ we get

$$H3 = \lambda_r tr(FF^T). \tag{18}$$

For $H4$, let $\mathrm{F}_{(u,k)} = \tilde{\mathrm{F}}_{(u,k)}\ \mathrm{Q}_{(u,k)}$ and we have the following.

$$\begin{aligned}
H4 &= \sum_{u=1}^{m}\sum_{i=1}^{n}\sum_{k=1}^{r}\sum_{l=1}^{r}\frac{\widetilde{F}_{(u,l)}(SG^T)_{(i,l)}W_{(u,i)}^2(GS^T)_{(i,k)}F_{(u,k)}^2}{\widetilde{F}_{(u,k)}} \\
&= \sum_{u=1}^{m}\sum_{i=1}^{n}\sum_{k=1}^{r}\sum_{l=1}^{r}\widetilde{F}_{(u,l)}(SG^T)_{(i,l)}W_{(u,i)}^2(GS^T)_{(i,k)}\widetilde{F}_{(u,k)}Q_{(u,k)}^2 \\
&= \sum_{u=1}^{m}\sum_{i=1}^{n}\sum_{k=1}^{r}\sum_{l=1}^{r}\widetilde{F}_{(u,l)}(SG^T)_{(i,l)}W_{(u,i)}^2(GS^T)_{(i,k)}\widetilde{F}_{(u,k)}\left(\frac{Q_{(u,k)}^2+Q_{(u,l)}^2}{2}\right) \\
&\geq \sum_{u=1}^{m}\sum_{i=1}^{n}\sum_{k=1}^{r}\sum_{l=1}^{r}\widetilde{F}_{(u,l)}(SG^T)_{(i,l)}W_{(u,i)}^2(GS^T)_{(i,k)}\widetilde{F}_{(u,k)}(Q_{(u,k)}+Q_{(u,l)}) \\
&= \sum_{u=1}^{m}\sum_{i=1}^{n}\sum_{k=1}^{r}\sum_{l=1}^{r}F_{(u,l)}(SG^T)_{(i,l)}W_{(u,i)}^2(GS^T)_{(i,k)}F_{(u,k)} \\
&= tr[(W \odot W \odot (FSG^T))GS^T F^T]
\end{aligned} \tag{19}$$

We use the inequality below, where $\mathrm{A}_{n\times n}$, $\mathrm{B}_{k\times k}$, $\mathrm{S}_{n\times k}$, and $\mathrm{S}^*_{n\times k}$ are nonnegative, and A and B are symmetric [48].

$$\sum_{i=1}^{n}\sum_{p=1}^{k}\frac{(AS^*B)S_{(i,p)}^2}{S_{(i,p)}^*} \geq tr(S^T ASB)$$

$$Thus, H5 = \sum_{u=1}^{m}\sum_{k=1}^{r}\frac{[\lambda_F D_M \widetilde{F}]_{(u,k)}F_{(u,k)}^2}{\widetilde{F}_{(u,k)}} \geq \lambda_F tr(F^T D_M F). \tag{20}$$

Substituting Eqs. (16), (17), (18), (19), (20) into Eq. (15) shows that the auxiliary function satisfies the second condition in Eq. (13). The gradient of the auxiliary function is the following.

$$\begin{aligned}
&\frac{1}{2}\frac{\partial H(F,\widetilde{F})}{\partial F_{(u,k)}} \\
&= -\frac{[W \odot W \odot (R+P)GS^T]_{(u,k)} \cdot \widetilde{F}_{(u,k)}}{F_{(u,k)}} \\
&\quad -\frac{[\lambda_F M\widetilde{F}]_{(u,k)} \cdot \widetilde{F}_{(u,k)}}{F_{(u,k)}} + \frac{\lambda_r F_{(u,k)} \cdot \widetilde{F}_{(u,k)}}{\widetilde{F}_{(u,k)}} \\
&\quad +\frac{[W \odot W \odot \widetilde{F}SG^T GS^T]_{(u,k)} \cdot F_{(u,k)}}{\widetilde{F}_{(u,k)}} + \frac{[\lambda_F D_M \widetilde{F}]_{(u,k)} \cdot F_{(u,k)}}{\widetilde{F}_{(u,k)}} \\
&= -\frac{[W \odot W \odot (R+P)GS^T + \lambda_F M\widetilde{F}]_{(u,k)} \cdot \widetilde{F}_{(u,k)}}{F_{(u,k)}} \\
&\quad +\frac{[W \odot W \odot \widetilde{F}SG^T GS^T + \lambda_r \widetilde{F} + \lambda_F D_M \widetilde{F}]_{(u,k)} \cdot F_{(u,k)}}{\widetilde{F}_{(u,k)}}
\end{aligned} \tag{21}$$

The Hessian of $H(\mathrm{F}, \tilde{\mathrm{F}})$ is a diagonal matrix with positive diagonal elements. Thus, we can obtain the global minimum by setting Eq. (21) to be zero, which results in the following solution.

$$F_{(u,k)}^2 = \widetilde{F}_{(u,k)}^2 \cdot \frac{[W \odot W \odot (R+P)GS^T + \lambda_F M\widetilde{F}]_{(u,k)}}{[W \odot W \odot (\widetilde{F}SG^T GS^T + \lambda_r \widetilde{F} + \lambda_F D_M \widetilde{F}]_{(u,k)}} \tag{22}$$

Setting $F^{(t+1)} = F$ and $F^{(t)} = \widetilde{F}$ proves that the update rule Eq. (8) monotonically decreases the cost function. With the equivalence between Eqs. (2) and (8), (6) monotonically decreases under the update rule Eq. (2).

The update rule for G can be proved analogously to the proof above. The matrix S-equivalent of the cost function Eq. (6) is the following.

$$\begin{aligned}
J(S) = &tr(-2W \odot W \odot (R+P)GS^T F^T) \\
&+ tr(W \odot W \odot (FSG^T)GS^T F^T) + \lambda_r tr(S^T S).
\end{aligned}$$

Therefore, we choose an auxiliary function for matrix S, which is missing two terms corresponding to $H2$ and $H5$ in Eq. (15). The auxiliary function and its gradient are the following.

$$
\begin{aligned}
&H(S,\widetilde{S})\\
&= -2\sum_{i=1}^{r}\sum_{j=1}^{s}[F^T(W\odot W\odot(R+P))G]_{(i,j)}\cdot\widetilde{S}_{(i,j)}\cdot\left(1+\log\frac{S_{(i,j)}}{\widetilde{S}_{(i,j)}}\right)\\
&\quad+\sum_{i=1}^{r}\sum_{j=1}^{s}\lambda_r S_{(i,j)}^2\\
&\quad+\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{[F^T(W\odot W\odot(F\widetilde{S}G^T))G]_{(i,j)}\cdot S_{(i,j)}^2}{\widetilde{S}_{(i,j)}}\frac{1}{2}\cdot\frac{\partial H(S,\widetilde{S})}{\partial S_{(i,j)}}\\
&= -\frac{[F^T(W\odot W\odot(R+P))G]_{(i,j)}\cdot\widetilde{S}_{(i,j)}}{S_{(i,j)}}+\frac{[\lambda_r\widetilde{S}]_{(i,j)}\cdot S_{(i,j)}}{\widetilde{S}_{(i,j)}}\\
&\quad+\frac{[F^T(W\odot W\odot(F\widetilde{S}G^T))G]_{(i,j)}\cdot S_{(i,j)}}{\widetilde{S}_{(i,j)}}
\end{aligned}
$$

Setting the gradient to zero, we obtain the global minimum solution.

$$S_{(i,j)}^2=\widetilde{S}_{(i,j)}^2\cdot\frac{[F^T(W\odot W\odot(R+P))G]_{(i,j)}}{[F^T(W\odot W\odot(F\widetilde{S}G^T))G+\lambda_r\widetilde{S}]_{(i,j)}}.$$

Combining the theorems 1 and 2, we proved that the proposed update rules satisfy the KKT condition and converge to the solution.

4 Experimental Setup

The TF-gene association data of our choice in this study is from ChEA, which contains manually curated as well as computationally extracted associations from more than 100 publications for ChIP-X experiments [13]. The ChEA dataset contains 386,776 TF-gene associations for 21,585 genes and 199 TFs for human (approximate density 9 percent). The gene-gene and TF-TF similarity matrices are calculated by assuming two interacting proteins are related to each other. The protein-protein interaction data is from the STRING database, which contains experimentally known and computationally predicted protein-protein physical interactions and functional associations (e.g., co-expressed genes) with reliability scores [49]. Together, they are termed as protein-protein interaction (PPI). The similarity score between the i^{th} and j^{th} genes (or TFs) is the PPI reliability score divided by the maximum available score (1,000). If multiple reliability scores exist for a pair of proteins, they are averaged. This makes all similarity scores in between 0 and 1, standing for minimum and maximum similarity, respectively. Sequence-based protein-protein similarity scores can be used as it was done in the REMAP application [25]. Two proteins will have a high similarity score if their BLAST [50] alignment returns a high score. As a result, 9,207,162 and 12,775 PPI-based non-zero scores are obtained for gene-gene and TF-TF similarity, respectively.

To compare WINTF with REMAP, we evaluated the performances of the two methods for the ChEA dataset described above. We performed 5-fold cross validations to measure four different performance metrics: area under receiver operating characteristic curve (AUC), mean average precision (MAP), half-life utility (HLU), and mean percentile rank (MPR). AUC is one of the most widely used performance measurements that measures how quickly an algorithm achieves a high true positive rate while keeping low false positive rates. HLU measures the likelihood that a user accepts recommendation if the likelihood exponentially decreases with the ranking of recommended items [51]. MAP measures the average precision for all users at different true positive rates [52]. MPR is the average percentile rank of positive associations in the test samples [53]. The higher AUC, MAP, HLU, and the lower MPR, the better performance. We compared the performance with and without the similarity score matrices derived from protein-protein interactions.

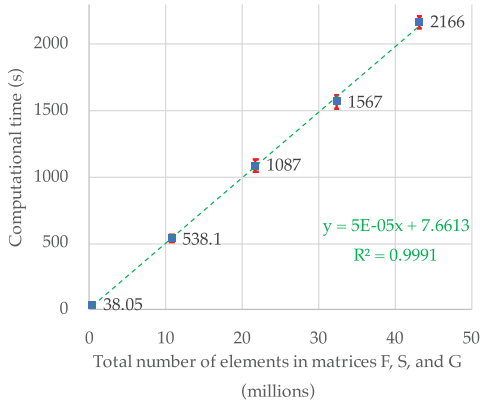


Fig. 2. Time complexity of WINTF using ChEA data set. Average running time (blue squares) from five runs of WINTF are plotted according to the total number of elements in the three low-rank matrices, F, S, and G. The *x-axis* is the total number of elements in the three matrices in millions (i.e., $m\times r+r\times s+n\times s$). Within the tested range, WINTF shows approximately linear time complexity (green dashed line) to the number of matrix elements. Error bars represent $\pm$ five times the standard deviations.

5 Results

5.1 Time Complexity

Compared to REMAP, WINTF necessarily require longer computational time for an additional matrix,S. As we have done in our previous application using WINTF algorithm [54], we measured computational time of WINTF using the ChEA data set. Fig. 2 shows the average running time according to increasing number of total elements in the matrices F, S, and G. The empirical time complexity of WINTF suggests that the running time approximately linearly increases as the number of matrix elements increases. It also shows that WINTF jobs on ChEA data set takes approximately 1,100 seconds at the default rank parameters ($r=1000, s=100$).

5.2 Benchmark Evaluation of Prediction Accuracy

Our benchmark tests under different conditions (e.g., different parameters and with/without similarity information) show WINTF outperforms REMAP under all tested conditions (Tables 2 and 3). Table 2 shows that regardless of the similarity matrices used, WINTF performs significantly better than REMAP in all four metrics. Table 3 shows that the rank parameters affect the performances of both algorithms, and that WINTF outperforms REMAP under any tested hyper-parameters. Due to the number of parameters for both algorithms, it is impractical to compare the two algorithms with all possible combinations. Thus, we tested a limited number of combinations, evaluating the usefulness of different similarity measurements, and the effect of an additional low-rank

TABLE 2
Performance Comparison for WINTF and REMAP With Different Similarity Information

[1]Con.	Algo.	AUC	HLU	MAP	MPR
A	WINTF	.763(8.0e-4)	40.7(.394)	.295(.003)	.259(5.0e-4)
	REMAP	.717(.001)	31.1(.667)	.231(.002)	.300(.001)
B	WINTF	.766(6.0e-4)	40.7(.23)	.295(2.6e-3)	.259(4.6e-4)
	REMAP	.726(1.7e-3)	32.0(1.7)	.237(7.4e-3)	.294(1.5e-3)
C	WINTF	.762(7.8e-4)	40.7(.394)	.295(.003)	.259(5.0e-4)
	REMAP	.727(0.002)	33.6(.150)	.243(.001)	.291(.002)
D	WINTF	.762(7.8e-4)	40.7(.394)	.295(.003)	.259(5.0e-4)
	REMAP	.717(.001)	30.1(.672)	.231(.002)	.300(.001)
E	Vanilla MTF	.500(0.0)	5.96(.06)	.155(5.0e-4)	.320(6.6e-4)

Values are mean and (standard deviation) for 5-fold cross validation
*[1]**Condition A**: TF similarity scores are based on sequence similarity only, and gene similarity scores are not used. **Condition B**: TF similarity scores are the average of sequence-based and protein-protein interaction-based scores, and gene similarity scores are based on PPIs only. **Condition C**: TF similarity scores are based on PPIs only, and gene similarity scores are not used. **Condition D**: No similarity information used. **Condition E:** Vanilla MTF ($w = 1.0, p = \lambda_r = \lambda_F = \lambda_G = 0$). $r = 1000$ and $s = 100$ for WINTF and vanilla MTF, and $r = 100$ for REMAP, respectively.*

matrix in WINTF. In our previous study with REMAP [25], we performed extensive grid searches on the hyperparameter space, where we found that the optimal parameters are $w = p = \lambda_r = 0.1$, and $\max\_iter = 100$. In our previous study, alterations on these parameters did not significantly affect the performance, unless w is set to 1.0 and p, or λ_r is set to 0.0, respectively (Table 2, Condition E), or $\max\_iter$ is fewer than 50. We found similar trends from grid searches on WINTF hyperparameter space (Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2020.2968442). Small, nonzero w, p, and λ_r values showed reliable performances with $\max\_iter = 100$. The dimension of the feature interaction matrix (e.g., r, s) are important hyperparameters and known to be data-dependent. Our experiments suggest that it is best to keep $r \geq 500$ and $s \leq 150$. Therefore, we set the default hyperparameters as $w = p = \lambda_r = 0.1$, $\max\_iter = 100$, $r = 1000$, and $s = 100$. With these default parameters, we observed that $\lambda_F \sim 0.01$ works best, and λ_G has less impact on performances compared to other hyperparameters.

TABLE 3
Performance Comparison for WINTF and REMAP With Different Hyperparameters

[1]Con.	Algo.	AUC	HLU	MAP	MPR
A	WINTF	.766(6.0e-4)	40.7(.23)	.295(2.6e-3)	.259(4.6e-4)
	REMAP	.726(1.7e-3)	32.0(1.7)	.237(7.4e-3)	.294(1.5e-3)
B	WINTF	.764(5.1e-6)	40.7(.09)	.295(1.6e-3)	.259(1.1e-6)
	REMAP	.726(2.9e-3)	32.2(1.6)	.238(3.7e-3)	.294(2.7e-4)
C	WINTF	.765(7.6e-4)	40.7(.10)	.295(1.5e-3)	.259(6.4e-4)
	REMAP	.730(1.8e-3)	29.8(.30)	.233(1.6e-3)	.292(1.6e-3)
D	WINTF	.762(6.6e-4)	40.7(.09)	.295(1.6e-3)	.259(6.4e-4)
	REMAP	.717(3.2e-3)	30.9(1.7)	.231(4.2e-3)	.300(3.0e-3)

Values are mean and (standard deviation) for 5-fold cross validation
*[1]**Condition A**: Default parameters. **Condition B**: WINTF ranks = (100,100), REMAP rank = 100. **Condition C**: WINTF ranks = (100,50), REMAP rank = 50. **Condition D**: $\lambda_F = \lambda_G = 0$.*

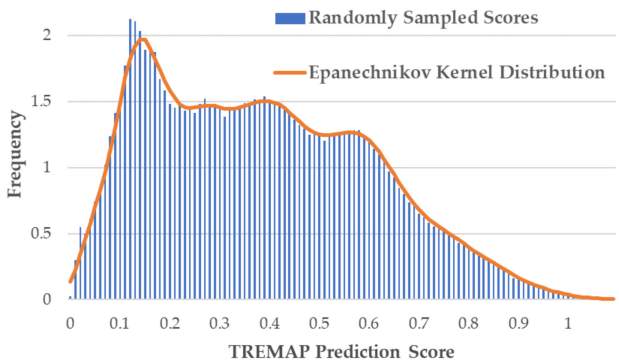


Fig. 3. Probability density of 909,924 randomly sampled WINTF prediction scores for TF-gene associations that are excluded from the training data. Epanechnikov kernel fits to the sampled scores.

Therefore, we set $\lambda_F = 0.01$ and $\lambda_G = 0.7$ as the default hyperparameters. Both of the gene-gene and TF-TF similarity matrices used for our WINTF application are based on protein-protein interactions from STRING, as described in the method section.

The results in Table 2 further demonstrates that the novelty of our method comes from the weight, regularization, imputation, and side information parameters. It is noted that when $w = 1.0, p = \lambda_r = \lambda_F = \lambda_G = 0$, the optimization function Eq. (1) is equivalent to that of the plain matrix tri-factorization, $J = \sum (R - FSG^T)^2$. We observe clear performance improvements when the weight, regularization, imputation, and side information parameters are introduced into the matrix (tri-) factorization methods.

5.3 Evaluation on Independent Test Data

With the choice of parameters and similarity measurements described above, we performed WINTF on the full ChEA TF-gene dataset. We first obtained the predicted score matrix $\hat{R}$ as described in the method section. To assess the statistical significance of the predicted scores, we randomly selected 1,000,000 scores in $\hat{R}$. Removing the scores for 90,076 known associations, we plotted a histogram of the 909,924 predicted scores, which suggested that the predicted scores are not following a simple distribution, such as Gaussian or exponential distribution (Fig. 3). Thus, we first removed the prediction scores for TF-gene associations that were already included in ChEA dataset, and we used Epanechnikov kernel to create a distribution that fits the sampled scores as shown in Fig. 3. Then, we selected the predicted TF-gene pairs that ranked approximately within top 2 percent (i.e., cumulative density is above 0.9808 under the kernel distribution). Our prediction and selection method returned 495 TF-gene associations that were not included in ChEA dataset (Appendix B, available in the online supplemental material). We searched for TRANSFAC [33], ENCODE [36], and TRRUST2 [38] databases to evaluate the final prediction accuracy. As a result, 187 of the 495 (37.8 percent) associations were found in at least one of the three databases. Considering that the chance of correct predictions for random guesses is 9.0 percent based on the density of the ChEA data set, we obtain an enrichment factor of 4.19 (37.8 percent divided by 9.0 percent) for our prediction accuracy.

Among 495 predicted novel associations that are not included in the ChEA training data, a number of them are

TABLE 4
Predicted TF-Gene Associations by WINTF

TF	Gene	[1]CDF	Database	Reference
MYC	NOTCH2	0.99213	ENCODE	[55], [56]
MYC	ZMIZ1	0.99906	ENCODE	[57]
MYC	ARID5B	0.99928	ENCODE	[58]
MYC	BCL6	0.99909	ENCODE, TRANSFAC	[59], [60]
MYC	NDRG1	0.98982	ENCODE, TRRUST2	[62]
MYC	ST3GAL1	0.9992	ENCODE, TRRUST2	[63]
MYC	EFNA5	0.99607	ENCODE	[64]
SPI1	BCL6	0.99864	ENCODE, TRRUST2	[65]
SOX2	HES1	0.99993	None	[66], [67]
SOX2	NOTCH2	0.99335	None	[67]
CREM	MEIS1	0.99955	None	[68]
AR	RUNX1	0.99837	None	[69]

[1]*Cumulative distribution function of the Epanechnikov kernel fitted to the WINTF prediction scores.*

strongly supported by published studies. The associations are listed in Table 4. While the association between NOTCH1 and MYC was previously known from studies regarding T-cell acute lymphoblastic leukemia and included in ChEA dataset, NOTCH2-MYC association was not included. Our prediction method suggests that NOTCH2 may also be association with MYC. It was suggested that NOTCH2 and MYC are related in terms of cellular proliferation in mouse thymic lymphoma without strong evidence to conclude their association [55]. More recently, a study concerned with hypoxia-induced signaling pathway showed that NOTCH2-knockdown murine mesenchymal stem cells cannot properly proliferate, which can be reverted by overexpression of MYC [56]. The collaboration of ZMIZ1 and activated NOTCH1 was found to cause T-cell acute lymphoblastic leukemia in mouse models, which was proposed to be a result of the interaction between ZMIZ1 and MYC at downstream [57]. ARID5B gene, whose role in T-cell acute lymphoblastic leukemia has been previously unknown, was found to directly bind MYC enhancer to promote the expression of MYC, which is a required step for the disease [58]. The concept of double protein lymphoma, characterized by the co-expression of MYC and BCL2 or BCL6, has been known to be aggressive [59], although the MYC/BCL6 biomarker is of less prognostic value [60]. Possibly due to rarity of studies involving MYC/BCL6, the association was not included in the ChEA dataset, while the MYC-BCL2 association was included. NDRG1, whose overexpression in tumor cells decreases the proliferation rate [61], is known to be suppressed by MYC in embryonic cells [62]. In a study concerned with genetic linkage in colon cancer cells, upregulation of glycosyltransferase genes, including ST3GAL1 by MYC was observed [63]. It was reported that EFNA5 was upregulated along with other genes in MYC-knockout mice neural stem and precursor cells [64]. The physical interaction between SPI1 and BCL6 is known. Interestingly, BCL6 acts as a repressor that binds to SPI1 in germinal center B cells [65]. Although the direct association is unknown and thus excluded from ChEA dataset, SOX2 and HES1 (with other genes) have been studied as markers of neural stem cells [66]. A more recent study added evidence that SOX2 and HES1 are at least members of the same regulatory pathway in rat anterior pituitary cells [67]. In the study, it was also found that SOX2-expressing cells have significantly lower levels of NOTCH2 expression, suggesting a potential repression of NOTCH2 by SOX2 [67]. The direct association between CREM and MEIS1 was not known although they are involved in the myogenesis, the growth of skeletal muscle. A recent study suggests that although CREM and MEIS1 may not directly interact, they seem to regulate the growth process through another transcription factor, NF-Y [68]. A recent ChIP-seq experiment showed that RUNX1 is a target of AR, which is important for AR-dependent transcription and cell growth in androgen-dependent prostate cancer [69]. These studies support our claim that WINTF can predict unobserved, but positive associations based on the known associations. NOTCH2-MYC, ZMIZ1-MYC, BCL6-MYC, and EFNA5-MYC associations are in the ENCODE database [36], but not in the TRANSFAC [33] or TRRUST2 [38] database. MYC-NDRG1, MYC-ST3GAL1, and MYC-BCL6 associations are found in both ENCODE and TRRUST2 databases. ARID5B-MYC, SOX2-HES1, and CREM-MEIS1 associations are not found in any of the three databases, suggesting that our method can predict novel TF-gene associations from known ones with proper similarity measurements.

6 Discussions

The results in Table 2 suggest that TF-gene associations may be better modeled by using both protein-protein interaction network and sequence comparison. It is important to note that using only one type of them (e.g., sequence-based similarity only) does not improve the predictive power of WINTF. In addition, the predictive performances were not sensitive to the PPI-based TF-TF network importance weights (Appendix A, available in the online supplemental material). It implies that the canonical PPIs included in the database are insufficient to model gene regulation scenarios where multiple TFs form a complex to regulare a gene. A more comprehensive list of PPIs may address such issue. For better performance as well as interpretability, other types of gene-gene similarity scores may be used. The similarity may be based on the sequence alignment scores of the regulatory elements of the genes, which assumes that the DNA sequences of the regulatory elements have evolved to efficiently recruit the TFs. Differential gene expression data can also be used to measure similarities between genes. The hypothesis in such a case is that two genes showing similar patterns of expression under the same conditions are likely to be regulated by the same TFs. A combination of the two types of similarity scores may improve the predictions. Other biological constraints such as the relative location of TF and gene in the genome can be incorporated into the weight and imputation matrices. However, many of such similarity measurements are tissue- and context-specific. Unfortunately, we do not have large-scale tissue-specific TF-gene association data yet. It is noted the canonical PPIs used in this study are mainly functional associations including the co-expressed genes. Although they are not perfect measurement for the gene-gene similarity, our benchmark studies and independent validations demonstrate their utilization.

Our benchmarks in Table 2 demonstrate that the weight, regularization, imputation, and side information parameters are essential. Without these parameters, the traditional matrix tri-factorization method shows no predictive power.

Table 3 also suggest that the improved performance of WINTF compared to REMAP is from the existence of the matrix S. While the main purpose of introducing matrix S is to set different ranks for genes and TFs, it is not clear whether the ranks must be very different. The Condition B in Table 3 shows that WINTF performs better than REMAP even if all rank parameters are set to 100. In practice, the rank parameters are heuristically optimized. On the other hand, the matrix S can be viewed as a hidden layer introduced to REMAP. Thus, the matrix S may have worked similarly to the hidden layers for the popular deep learning methods, characterized by multiple layers of neural networks with activation functions and regularization steps. Increasing the number of low-rank matrices in WINTF to mimic deep learning may be an interesting future study. A more interesting combination is to integrate neural network techniques with matrix factorization, as shown in a recent study where the matrix inner product is considered an additional layer to a multilayer neural network [70]. The time complexity due to the introduction of an additional low-rank matrix as well as large number of parameters from multilayer neural network can be overcome by factorizing smaller submatrices and projecting to the original feature space [71]. In addition, the algorithm to optimize the cost function of matrix factorization may be improved. Simultaneous perturbation stochastic approximation is a potential algorithm to improve the performance as well as the speed of optimization since it requires a dramatically low number of evaluations per iteration and randomness to potentially find the global minimum solution [72], [73]. Such work will enable larger scale applications of the association prediction method with improved accuracy and interpretability.

7 Conclusions

In this study, we develop a tri-factorization-based collaborative filtering algorithm, WINTF, that allows users to set different low-ranks for users and items. Compared with its single-rank analog, WINTF showed better performances measured by four different metrics. We apply WINTF to predict unobserved TF-gene associations using a collection of known associations. Many of the predicted associations by WINTF are supported by evidence in the literature or listed in existing databases. Therefore, WINTF is a powerful tool for TF-gene association prediction, and it can be directly applied to tissue-specific tasks to yield further refined predictions.

Acknowledgments

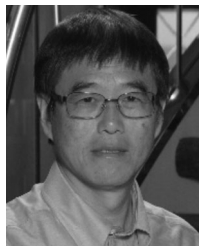
This work was supported by Grant Number R01LM011986 from the National Library of Medicine (NLM) of the National Institute of Health (NIH), and Grant Number R01GM122845 from the National Institute of General Medical Sciences (NIGMS) of the National Institute of Health (NIH).

References

[1] M. Levine and R. Tjian, "Transcription regulation and animal diversity," *Nature*, vol. 424, no. 6945, pp. 147–51, Jul. 10, 2003.
[2] D. Baltimore, "Our genome unveiled," *Nature*, vol. 409, no. 6822, pp. 814–816, Feb. 15, 2001.
[3] E. S. Lander *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, Feb. 15, 2001.
[4] T. Phillips, and L. Hoopes, "Transcription factors and transcriptional control in eukaryotic cells," *Nature Educ.*, vol. 1, no. 1, 2008, Art. no. 119.
[5] P. Collas, "The current state of chromatin immunoprecipitation," *Mol. Biotechnology*, vol. 45, no. 1, pp. 87–100, May, 2010.
[6] V. R. Iyer *et al.*, "Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF," *Nature*, vol. 409, no. 6819, pp. 533–538, 2001.
[7] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-DNA interactions," *Science*, vol. 316, no. 5830, pp. 1497–502, Jun. 8, 2007.
[8] C. L. Wei *et al.*, "A global map of p53 transcription-factor binding sites in the human genome," *Cell*, vol. 124, no. 1, pp. 207–19, Jan. 13, 2006.
[9] P. V. Kharchenko, M. Y. Tolstorukov, and P. J. Park, "Design and analysis of ChIP-seq experiments for DNA-binding proteins," *Nature Biotechnol.*, vol. 26, no. 12, pp. 1351–9, Dec. 2008.
[10] D. A. Nix, S. J. Courdy, and K. M. Boucher, "Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks," *BMC Bioinformatics*, vol. 9, Dec. 5, 2008, Art. no. 523.
[11] G. Tuteja, P. White, J. Schug, and K. H. Kaestner, "Extracting transcription factor targets from ChIP-Seq data," *Nucleic Acids Res.*, vol. 37, no. 17, pp. e113, Sep. 2009.
[12] M. J. Vogel, D. Peric-Hupkes, and B. van Steensel, "Detection of in vivo protein-DNA interactions using DamID in mammalian cells," *Nat. Protoc.*, vol. 2, no. 6, pp. 1467–78, 2007.
[13] A. Lachmann *et al.*, "ChEA: Transcription factor regulation inferred from integrating genome-wide ChIP-X experiments," *Bioinformatics*, vol. 26, no. 19, pp. 2438–44, Oct. 1, 2010.
[14] T. S. Furey, "ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions," *Nature Rev. Genetics*, vol. 13, no. 12, pp. 840–52, Dec, 2012.
[15] C. D. McClure and T. D. Southall, "Getting down to specifics: Profiling gene expression and protein-DNA interactions in a cell type-specific manner," *Adv. Genetics*, vol. 91, pp. 103–51, 2015.
[16] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Adv. Artif. Intell.*, vol. 2009, 2009, Art. no. 4.
[17] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet Comput.*, vol. 7, no. 1, pp. 76–80, Jan./Feb. 2003.
[18] T. Hofmann, "Latent semantic models for collaborative filtering," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 89–115, 2004.
[19] P. Paatero and U. Tapper, "Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.
[20] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–91, Oct. 21, 1999.
[21] J. Lee, M. Sun, and G. Lebanon, "A comparative study of collaborative filtering algorithms," 2012, *arXiv:1205.3193*.
[22] J. Bennett and S. Lanning, "The netflix prize," in *Proc. KDD Cup Workshop*, 2007, Art. no. 35.
[23] Y. Yao *et al.*, "Dual-regularized one-class collaborative filtering," in *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manage.*, 2014, pp. 759–768.
[24] C. Chen, H. Tong, L. Xie, L. Ying, and Q. He, "FASCINATE: Fast cross-layer dependency inference on multi-layered networks," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 765–774.
[25] H. Lim *et al.*, "Large-scale off-target identification using fast and accurate dual regularized one-class collaborative filtering and its application to drug repurposing," *PLoS Comput. Biol.*, vol. 12, no. 10, Oct. 2016, Art. no. e1005135.
[26] C. Cheng, R. Min, and M. Gerstein, "TIP: A probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles," *Bioinformatics*, vol. 27, no. 23, pp. 3221–3227, 2011.
[27] C.-C. Yang *et al.*, "iTAR: A web server for identifying target genes of transcription factors using ChIP-seq or ChIP-chip data," *BMC Genomics*, vol. 17, 2016, Art. no. 632.
[28] H. Redestig, D. Weicht, J. Selbig, and M. A. Hannah, "Transcription factor target prediction using multiple short expression time series from Arabidopsis thaliana," *BMC Bioinformatics*, vol. 8, Nov 18, 2007, Art. no. 454.

[29] S. M. Kielbasa, N. Bluthgen, M. Fahling, and R. Mrowka, "Targetfinder.org: A resource for systematic discovery of transcription factor target genes," *Nucleic Acids Res.*, vol. 38, no. Web Server issue, pp. W233–W238, Jul. 2010.
[30] C. J. Banks, A. Joshi, and T. Michoel, "Functional transcription factor target discovery via compendia of binding and expression profiles," *Sci. Rep.*, vol. 6, Feb. 9, 2016, Art. no. 20649.
[31] E. Wingender, "Compilation of transcription regulating proteins," *Nucleic Acids Res.*, vol. 16, no. 5 Pt B, 1988, Art. no. 1879.
[32] V. Matys *et al.*, "TRANSFAC : Transcriptional regulation, from patterns to profiles," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 374–378, 2003.
[33] V. Matys *et al.*, "TRANSFAC and its module TRANSCompel: Transcriptional gene regulation in eukaryotes," *Nucleic Acids Res.*, vol. 34, no. suppl_1, pp. D108–D110, 2006.
[34] A. Sandelin, W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard, "JASPAR: An open-access database for eukaryotic transcription factor binding profiles," *Nucleic Acids Res.*, vol. 32, no. suppl_1, pp. D91–D94, 2004.
[35] A. Mathelier *et al.*, "JASPAR 2016: A major expansion and update of the open-access database of transcription factor binding profiles," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D110–D115, 2016.
[36] E. P. Consortium, "The ENCODE (Encyclopedia of DNA elements) project," *Science*, vol. 306, no. 5696, pp. 636–640, Oct. 22, 2004.
[37] H. Han *et al.*, "TRRUST: a reference database of human transcriptional regulatory interactions," *Sci. Rep.*, vol. 5, Jun. 12, 2015, Art. no. 11432.
[38] H. Han *et al.*, "TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions," *Nucleic Acids Res.*, vol. 46, pp. D380–D386, Oct. 26, 2017.
[39] A. D. Rouillard *et al.*, "The harmonizome: A collection of processed datasets gathered to serve and mine knowledge about genes and proteins," *Database*, vol. 2016, no. 2016, pp. baw100–baw100, 2016.
[40] L. A. Bovolenta, M. L. Acencio, and N. Lemke, "HTRIdb: An open-access database for experimentally verified human transcriptional regulation interactions," *BMC Genomics*, vol. 13, Aug, 17, 2012, Art. no. 405.
[41] H. Chen and J. Li, "A flexible and robust multi-source learning algorithm for drug repositioning," in *Proc. 8th ACM Int. Conf. Bioinf. Comput. Biol. Health Inf.*, 2017, pp. 510–515.
[42] T. Hwang *et al.*, "Co-clustering phenome-genome for phenotype classification and disease gene discovery," *Nucleic Acids Res.*, vol. 40, no. 19, Oct. 2012, Art. no. e146.
[43] S. Park *et al.*, "An integrative somatic mutation analysis to identify pathways linked with survival outcomes across 19 cancer types," *Bioinformatics*, vol. 32, no. 11, pp. 1643–1651, Jun. 1, 2016.
[44] M. Zitnik and B. Zupan, "Matrix factorization-based data fusion for gene function prediction in baker's yeast and slime mold," in *Proc. Pacfic Symp. Biocomput.*, 2014, pp. 400–411.
[45] A. Copar, M. Zitnik, and B. Zupan, "Scalable non-negative matrix tri-factorization," *BioData Min*, vol. 10, 2017, Art. no. 41.
[46] Y. Yao *et al.*, "Dual-regularized one-class collaborative filtering," in *Proc. 23rd ACM Int. Conf. Inf. Knowl. Manage.*, 2014, pp. 759–768.
[47] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2000, pp. 556–562.
[48] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 126–135.
[49] D. Szklarczyk *et al.*, "STRING v10: Protein-protein interaction networks, integrated over the tree of life," *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D447–452, Jan. 2015.
[50] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–10, Oct. 5, 1990.
[51] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," in *Proc. 14th Conf. Uncertainty Artif. Intell.*, 1998, pp. 43–52.
[52] Y. Li, J. Hu, C. Zhai, and Y. Chen, "Improving one-class collaborative filtering by incorporating rich user information," *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage.*, pp. 959–968, 2010.
[53] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *Proc. 8th IEEE Int. Conf. Data Mining*, 2008, pp. 263–272.
[54] A. Wang, H. Lim, S. Y. Cheng, and L. Xie, "ANTENNA, a multi-rank, multi-layered recommender system for inferring reliable drug-gene-disease associations: Repurposing diazoxide as a targeted anti-cancer therapy," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 6, pp. 1960–1967, Nov./Dec. 2018.
[55] P. Lopez-Nieva, J. Santos, and J. Fernandez-Piqueras, "Defective expression of Notch1 and Notch2 in connection to alterations of c-Myc and Ikaros in gamma-radiation-induced mouse thymic lymphomas," *Carcinogenesis*, vol. 25, no. 7, pp. 1299–304, Jul. 2004.
[56] Y. Sato *et al.*, "Notch2 signaling regulates the proliferation of murine bone marrow-derived mesenchymal stem/stromal cells via c-Myc expression," *PLoS One*, vol. 11, no. 11, 2016, Art. no. e0165946.
[57] L. A. Rakowski *et al.*, "Convergence of the ZMIZ1 and NOTCH1 pathways at C-MYC in acute T lymphoblastic leukemias," *Cancer Res.*, vol. 73, no. 2, pp. 930–941, Jan. 15, 2013.
[58] W. Z. Leong *et al.*, "ARID5B activates the TAL1-induced core regulatory circuit and the MYC oncogene in T-cell acute lymphoblastic leukemia," in *Proc. Amer. Soc. Hematol.*, 2017, pp. 2343–2360.
[59] R. K. Pillai, M. Sathanoori, S. B. Van Oss, and S. H. Swerdlow, "Double-hit B-cell lymphomas with BCL6 and MYC translocations are aggressive, frequently extranodal lymphomas distinct from BCL2 double-hit B-cell lymphomas," *Amer. J. Surgical Pathol.*, vol. 37, no. 3, pp. 323–32, Mar. 2013.
[60] Q. Ye *et al.*, "Prognostic impact of concurrent MYC and BCL6 rearrangements and expression in de novo diffuse large B-cell lymphoma," *Oncotarget*, vol. 7, no. 3, pp. 2401–2416, Jan. 19, 2016.
[61] R. J. Guan *et al.*, "Drg-1 as a differentiation-related, putative metastatic suppressor gene in human colon cancer," *Cancer Res.*, vol. 60, no. 3, pp. 749–755, 2000.
[62] X. Qu *et al.*, "Characterization and expression of three novel differentiation-related genes belong to the human NDRG gene family," *Mol. Cell Biochem.*, vol. 229, no. 1-2, pp. 35–44, Jan. 2002.
[63] K. Sakuma, M. Aoki, and R. Kannagi, "Transcription factors c-Myc and CDX2 mediate E-selectin ligand expression in colon cancer cells undergoing EGF/bFGF-induced epithelial-mesenchymal transition," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 20, pp. 7776–7781, May 15, 2012.
[64] V. Martinez-Cerdeno *et al.*, "N-Myc and GCN5 regulate significantly overlapping transcriptional programs in neural stem cells," *PLoS One*, vol. 7, no. 6, 2012, Art. no. e39456.
[65] F. Wei, K. Zaprazna, J. Wang, and M. L. Atchison, "PU.1 can recruit BCL6 to DNA to repress gene expression in germinal center B cells," *Mol. Cell Biol.*, vol. 29, no. 17, pp. 4612–4622, Sep. 2009.
[66] H. Takanaga *et al.*, "Gli2 is a novel regulator of sox2 expression in telencephalic neuroepithelial cells," *Stem Cells*, vol. 27, no. 1, pp. 165–74, Jan. 2009.
[67] K. Batchuluun, M. Azuma, K. Fujiwara, T. Yashiro, and M. Kikuchi, "Notch signaling and maintenance of SOX2 expression in rat anterior pituitary cells," *Acta Histochem Cytochem*, vol. 50, no. 2, pp. 63–69, Apr. 27, 2017.
[68] C. V. C. Grade *et al.*, "CREB, NF-Y and MEIS1 conserved binding sites are essential to balance Myostatin promoter/enhancer activity during early myogenesis," *Mol. Biol. Rep.*, vol. 44, no. 5, pp. 419–427, Oct. 2017.
[69] K. Takayama *et al.*, "RUNX1, an androgen- and EZH2-regulated gene, has differential roles in AR-dependent and -independent prostate cancer," *Oncotarget*, vol. 6, no. 4, pp. 2263–2276, Feb. 10, 2015.
[70] X. He *et al.*, "Neural collaborative filtering," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 173–182.
[71] L. W. Mackey, A. Talwalkar, and M. I. Jordan, "Distributed matrix completion and robust factorization," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 913–960, 2015.
[72] J. C. Spall, "An overview of the simultaneous perturbation method for efficient optimization," *Airport Model. Simul.*, vol. 19, no. 4, pp. 141–154, 1999.
[73] J. L. Maryak and D. C. Chin, Efficient global optimization using SPSA, in *Proc. Amer. Control Conf.*, (Cat. No. 99CH36251), vol. 2, pp. 890–894, 1999.

Hansaim Lim received the BA degree in chemistry with bioinformatics concentration from the Hunter College of the City University of New York, in 2014, the MPhil degree in biochemistry from the CUNY Graduate Center, in 2019. He is currently working toward the PhD degree in biochemistry in the Graduate Center of the City University of New York, since 2015. He joined Dr. Lei Xie's lab at Hunter College for his dissertation project. His research interests cover machine learning and deep learning-based drug discovery.



Lei Xie received the BS degree in polymer physics from the University of Science and Technology of China, P. R. China., in 1990, the MSc degree in computer science, and the PhD degree in chemistry from Rutgers University, USA, in 2000. He was an associate scientist at Columbia University and Howard Hughes Medical Institute, USA. He has worked in pharmaceutical and biotechnology companies Roche and Eidogen, USA for several years. He was a principal scientist at San Diego Supercomputer Center from 2006 to 2011. He is currently a professor with the Department of Computer Science, Hunter College, and The Graduate Center, The City University of New York, and adjuct professor of neuroscience with Weill Cornell Medical College, Cornell University, USA. His research interests include data mining, machine learning, biophysics, systems biology, drug discovery, and precision medicine with more than 70 technical publications.