Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Matrix (Factorization) Reloaded: Flexible Methods for Imputing Genetic Interactions with Cross-Species and Side Information

Jason Fan [1,*], Xuan Cindy Li [2], Mark Crovella[3] and Mark D.M. Leiserson [1,*]

[1]Department of Computer Science and Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA, [2]Program in Computational Biology, Bioinformatics, and Genomics, University of Maryland, College Park, MD 20742, USA, [3]Department of Computer Science, Boston University, MA 02215, USA.

*To whom correspondence should be addressed.

Abstract

Motivation: Mapping genetic interactions (GIs) can reveal important insights into cellular function, and has potential translational applications. There has been great progress in developing high-throughput experimental systems for measuring GIs (e.g. with double knockouts) as well as in defining computational methods for inferring (imputing) unknown interactions. However, existing computational methods for imputation have largely been developed for and applied in baker's yeast, even as experimental systems have begun to allow measurements in other contexts. Importantly, existing methods face a number of limitations in requiring specific side information and with respect to computational cost. Further, few have addressed how GIs can be imputed when data is scarce.

Results: In this paper we address these limitations by presenting a new imputation framework, called *Extensible Matrix Factorization* (EMF). EMF is a framework of composable models that flexibly exploit cross-species information in the form of GI data across multiple species, and arbitrary side information in the form of kernels (e.g. from protein-protein interaction networks). We perform a rigorous set of experiments on these models in matched GI datasets from baker's and fission yeast. These include the first such experiments on genome-scale GI datasets in multiple species in the same study. We find that EMF models that exploit side and cross-species information improve imputation, especially in data-scarce settings. Further, we show that EMF outperforms the state-of-the-art deep learning method, even when using strictly less data, and incurs orders of magnitude less computational cost.

Availability: Implementations of models and experiments are available at: `github.com/lrgr/emf`

Contact: mdml@umd.edu

1 Introduction

A genetic interaction (GI) is a measure of how a *combination* of gene variants produces a phenotype that is different than expected, given the phenotypes of each *independent* gene variant. Most commonly, a GI is measured for a pair of gene knockouts with a measure of cell viability as the phenotype. Although a single GI provides only limited phenotypic information, mapping a set of GIs in a model organism is thought to be able to resolve fundamental biological questions such as the minimum number of genes required for a viable cell (Kuzmin *et al.*, 2018; Hutchison *et al.*, 2016). Furthermore, knowledge of GIs has enabled promising new strategies for cancer treatment (Ashworth and Lord, 2018; Lee *et al.*, 2018), and may expand opportunities for treating infectious diseases (Patrick *et al.*, 2018).

Consequently, identifying and characterizing GIs has been a major focus in systems biology for the past two decades, spurring innovations in experimental systems and computational methods. Recently, researchers have sought to go beyond measuring interactions for small sets of specific genes or gene pairs, to develop approaches for generating what are referred to as "unbiased" maps of pairwise *quantitative* GIs between large sets of genes (Costanzo *et al.*, 2019). A quantitative GI for a tested pair of genes is a real-valued score for the direction (positive / alleviating versus negative / aggravating) and strength of the interaction. For example, treated quantitatively, a synthetic lethal interaction is a GI with a score much less than zero. *In vitro* efforts began in baker's yeast with small maps for all pairs of genes involved in key biological functions (Schuldiner *et al.*, 2005; Collins *et al.*, 2007; Roguev *et al.*, 2008). These efforts culminated in a landmark study (Costanzo *et al.*, 2016) that published a map of GIs for over 90% of all genes in baker's yeast (*S. cerevisiae*).

Despite impressive progress, many challenges remain. These challenges include measuring interactions in species other than baker's yeast, examining higher-order interactions for sets of more than two genes, and measuring GIs for different phenotypes. In fact, in each of these cases, there have been recent experimental studies (Ryan *et al.* (2012); Kuzmin *et al.* (2018); Dixit *et al.* (2016), respectively). However, the landscape of yet unmeasured GIs remains vast and will not be fully explored through *in vitro* experimentation alone. Thus, there is an enormous need for *in silico* methods to complement the recent and ongoing experimental advances.

To address this need, methods have been developed along a number of dimensions. First, it is important to note the critical difference between the classification problem posed by *binary* classes of extreme GIs, and the regression problem associated with the larger information content contained in *quantitative* GI data. With respect to binary GIs, much work has been focused on the prediction of synthetic lethal interactions (Wong *et al.*, 2004; Paladugu *et al.*, 2008; Pandey *et al.*, 2010; Wu *et al.*, 2014; Jacunski *et al.*, 2015; Benstead-Hume *et al.*, 2019), sometimes treating the classification problem as standard link prediction (Liben-Nowell and Kleinberg, 2003; Lü and Zhou, 2011). However, genome-scale work in yeast has gone beyond identifying the most extreme interactions to identifying correlations between genes' profiles of quantitative GI scores regardless of magnitude, in order to create genome-wide maps of gene function (Costanzo *et al.*, 2010). Hence, in this paper, we study imputing real valued, quantitative GI scores for all gene-pairs.

Ulitsky *et al.* (2009) were the first to develop methods for the *regression* problem of predicting quantitative GIs using features derived from functional annotations, protein-protein interactions, and the Gene Ontology (GO) (The Gene Ontology Consortium, 2018). More recently, Ma *et al.* (2018) – building off of the work of Yu *et al.* (2016) – introduced an interpretable deep learning method that uses GO to achieve state-of-the-art performance in predicting GI scores in baker's yeast.

In this work, we present a new computational framework for the quantitative GI regression problem, termed *Extensible Matrix Factorization* (EMF), and show its utility in both baker's yeast *and* fission yeast (*S. pombe*). In developing EMF, we seek to overcome a number of limitations of existing methods.

First, we note that existing state-of-the-art methods are predicated on the availability of *specific* kinds of side information as a necessary input for feature generation. That is, computational methods such as DCell (Ma *et al.*, 2018) require annotations from GO. However, the availability and quality of GO annotations vary widely across species. For example, baker's yeast has more than double the number of annotations in fission yeast (The Gene Ontology Consortium, 2018). Thus, the reliance on specific *side information* as input to DCell limits its ability to be used for a wide range of species. Furthermore, no methods have exploited known correlations across GI data (Roguev *et al.*, 2008; Ryan *et al.*, 2012; Koch *et al.*, 2012) in related species (*cross-species information*) to make predictions in species in which data is scarce.

Second, training state-of-the-art methods is computationally intensive. For example, DCell took two to three days to train on data from Costanzo *et al.* (2016). Methods that require significant time to train can impede efforts to develop and benchmark new models; this bottleneck may grow further as the sizes of GI datasets increase.

To address these limitations, EMF is designed to be a more broadly useful approach that can *flexibly incorporate* various kinds of side information, as available. The EMF framework consists of a collection of composable matrix factorization (MF) models that can optionally exploit known non-uniformities in GIs, within-species side information (via kernelization), and cross-species information (via gene-gene similarities). A core contribution within EMF is *cross-species matrix factorization* (XSMF), a new method for using information from one (*source*) species to improve GI imputation in a second (*target*) species.

We also designed EMF to have low computational cost – EMF models typically takes less than one minute to train on genome-scale data. As evidence of the scalability *and* flexibility of EMF, we use it to impute GIs in baker's *and* fission yeast at genome-scale. To the best of our knowledge, ours is the first study to do so in fission yeast.

Further, we note that recent evidence from data mining literature shows that MF can be competitive with deep learning for some problems when attention is paid to details such as hyperparameter tuning (Rendle *et al.*, 2019). In light of this, we also present in this study a principled approach to composing EMF models and a rigorous approach to hyperparameter optimization to properly weight model combinations.

In the remainder of this work, we show that (when properly applied) MF compares favorably to previous methods. Our contributions include:

1. **Extensible Matrix Factorization (EMF): a framework of composable matrix factorization models for imputing genetic interactions**. EMF extends and unifies several existing MF methods that have not previously been applied to GIs. EMF consists of: a cross-species model that regularizes learned factors across species based on a gene-gene similarity measure; a kernelized model that regularizes learned factors within species; and, a bias model that learns the mean GI score per gene and (motivated by Koch *et al.* (2012)) regularizes biases across species.
2. **Rigorous benchmarking of EMF on matched datasets from baker's and fission yeast**. We compare EMF models on matched GI datasets for chromosome biology genes from baker's and fission yeast using automated approaches for hyperparameter selection (Bergstra *et al.*, 2011), and show that each component of the EMF framework captures additional and complementary signal in data-scarce settings. We also compare directly to the one earlier MF method for imputing GIs, and find EMF to be superior in performance.
3. **Application of EMF to genome-scale datasets.** We apply the best performing models from our benchmarking experiment on datasets covering 75% and 60% of all non-essential genes in baker's and fission yeast, respectively (Costanzo *et al.*, 2010; Ryan *et al.*, 2012). Compared to the state-of-the-art as reported in literature, EMF models show superior performance, and train in minutes instead of days.

2 Methods

Matrix factorization (MF) (Salakhutdinov and Mnih, 2008; Koren *et al.*, 2009), also referred to as *matrix completion* (Candès and Recht, 2009), is a strategy for imputing missing values in a matrix. The matrix is generally assumed to contain redundancies and potentially other regularities or correlations. In other words, a subset of visible values suffices to approximately infer some or all missing values.

MF has proven to be a broadly effective technique in a wide range of problem areas (Koren *et al.*, 2009; Lee and Seung, 1999; Stein-O'Brien *et al.*, 2018). Furthermore, it can have a number of advantages over more recently developed methods such as deep learning (Rendle *et al.*, 2019). A goal of this study is to demonstrate how MF can be an advantageous strategy for the problem of genetic interaction (GI) prediction.

MF takes as input a $n \times m$ matrix X which is *partially observed*. We use Ω to denote the set of indices in X whose values are known. The goal of MF is to impute missing values in X (i.e., $(i,j) \notin \Omega$). The basic MF framework starts from the assumption that X, were it fully-known, would be *effectively low-rank*. That is, X can be well approximated by a matrix R of rank $k \ll \min(m,n)$. MF methods seek to estimate R and use the values of R to estimate the missing values of X.

This suggests the following optimization problem:

$$U, V = \underset{U^*, V^*}{\operatorname{argmin}} \sum_{(i,j)\in\Omega} (x_{ij} - u_i^{T*} v_j^*)^2 \tag{1}$$

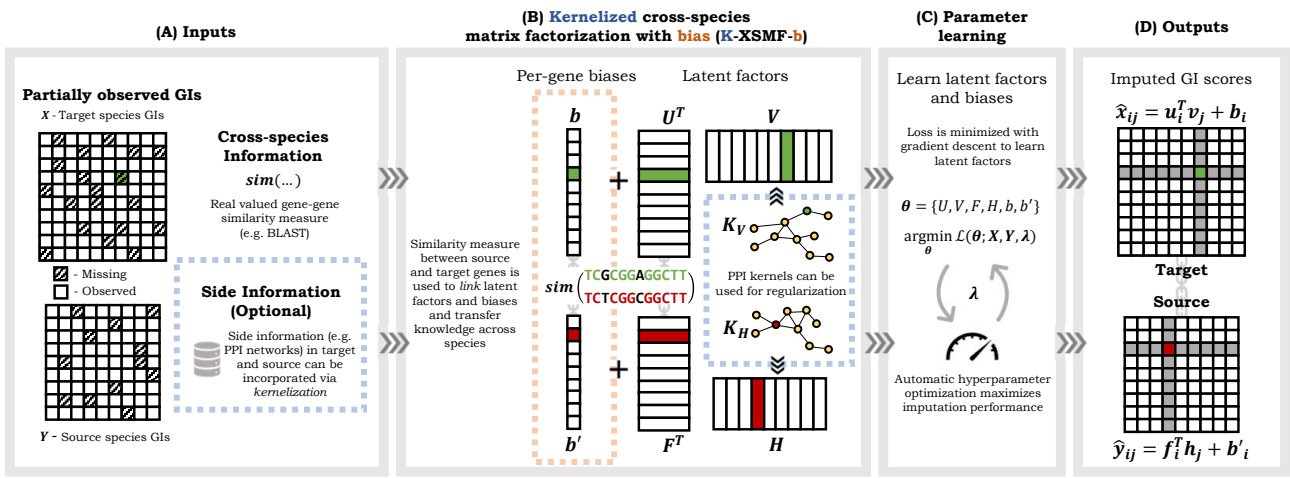

Fig. 1. (A) Extensible Matrix Factorization (EMF) is a composable framework of matrix factorization models that takes as input partially observed matrices in a target species and, optionally, a source species. Here we use K-XSMF-b, one realization of EMF, to illustrate the range of EMF models as they apply to imputing genetic interactions (GIs). The XSMF model uses cross-species information in the form of a gene-gene similarity measure (e.g. BLAST). Side information (blue box, which can include PPI networks, GO annotations, etc.) forms additional optional input, yielding K-XSMF. (B) To model per-gene biases in mean GI score, bias terms (orange box) can be introduced in the target and source species. Similarities from the provided cross-species similarity measure are used to link both biases and latent factors across species, resulting in K-XSMF-b. (C) Latent factors and biases are learned using gradient descent. Importantly, to ensure best possible test-time performance, `hyperopt` is used to automatically select and optimize hyperparameters (Bergstra et al., 2011). (D) After hyperparameters are selected and latent factors and biases are learned, missing interactions can be imputed.

in which U and V are matrices with k rows, where k is a hyperparameter chosen to model the effective rank of X. This framework allows one to recover $R = U^T V$, and admits an interpretation of corresponding columns u_i and v_j as *latent factors* representing the entities on the i-th row and j-th column of X respectively. To impute missing value $(i, j) \notin \Omega$, one simply computes the inner-product, $u_i^T v_j$ of the learned latent factors.

Regularization to reduce overfitting can be achieved by including additional terms, such as the ℓ_2-regularizer from Koren *et al.* (2009) and Salakhutdinov and Mnih (2008):

$$U, V = \underset{U^*, V^*}{\operatorname{argmin}} \sum_{(i,j)\in\Omega} (x_{ij} - u_i^{T*} v_j^*)^2 + \lambda \left(\|U^*\|_F^2 + \|V^*\|_F^2 \right). \quad (2)$$

The basic MF framework succeeds by exploiting the inherent low effective rank of the data. Moreover, an important advantage of MF is the straightforward and principled ways in which it can be adapted to incorporate additional regularities in the data. For example, "side" information (additional data features) may be predictive in a manner that is synergistic with the basic low-rank assumption.

2.1 Extensible Matrix Factorization (EMF): a composable *class* of matrix factorization models

We present a set of composable components for MF that exploit *cross-species* and *side* information. We derive these from biological observations and ultimately incorporate these into a unified *Extensible Matrix Factorization* (EMF) framework.

EMF encompasses several existing MF models, including the basic MF model given in (2) (Salakhutdinov and Mnih, 2008), MF with bias (MF-b) (Koren *et al.*, 2009), and kernelized probabilistic MF (K-PMF) (Zhou *et al.*, 2012). Our contribution with EMF is in presenting a unified view of these models, and expanding their formulations to cross-species settings.

We describe the framework generally as it applies to matrices of biological data where the rows and columns are indexed by genes. We begin by introducing two novel components for exploiting cross-species information, and then present a component for exploiting side information (i.e. within a species). While we apply the EMF components in both the single-species and cross-species settings, we describe all components as they apply to a cross-species setting. Where emphasis is useful, we describe how the models can be leveraged specifically for imputing genetic interactions (GIs).

2.1.1 Cross-species matrix factorization (XSMF)

The first extension to MF that we propose is a *cross-species matrix factorization (XSMF) component*, a novel MF scheme that jointly factorizes matrices in a *target* and a *source* species to better impute missing values in the *target*.

Let $X \in \mathcal{R}^{n\times m}$ be a partially observed matrix for a target species, and $Y \in \mathcal{R}^{n'\times m'}$ be a partially observed matrix for a source species. We use Ω_X and Ω_Y to denote the indices in X and Y whose values are known. We present, piecewise, the optimization objective that defines XSMF.

First, the primary objective of XSMF is to estimate latent factors $U \in \mathcal{R}^{k\times n}$ and $V \in \mathcal{R}^{k\times m}$ that best reconstruct observed values in the target species. To do so, XSMF minimizes the objective:

$$\mathcal{L}_t = \sum_{(i,j)\in\Omega_X} (x_{ij} - \hat{x}_{ij})^2, \quad (3)$$

where $\hat{x}_{ij} = u_i^T v_j$.

Second, XSMF simultaneously estimates latent factors $F \in \mathcal{R}^{k\times n'}$ and $H \in \mathcal{R}^{k\times m'}$ that reconstruct observed values in the source species. To do so, XSMF also minimizes the objective:

$$\mathcal{L}_s = \sum_{(i,j)\in\Omega_Y} (y_{ij} - \hat{y}_{ij})^2, \quad (4)$$

where $\hat{y}_{ij} = f_i^T h_j$.

Third, given a *similarity* measure between target species genes and source species genes, $sim(\cdot,\cdot)$, with corresponding similarity score matrix, S, XSMF *links* the factorizations sought by (3) and (4).

It is important to observe here that matrices belonging to target and source species cannot be naively merged because there is no complete one-to-one correspondence between genes in the rows and columns of the target and source. It is also not useful to naively merge matrices by adding source species values, via new rows and columns for source genes, to the

target matrix. In such a merged matrix, the latent factors between source and target genes would be independent and not interact.

Thus, XSMF seeks to maximize *weighted inner products* between the latent factors of genes by minimizing the objective:

$$\mathcal{L}_{\mathbf{x}} = -\sum_i \sum_j sim(i,j) \cdot u_i^T f_j \tag{5}$$

or equivalently,

$$\mathcal{L}_{\mathbf{x}} = -\text{tr}(USF^T). \tag{6}$$

Finally, ℓ_2 regularization is also added to reduce overfitting and XSMF also minimizes the regularizer:

$$\mathcal{L}_r = \|U\|_F^2 + \|V\|_F^2 + \|F\|_F^2 + \|H\|_F^2 \tag{7}$$

The full objective function that XSMF minimizes, with respect to latent factors, can then be written as:

$$\mathcal{L}_{\text{XSMF}} = \mathcal{L}_t + \lambda_s \mathcal{L}_s + \lambda_{\mathbf{x}} \mathcal{L}_{\mathbf{x}} + \lambda_r \mathcal{L}_r \tag{8}$$

with the introduction of user-defined hyperparameters λ_s, $\lambda_{\mathbf{x}}$, and λ_r. In the XSMF model, the parameters λ_s and $\lambda_{\mathbf{x}}$ have useful interpretations. The hyperparameter λ_s controls the tradeoff between reconstructing the target and source species values, and $\lambda_{\mathbf{x}}$ controls the degree to which latent factors of similar genes across species ought to be close in representation.

We highlight that $sim(\cdot,\cdot)$ can be *any* reasonable similarity measure of homology. For example, similarity measures like BLAST bitscores (Altschul *et al.*, 1990), string kernels for protein and DNA sequences (Leslie *et al.*, 2001), or similarity scores based on biological networks (Fan *et al.*, 2019), that have proven to be informative in other contexts can be utilized with little to no modification. Unlike researcher-provided labels in GO, many such similarity measures can be computed with only minimal researcher supervision.

2.1.2 Modeling per-gene biases in average values

It has been observed in other settings that MF models that explicitly account for per-column and per-row 'biases' have been shown to outperform MF models that do not (Koren *et al.*, 2009). In fact, for GI data, the average GI score (i.e., the propensity for a given gene to genetically interact with any other gene) is known to be non-uniform across yeast genomes (Costanzo *et al.*, 2010, 2016; Ryan *et al.*, 2012).

Thus, all models in the EMF framework can be extended to account for per-gene biases. For cross-species models, per-gene latent bias terms can be introduced and an imputed value in the target species between genes (i,j) can instead be modified to:

$$\hat{x}_{ij} = u_i^T v_j + b_i. \tag{9}$$

In the source species an imputed value can be modified to:

$$\hat{y}_{ij} = f_i^T h_j + b'_i. \tag{10}$$

Naturally, ℓ_2 regularization over corresponding vectors of biases, b and b', can be added to the final optimization objective to reduce overfitting.

2.1.3 Modeling the conservation of biases across species

EMF can also 'link' biases if one expects biases to be correlated across species. In fact, this is the case for GIs. Koch *et al.* (2012) showed that the total number of extreme, synthetic lethal interactions can be correlated between similar genes in baker's and fission yeast. This observation motivates an additional way to exploit cross-species similarities in the EMF framework. The following regularization term that links biases in the source and target can be added to cross-species models:

$$\mathcal{L}_b = -b^T S b'. \tag{11}$$

Adding the regularization term $\mathcal{L}_b$ to the objective of an EMF model encourages biases of similar genes to also be similar.

2.1.4 Incorporating arbitrary side information

Recent work in MF has introduced a number of additional ways to incorporate side information – such as networks (Zitnik and Zupan, 2015) or kernels (Zhou *et al.*, 2012) – to further improve model performance.

We adapt the kernelized approach taken by Zhou *et al.* (2012) to extend both single-species and cross-species models in EMF. To exploit side information in the target species, kernels that regularize latent factors U and V are introduced. Kernelization enables incorporation of any arbitrary side information about known similarities between (same-species) genes, as long as appropriate kernels K_U and K_V can be computed for genes in the target species. Concretely, the following quadratic terms can either be added in addition to, or replace the usual ℓ_2 regularizers on U and V:

$$\mathcal{L}_{kt} = \text{tr}\left(UK_U^{-1}U^T\right) + \text{tr}\left(VK_V^{-1}V^T\right). \tag{12}$$

Intuitively, these regularizers encourage corresponding latent factors for two genes to be close if two genes are similar, given a particular kernel.

For cross-species models, assuming the availability of appropriate kernels in the source species, the same technique can be applied to factors F and H, and the following quadratic term can be added for regularization:

$$\mathcal{L}_{ks} = \text{tr}\left(FK_F^{-1}F^T\right) + \text{tr}\left(HK_H^{-1}H^T\right). \tag{13}$$

2.2 A kernelized cross-species model including bias for imputing genetic interactions

In the sections above, we have described, in abstract terms, how loss terms can be composed to form EMF models with varying complexity. As an example, we describe in detail an instantiation of the EMF framework designed specifically to impute GIs. *Kernelized cross-species matrix factorization with bias* (K-XSMF-b) is a cross-species EMF model that imputes missing GIs in a target species. K-XSMF-b takes as input partially observed matrices of GIs of the target and a source species, computed cross-species gene-gene similarities, and side information in both target and source species. We graphically illustrate the components of K-XSMF-b and the greater EMF framework in Figure 1, and describe the optimization objective for K-XSMF-b in parts.

First, K-XSMF-b models per-gene biases in target and source GIs, and thus aims to minimize:

$$\mathcal{L}_1 = \sum_{(i,j)\in\Omega_x} (x_{ij} - u_i^T v_j - b_i)^2 + \lambda_s \sum_{(i,j)\in\Omega_y} (y_{ij} - f_i^T h_j - b'_i)^2. \tag{14}$$

Then, given kernels K_V and K_H over source and target genes, K-XSMF-b regularizes its factorization with:

$$\mathcal{L}_2 = \text{tr}\left(VK_V^{-1}V^T\right) + \text{tr}\left(HK_H^{-1}H^T\right) + \|U\|_F^2 + \|F\|_F^2. \tag{15}$$

Finally, loss terms that link latent factors and biases across species $\mathcal{L}_{\mathbf{x}}$, as in (5), and $\mathcal{L}_b$, as in (11), are added. Given hyperparameters λ_s, $\lambda_{\mathbf{x}}$ and λ, the full loss function that of the K-XMSF-b aims to minimize is:

$$\mathcal{L} = \mathcal{L}_1 + \lambda_{\mathbf{x}} \mathcal{L}_{\mathbf{x}} + \lambda(\mathcal{L}_2 + \mathcal{L}_b). \tag{16}$$

Thus, K-XSMF-b is a fully featured EMF model that simultaneously exploits cross-species information, side information in the source and target, and models the effect and conservation of per-gene biases.

2.3 Parameter Learning and Hyperparameter Selection

Each loss term in the various EMF models described above is differentiable. Thus all the objective functions we work with are amenable to typical gradient-based optimization algorithms. In this work we use the popular method ADAM to learn our models (Kingma and Ba, 2017).

For all models, all input GI scores in the target and source species (where applicable) are normalized to zero mean and unit variance prior to training. Accordingly, for imputed GI scores, this normalization operation is inverted prior to evaluation. The input cross-species similarity score matrix is also scaled element-wise to $[0, 1]$ prior to training.

We take care to ensure fair benchmarking of every MF model in our experiments. We use `hyperopt` to automatically tune and optimize model hyperparameters to maximize the performance of each benchmarked model (Bergstra *et al.*, 2011). Furthermore, a consistent early-stopping strategy is adopted for all models for the same purpose. Models are early-stopped when the R^2 score evaluated on the validation set fails to decrease for five consecutive iterations, or when the user-defined maximum number of iterations is reached.

For each combination of model, dataset, and proportion of training examples used, a validation set of 10% of training examples is first held out. Using this validation set, `hyperopt` (50 iterations) is used to determine the best hyperparameters to be used across multiple repeats.

2.4 Evaluation

In this work, we primarily evaluate imputation performance of models using the R^2 measure (the coefficient of determination), in contrast to prior studies that have used Pearson's ρ (the correlation coefficient). We report Pearson's ρ where context and comparison to prior work is necessary.

In evaluating EMF models, we rely on R^2 because it is a measure of *goodness-of-fit* while Pearson's ρ is a measure of *correlation* – and the latter does not imply the former. Critically, R^2 correctly rejects a model that systematically mis-estimates the magnitude of predictions while Pearson ρ fails to do so. For example, consider a poor model that systematically predicts values that are exactly half of the ground truth. Despite being very wrong, such a model would output values that have perfect correlation but low (or even negative) R^2 when compared to ground truth.

When imputing GI scores, the difference between goodness-of-fit and correlation is critical because extreme classes of GIs (e.g. synthetic lethal interactions) are binarized on strict numerical thresholds in the literature (Costanzo *et al.*, 2010). Thus, a model that systematically underestimates GI scores will also systematically under-report the number of predicted extreme GIs.

On the Costanzo *et al.* (2010) dataset, we also evaluate the ability of models to correctly classify "negative GIs" (analogous to synthetic sick or lethal) as defined by (Costanzo *et al.*, 2010). We follow Ma *et al.* (2018) and Yu *et al.* (2016) for these evaluations. That is, we impute interaction scores directly and, afterwards, vary the binarization threshold to compute the area under the precision-recall curve (AUPR).

Unless stated otherwise, we use Monte Carlo cross-validation to evaluate all experiments. For training and evaluation, GIs for *unique gene-pairs* are partitioned. Following Zitnik and Zupan (2015), if two genes A and B are in both rows and columns of an input matrix and two values are imputed (e.g. across the diagonal of the imputed matrix), the imputed scores are averaged for evaluation. All reported evaluation measures are averaged over 10 random repeats.

2.4.1 Comparison against existing factorization based methods

In our experiments, we compare our cross-species models against two existing factorization-based models. We compare our models to K-PMF (Zhou *et al.*, 2012), a model originally developed for recommender systems. To the best of our knowledge, we are the first to use K-PMF to impute GIs. We also compare our models to Network Guided Matrix Completion, a method that incorporates network information (from PPI networks or the Gene Ontology) to impute GIs (Zitnik and Zupan, 2015). We note that both K-PMF and NGMC do not account for per-gene biases and cannot incorporate information across species. Zitnik and Zupan (2015) also did not evaluate NGMC on genome-scale datasets available at the time of publication.

Hyperparameter optimization described in Section 2.3 is applied to both K-PMF and NGMC. The same early stopping criterion described in Section 2.3 is applied to K-PMF but not NGMC; all NGMC models run for 500 iterations.

2.4.2 Comparison against Gene Ontology based methods

We also compare EMF to DCell, the current state-of-the-art neural-network based approach developed by Ma *et al.* (2018), and Ontotype, the best non-deep-learning based method developed by Yu *et al.* (2016). Both methods featurize labels from GO to predict GIs in baker's yeast at genome-scale. We downloaded published data and predictions from Yu *et al.* (2016) and Ma *et al.* (2018), and for these comparisons evaluate EMF using the same 4-fold cross-validation procedure carried out by these studies.[1]

2.5 Implementation

EMF models are implemented using TensorFlow (Abadi *et al.*, 2016). For NGMC, we use the implementation released by the authors (Zitnik and Zupan, 2015). `Snakemake` is used extensively to configure and manage experiments (Köster and Rahmann, 2012). Models and scripts to reproduce experiments are publicly available at: `github.com/lrgr/emf`.

3 Experiments and results

Armed with the EMF framework defined in the previous section, we now evaluate it in three ways: first, we demonstrate its superiority to the state-of-art methods for predicting genetic interactions (GIs); next, in chromosome biology GI datasets for baker's and fission yeast, we perform a systematic ablation analysis to identify the components of EMF that capture additional signal to better impute GIs; and finally, we apply EMF to impute GIs on genome-scale datasets in both yeast species.

3.1 Data

Our experiments were performed on two pairs of GI datasets from baker's and fission yeast. The first pair of GI datasets consists of published epistatic miniarray profiles (E-MAPs) for chromosome biology genes in baker's and fission yeast (Collins *et al.*, 2007; Roguev *et al.*, 2008).

The second pair are genome-scale GI datasets in baker's and fission yeast. Ryan *et al.* (2012) produced an E-MAP covering ~60% of all non-essential genes in fission yeast. Costanzo *et al.* (2010) produced a synthetic genetic array (SGA) covering ~75% of all non-essential genes in baker's yeast. We note that in the SGA dataset for baker's yeast, a confidence measure (p-value) computed from technical replicates is also assigned to each reported GI score (Baryshnikova *et al.*, 2010).

3.1.1 Genetic interaction scores

All four datasets measure GI scores with respect to cell growth. Each yields a matrix of real valued GI scores where index (i, j) corresponds to the interaction of column gene i and row gene j. For chromosome biology datasets, matrices of GI scores are symmetric. For genome-scale datasets, the GI scores between a set of array (columns) and query genes (rows) are measured and the set of array and query genes have non-zero intersection.

[1] Published predictions from these studies were not stratified by fold. Thus, while we follow the same experimental procedure, we train our models on different folds.

Thus a unique gene-pair can correspond to two measured GI scores. We follow Ma *et al.* (2018) to associate each unique gene pair to a unique GI score. That is, if a gene-pair corresponds to GI scores of opposite signs, the GI scores are discarded. Otherwise, for baker's yeast the GI score with lower P-value is retained, and for fission yeast the average GI score is retained (as significance is not reported for this dataset).

We note that E-MAPs and SGAs both quantify a GI score between a pair of genes using similar principles. Both technologies use imaging to quantify the fitness of the double and corresponding single mutants. The GI score is then defined to be the deviation of the fitness of the double mutant from the multiplicative product of the fitnesses of the single mutants (Costanzo *et al.*, 2019). However, since E-MAPs and SGAs are different technologies, raw GI scores cannot be directly compared. Hence, we applied the normalization strategy described in Section 2.3.

All datasets are restricted to GIs between non-essential genes only. For Costanzo *et al.* (2010) data, as in Ma *et al.* (2018), GIs for temperature sensitive alleles are also removed. The matrices of GI scores for the datasets described above are all partially observed. The percentage of missing entries and size of each processed dataset are listed in Table 1.

Species	Reference	# Rows	# Columns	% Missing
Baker's yeast	Collins *et al.* (2007)	664	664	32%
	Costanzo *et al.* (2010)	3,885	1,377	19%
Fission yeast	Roguev *et al.* (2008)	536	536	21%
	Ryan *et al.* (2012)	1955	862	16%

Table 1. Summary statistics for genetic interaction datasets.

For experiments with Costanzo *et al.* (2010), all pairs regardless of significance were used for training. We report imputation performance on all scores as well as scores restricted to significant pairs ($p < 0.05$).

3.1.2 BLASTp, protein sequences, and PPI networks

We use BLASTp bitscores between proteins sequences across species as the similarity measure for cross-species EMF models. Protein sequences for baker's and fission yeast were downloaded from the Saccharomyces Genome Database and PomBase (Cherry *et al.*, 2011; Lock *et al.*, 2018) and used to compute bitscores between genes in the rows of target and source species data. Bitscores between proteins without available sequences were set to zero. For chromosome biology datasets (Collins *et al.*, 2007; Roguev *et al.*, 2008), the set of downloaded sequences covered 99.2% and 99.3% of baker's and fission yeast genes. For genome-scale datasets (Costanzo *et al.*, 2010; Roguev *et al.*, 2008), downloaded sequences covered 86.9% and 99.8% of baker's and fission yeast genes.

We downloaded protein-protein interaction (PPI) networks for baker's and fission yeast from BioGRID database version 3.5.174 (Oughtred *et al.*, 2018). These PPI networks were used for all models that incorporated side information. PPI networks were restricted to genes in the columns of each GI dataset. For chromosome biology GI datasets, the PPI networks covered 99.1% and 73.5% of genes in baker's and fission yeast. For genome-scale GI datasets, the PPI networks covered 99.6% and 78.3% of genes in baker's and fission yeast. Singletons were then added for genes in GI data missing from PPI networks.

3.2 Evaluated models

In our experiments, we seek to investigate how composable components of EMF affect, and ultimately improve, GI imputation. We implement seven model instances of the EMF framework by progressively adding components that, model per-gene biases, link factorizations across a target and source species, and regularize with side information within each species.

Of the seven EMF models, four are *single-species* models that factorize GI data in the target species only[2]:

- **Matrix Factorization (MF)** is the simplest matrix factorization model. It uses the optimization objective described in Section 2 and equation (2) (Salakhutdinov and Mnih, 2008; Koren *et al.*, 2009).
- **MF with bias (MF-b)** is the extension to MF that incorporates a latent bias term, b, as described in Section 2.1.2 and (9). ℓ_2 regularization over b is also added to prevent overfitting (Koren *et al.*, 2009).
- **Kernelized Probabilistic Matrix Factorization (K-PMF)** is the model developed by Zhou *et al.* (2012) with regularizers described in Section 2.1.4 and (12). Here, $\mathcal{L}_{kt}$ from (12) replaces the corresponding ℓ_2 regularizers in MF.

Of these models, only MF has been used to impute missing GIs in prior work (Zitnik and Zupan, 2015). To the best of our knowledge, our work is the first to evaluate MF-b and KPMF for imputing GIs. For context, we also compare EMF models to NGMC, a matrix factorization based model not that is not encompassed by the EMF framework but does utilize PPI networks for GI imputation (Zitnik and Zupan, 2015).

Additionally, we implement one other single-species model that is a novel extension to K-PMF that has not been explored in prior work:

- **K-PMF with bias (K-PMF-b)**, is an extension of the K-PMF model that incorporates per gene biases. K-PMF-b applies the same modification to K-PMF that MF-b does to MF.

To determine how EMF components which incorporate cross-species information capture complementary signal to improve performance, we evaluate three *cross-species* models of increasing complexity. These cross-species models use BLASTp bitscores to link the factorizations of GI scores in a target and source species to better impute GIs in the target. One model additionally uses PPI network information in each species to regularize factorizations. Another both models and links per-gene biases across species *and* incorporates PPI network information:

- **Cross-species Matrix Factorization (XSMF)** is the cross-species model described Section 2.1.1 with loss function as specified by (8).
- **Kernelized XSMF (K-XSMF)** is the cross-species model described Section 2.2 with model components that correspond to bias terms removed. Per-gene biases are not fitted and $\mathcal{L}_b$ is removed from the loss function defined by (16).
- **K-XSMF with bias (K-XSMF-b)** is the fully featured cross-species model described in Section 2.2.

A summary of the data and components used by NGMC and each EMF model is given in Table 2.

Algorithm		Target Species		Source Species		
Name / short description	**abbr.**	**Bias**	**PPI**	**GIs**	**Bias**	**PPI**
Matrix Factorization	MF	-	-	-	-	-
MF with bias	MF-b	✓	-	-	-	-
Kernelized Probabilistic MF	K-PMF	-	✓	-	-	-
Network Guided Matrix Completion	NGMC	-	✓	-	-	-
K-PMF with bias	K-PMF-b	✓	✓	-	-	-
Cross-species MF	XSMF	-	-	✓	-	-
Kernelized XSMF	K-XSMF	-	✓	✓	-	✓
Kernelized XSMF with bias	K-XSMF-b	✓	✓	✓	✓	✓

Table 2. Overview of benchmarked MF models. For each model, ✓ indicates the additional MF component and side information used.

[2] We note that MF, MF-b, and K-PMF were first introduced by other researchers.

Our focus is on imputing GIs in the target species; so for the three cross-species models, all available GIs in the source species are used for training while varying the proportions of the target species' GIs are held out for evaluation. For example, when baker's yeast is the target species, we train on the entire fission yeast dataset and part of the baker's yeast dataset, holding out some of the baker's yeast dataset for evaluation.

For all kernelized models, PPI network information is incorporated using regularized Laplacian kernels. For K-PMF, K-PMF-b, K-XSMF, and K-XSMF-b, the kernels used for target species factors are the identity matrix for K_U and the regularized Laplacian for K_V, respectively. Likewise, where applicable, the kernels for source species factors are the identity matrix for K_F and the regularized Laplacian for K_H, respectively. We note that hyperparameters for regularized Laplacian kernels used are also optimized via the same procedure described in Section 2.3.

During hyperparameter optimization, the maximum rank searched for in all matrix factorization algorithms is set to $k = 100$ and $k = 200$ for chromosome biology and genome-scale datasets, respectively. The ranges searched over and the selected hyperparameters for all the above-mentioned models are listed in our publicly available implementation.

3.3 Matrix factorization outperforms state-of-the-art Gene Ontology based models in baker's yeast

Surprisingly, EMF outperforms the best deep learning-based method for GI prediction even when using strictly less data.

We demonstrate this by establishing a baseline comparison, and contextual correspondence, between a simple EMF model that relies on GI data alone and Gene Ontology (GO) based state-of-the-art models, DCell and Ontotype (Ma *et al.*, 2018; Yu *et al.*, 2016). Specifically, since we expect mean GI scores across genes in genome-scale datasets to vary greatly, we choose to compare DCell and Ontotype to MF-b, the simplest model within the EMF framework that *only* requires GI data and also models per-gene biases.

We compare MF-b to DCell and Ontotype only in baker's yeast since both have only been applied to genome-scale data in baker's yeast and their predictions are publicly available. We note that DCell and Ontotype cannot predict GIs for *all* ~4.0 million unique gene-pairs with GI scores available from Costanzo *et al.* (2010) because only ~3.3 million gene-pairs can be featurized from GO (as some genes have no annotations). Even though MF-b does not have the same limitation, we nonetheless restrict MF-b to only use the 3.3 million GI scores used by Yu *et al.* (2016) and Ma *et al.* (2018) to perform an apples-to-apples comparison.

Though we argue in Section 2.4 that R^2 is a better metric, we report regression performance both in terms of R^2 and Pearson's ρ for context. Ma *et al.* (2018) and Yu *et al.* (2016) only report performance in terms of Pearson's ρ in their work. Following prior studies (Yu *et al.*, 2016; Ma *et al.*, 2018), we also evaluate how well each model predicts extreme GIs and report the AUPR achieved by each model for classifying negative GIs (see Section 2.4 for details).

First, following Yu *et al.* (2016) and Ma *et al.* (2018), we evaluate predictions restricted to the subset of GI scores deemed *significant* by Costanzo *et al.* (2010). When imputing significant GI scores, MF-b outperforms DCell and Ontotype by 17.5% and 75.1% in Pearson's ρ (0.604 versus 0.514 and 0.245), respectively. In terms of R^2, MF-b more than doubles the R^2 score of Ontotype (0.271 versus 0.112) and outperforms DCell (0.265). Furthermore, when classifying negative GIs, MF-b again outperforms DCell and Ontotype, achieving 18.8% and 66.2% improvement in AUPR (0.570 versus 0.480 and 0.343) over DCell and Ontotype.

Second, we hypothesize that methods that perform better at imputing *all* GI scores may be less sensitive to noise or variability in the data; hence we also evaluate models with respect to all imputed scores. When imputing all GI scores, MF-b achieves double the Pearson's ρ of Ontotype and improves over DCell by 19.0% (0.425 versus 0.191 and 0.358, respectively). When classifying negative GIs, MF-b again doubles the AUPR of Ontotype and improves over DCell by 31.5% (0.267 versus 0.104 and 0.203). Surprisingly, Ontotype and DCell both achieve *negative* R^2 scores while MF-b achieves an R^2 score of 0.187. These results indicate that, on all scores, DCell and Ontotype perform worse than a model that predicts the mean. One reason for this, shown by Fig. S1, is that when DCell predicts the sign of a GI score incorrectly, it does so more often with greater magnitude than MF-b.

MF-b and other matrix factorization based models presented in this study are also faster to train. On a machine with an NVidia GTX 1080Ti GPU, EMF models take less than 1 minute to train on the baker's yeast dataset. In fact, the benchmarked MF-b model trains in less than 3 seconds. In comparison, the authors of DCell report in their publicly available software release that the "running time on a standard Tesla K20 GPU takes 2-3 days" [3] on Costanzo *et al.* (2016).

Having established that MF-b, a simple model in the EMF framework, outperforms deep learning and GO-based methods under three different measures, we refrain from comparing other matrix factorization methods to DCell and Ontotype in subsequent sections. Furthermore, since Pearson's ρ fails to detect systematic mis-estimation of GI score magnitudes (see Section 2.4), subsequent experiments are evaluated using R^2 only.

3.4 Ablation analysis on matched chromosome biology datasets in baker's and fission yeast

Next, via an ablation analysis, we evaluate how components of the EMF framework affect GI imputation. We perform this analysis on GI datasets for chromosome biology genes in baker's and fission yeast (Collins *et al.*, 2007; Roguev *et al.*, 2008). To compare to prior work, we also compare EMF models to NGMC (Zitnik and Zupan, 2015). To explore the range of settings that may arise in practice, we evaluate models with varying amounts of training data in the target species. These allow us to assess both data-rich and data-scarce settings. Our results demonstrate that EMF models that jointly exploit cross-species and side information consistently impute GIs more accurately. We report these results in Tables 3, and standard deviations across repeats in Tables S1 and S2.

Algorithm	**% of GIs used in training**							
	Baker's yeast				Fission yeast			
	10%	*25%*	*50%*	*75%*	*10%*	*25%*	*50%*	*75%*
MF	0.054	0.178	0.303	0.380	0.093	0.220	0.370	0.464
MF-b	0.069	0.183	0.308	0.385	0.113	0.234	0.371	0.464
K-PMF	0.105	0.215	0.329	0.397	0.119	0.266	0.397	0.472
NGMC	0.050	0.207	0.304	0.329*	0.081	0.256	0.396*	0.479*
K-PMF-b	0.102	0.218	0.326	0.393	0.136	0.273	0.391	0.475
XSMF	0.070	0.181	0.304	0.386	0.106	0.232	0.373	0.466
K-XSMF	0.104	0.217	0.327	**0.399**	0.142	**0.278**	**0.405**	**0.480**†
K-XSMF-b	**0.116**	**0.225**	**0.330**	0.397	**0.155**	0.270	0.394	0.476

Table 3. R^2 score of imputed versus actual GI scores for chromosome biology datasets in baker's and fission yeast (Collins et al., 2007; Roguev et al., 2008). Models are evaluated with varying proportions of GI scores used during training. The best performing models are indicated in bold. † Standard deviations of best performing model and MF baseline overlap. * Folds that did not converge were excluded from evaluation.

Our results first show that modeling per-gene biases aids GI imputation. The improvement gained by modeling biases is most clear when comparing the performance of MF-b to MF. MF-b outperforms MF in all but one experiment. We note that, when cross-species and side information are

[3] `github.com/idekerlab/DCell`

used, improvement due to modeling biases is less consistent. When 50% and 75% of GIs are observed during training, the difference in performance when adding bias terms to K-XSMF and K-PMF is small. However when data is scarce, modeling biases more consistently improves imputation. For example, when 10% of GIs are used during training, K-XSMF-b outperforms K-XSMF by 11.5% and 9.2% in baker's and fission yeast, respectively. In fact, in these scenarios, K-XSMF-b outperforms K-XSMF, and K-PMF-b outperforms K-PMF, in six out of eight experiments.

Unsurprisingly, all models that exploit side information consistently outperform corresponding models that do not. We highlight some results for the most data-scarce and data-rich scenarios. In baker's yeast, when 10% and 75% of GIs are used, K-PMF-b outperforms MF-b by 47.8% and 3.4%, and K-XSMF outperform XSMF by double and by 3.4% percent, respectively. In fission yeast, when 10% and 75% of GIs are used, K-PMF-b outperforms MF-b by 20.4% and 2.4%, and K-XSMF outperforms XSMF by 30.2% and 1.1%. We note that while both models use the same side information, K-PMF outperforms NGMC in seven of eight experiments across both yeast species.

Moreover, our results show that cross-species models outperform single-species models. Exploiting cross-species information only, XSMF outperforms MF across the board, albeit by a small margin when data is abundant. Again, differences in performances are largest when data is scarce. When 10% of GIs are used, XSMF outperforms MF by 29.6% and 14.0% in baker's and fission yeast, respectively.

Most strikingly, models that exploit cross-species *and* side information (K-XSMF-b and K-XSMF) are the best performing models (bolded in Table 3) and outperform the MF baseline without overlapping standard deviations in all but one case. Again, data-scarce scenarios show the largest differences in performance. In baker's yeast, when 10% and 25% of target GIs are used during training, K-XSMF-b outperforms the next best single-species model by 10.5% and 3.2%, respectively. In fission yeast, K-XSMF-b outperforms K-PMF-b by 14.0% when 10% of target GIs are used, and K-XSMF outperforms K-PMF-b 1.8% when 25% of GIs are used. We highlight that K-XSMF and K-XSMF-b not only exploit cross-species information, but also side information in *both* target *and* source species. These results highlight the utility and versatility of the EMF framework.

It is particularly notable that EMF components (i.e., biases, exploiting side and cross-species information) offer largest improvements in imputation performance when data is scarce. Data-scarce scenarios are most likely to occur when new methods for measuring GIs for new phenotypes or new species are developed.

3.5 Results on genome-scale datasets in baker's and fission yeast

Finally, we evaluate a representative set of EMF models on genome-scale GI datasets in baker's and fission yeast (Costanzo *et al.*, 2010; Ryan *et al.*, 2012). On these datasets, our results show that incorporating cross-species information aids GI imputation when training examples are scarce.

Again, since EMF models do not depend on GO, EMF models are able to impute interactions between *all* 4.0 million unique baker's yeast gene-pairs published in Costanzo *et al.* (2010) as opposed to the 3.3 million featurizable gene-pairs in Yu *et al.* (2016) and Ma *et al.* (2018). To the best of our knowledge, our study is the first to predict GIs at genome-scale in fission yeast, and GIs for all 4.0 million gene-pairs measured by Costanzo *et al.* (2010) in baker's yeast.

In baker's yeast, as in Section 3.3, we evaluate both all imputed scores and the subset of scores deemed to be significant by Costanzo *et al.* (2010). In fission yeast, we evaluate imputed scores for all held-out gene-pairs since Ryan *et al.* (2012) do not report p-values for measured GI scores. We report the results in Tables 4 and 5, and standard deviations across repeats in Tables S3, S4 and S5.

Perhaps unsurprisingly, when a large amount of data is available, the differences in the best performing models are almost indistinguishable. In fission yeast, XSMF and K-XSMF-b outperforms MF by a small margin when 75% of GIs are used during training. When 50% of GIs are observed during training, K-PMF-b outperforms K-XSMF-b by just over 1%. In baker's yeast, when 75% of GIs are observed during training, MF and XSMF are the best performing models and outperform their kernelized counterparts by small margins. Here, one key observation is that in these data rich settings, cross-species components of the EMF framework do not impair the single-species matrix factorization models which they extend.

Algorithm	**% of GIs used in training**			
	10%	*25%*	*50%*	*75%*
MF	0.049	0.147	0.251	0.316
K-PMF-b	0.064	**0.159**	**0.257**	0.317
XSMF	0.062	0.153	0.252	**0.318**†
K-XSMF-b	**0.067**	0.158	0.254	**0.318**†

Table 4. R^2 score of imputed versus actual GI scores for EMF models in genome-scale fission yeast dataset (Ryan et al., 2012). Notation is the same as in Table 3.

Algorithm	**% of GIs used in training**			
	10%	*25%*	*50%*	*75%*
MF	0.004 (0.007)	0.088 (0.055)	0.180 (0.133)	**0.267** (**0.189**)
K-PMF-b	**0.026** (0.009)	**0.100** (**0.061**)	0.180 (0.126)	0.238 (0.176)
XSMF	0.005 (0.006)	0.084 (0.059)	0.190 (**0.134**†)	0.266 (**0.189**†)
K-XSMF-b	0.019 (**0.011**)	0.085 (**0.061**)	**0.200** (0.130)	0.250 (0.182)

Table 5. R^2 score of imputed versus actual GI scores for EMF models in genome-scale baker's yeast dataset (Costanzo et al., 2010). Scores for predictions restricted to significant GI scores as determined by Costanzo et al. (2010) appear on the left. Scores for predictions on all pairs to the right in parentheses. Other notation is the same as in Table 3.

However, when data is scarce, the improved performance of EMF models due to the inclusion of side information and cross-species information is clear. When fewer than 75% of observed GIs are used during training, the best performing EMF models outperform the MF baseline without overlapping standard deviations in all but one case. In fission yeast, when 10% and 25% of observed GIs are used during training, K-XSMF-b and K-PMF-b are the best performing models. Further, both cross-species models improve over their single-species counterparts: K-XSMF-b outperforms K-PMF-b by 5%, when 10% of observed GIs are used during training, and XSMF outperforms MF by 27% and 4%, when 10% and 25% of observed interactions are used for training.

Likewise, the inclusion of cross-species information and side information aids imputation in baker's yeast when data is scarce. When imputing significant pairs, both K-PMF-b and K-XSMF-b roughly quadruple the R^2 score of their non-kernelized counterparts when 10% of GIs are used during training. Here, K-PMF-b is clearly the best performing model when 10% and 25% of GIs are used during training. Finally, when imputing all GIs, cross-species models XSMF and K-XSMF-b achieve the best R^2 score, when 10%, 25% and 50% of GIs are used during training.

4 Discussion

In this work, we introduce *Extensible Matrix Factorization* EMF, a framework of composable matrix factorization (MF) models for imputing genetic interactions (GIs). The EMF framework unifies several MF strategies for improving imputation. EMF models can explicitly model per-gene biases, and can readily exploit available side information via kernelization. A novel contribution of EMF models is the ability to simultaneously exploit *cross-species* information. Given a cross-species

gene-gene similarity measure, EMF models can link factorizations in a *source* and *target* species to better impute missing values in the target.

Surprisingly, even a simple EMF model outperforms the state-of-the-art method for GI prediction. This simple model only requires GIs as input and does not require labels from the Gene Ontology. Via an ablation analysis in chromosome biology GI datasets in baker's and fission yeast, we show how components of the EMF framework improve GI imputation. Furthermore, our results show that EMF models are also effective in genome-scale datasets in both yeast species. To the best of our knowledge, our study is the first to impute GIs in fission yeast at genome-scale.

In sum, the EMF framework highlights the versatility, and surprising utility, of MF based approaches. Our results show that components in the EMF framework that exploit cross-species information are most effective when data is *scarce*. We also emphasize that data scarcity is relative. For example, 10% of available data in baker's yeast equates to approximately 400,000 observations, which is more than have been measured in all but a handful of species. Thus, we expect MF based approaches like EMF to be invaluable for efforts to map GIs in new species. In these scenarios, the incorporation of data across multiple contexts, be it species or phenotypes, may be fruitful if not necessary. Though not the focus of this work, we also anticipate that the performance of cross-species models could be improved via other cross-species similarity measures and other methodological optimizations (e.g. combining kernels via multiple kernel learning (Gönen and Alpaydin, 2011)).

Funding

We gratefully acknowledge support from NSF award CNS-1618207 to MC, and NSF award DGE-1632976 to JF. This project has been made possible in part by support to MDML and JF through grant number 2018-182608 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation.

Conflict of interest. MDML was a paid consultant for Microsoft during the time when part of this study was performed.

References

Abadi, M. *et al.* (2016). Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, Savannah, GA. USENIX Association.

Altschul, S. F. *et al.* (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3), 403 – 410.

Ashworth, A. and Lord, C. J. (2018). Synthetic lethal therapies for cancer: what's next after parp inhibitors? *Nature Reviews Clinical Oncology*, **15**(9), 564–576.

Baryshnikova, A. *et al.* (2010). Chapter 7 - synthetic genetic array (sga) analysis in saccharomyces cerevisiae and schizosaccharomyces pombe. In *Guide to Yeast Genetics: Functional Genomics, Proteomics, and Other Systems Analysis*, volume 470 of *Methods in Enzymology*, pages 145 – 179. Academic Press.

Benstead-Hume, G. *et al.* (2019). Predicting synthetic lethal interactions using conserved patterns in protein interaction networks. *PLOS Computational Biology*, **15**(4), 1–25.

Bergstra, J. S. *et al.* (2011). Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems*, pages 2546–2554.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, **9**(6), 717.

Cherry, J. M. *et al.* (2011). Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Research*, **40**(D1), D700–D705.

Collins, S. R. *et al.* (2007). Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature*, **446**(7137), 806–810.

Costanzo, M. *et al.* (2010). The genetic landscape of a cell. *Science*, **327**(5964), 425–431.

Costanzo, M. *et al.* (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science*, **353**(6306), aaf1420.

Costanzo, M. *et al.* (2019). Global genetic networks and the genotype-to-phenotype relationship. *Cell*, **177**(1), 85 – 100.

Dixit, A. *et al.* (2016). Perturb-seq: Dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, **167**(7), 1853 – 1866.e17.

Fan, J. *et al.* (2019). Functional protein representations from biological networks enable diverse cross-species inference. *Nucleic Acids Research*, **47**(9), e51–e51.

Gönen, M. and Alpaydin, E. (2011). Multiple kernel learning algorithms. *J. Mach. Learn. Res.*, **12**, 2211–2268.

Hutchison, C. A. *et al.* (2016). Design and synthesis of a minimal bacterial genome. *Science*, **351**(6280).

Jacunski, A. *et al.* (2015). Connectivity homology enables inter-species network models of synthetic lethality. *PLOS Computational Biology*, **11**(10), e1004506.

Kingma, D. P. and Ba, J. (2017). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. arXiv: 1412.6980.

Koch, E. N. *et al.* (2012). Conserved rules govern genetic interaction degree across species. *Genome Biology*, **13**(7), R57.

Koren, Y. *et al.* (2009). Matrix factorization techniques for recommender systems. *Computer*, **42**(8), 30–37.

Köster, J. and Rahmann, S. (2012). Snakemake - a scalable bioinformatics workflow engine. *Bioinformatics*, **28**(19), 2520–2522.

Kuzmin, E. *et al.* (2018). Systematic analysis of complex genetic interactions. *Science*, **360**(6386).

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**(6755), 788–791.

Lee, J.-S. *et al.* (2018). Harnessing synthetic lethality to predict the response to cancer treatment. *Nature Communications*, **9**(1), 2546.

Leslie, C. *et al.* (2001). The spectrum kernel: A string kernel for svm protein classification. In *Biocomputing 2002*, pages 564–575, Kauai, Hawaii, USA. World Scientific.

Liben-Nowell, D. and Kleinberg, J. (2003). The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 556–559, New York, NY, USA. Association for Computing Machinery.

Lock, A. *et al.* (2018). PomBase 2018: user-driven reimplementation of the fission yeast database provides rapid and intuitive access to diverse, interconnected information. *Nucleic Acids Research*, **47**(D1), D821–D827.

Lü, L. and Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, **390**(6), 1150 – 1170.

Ma, J. *et al.* (2018). Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, **15**(4).

Oughtred, R. *et al.* (2018). The BioGRID interaction database: 2019 update. *Nucleic Acids Research*, **47**(D1), D529–D541.

Paladugu, S. R. *et al.* (2008). Mining protein networks for synthetic genetic interactions. *BMC Bioinformatics*, **9**(1), 426.

Pandey, G. *et al.* (2010). An integrative multi-network and multi-classifier approach to predict genetic interactions. *PLOS Computational Biology*, **6**(9), 1–14.

Patrick, K. L. *et al.* (2018). Quantitative yeast genetic interaction profiling of bacterial effector proteins uncovers a role for the human retromer in salmonella infection. *Cell Systems*, **7**(3), 323 – 338.e6.

Rendle, S. *et al.* (2019). On the difficulty of evaluating baselines: A study on recommender systems. *CoRR*, **abs/1905.01395**.

Roguev, A. *et al.* (2008). Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science*, **322**(5900), 405–410.

Ryan, C. *et al.* (2012). Hierarchical modularity and the evolution of genetic interactomes across species. *Molecular Cell*, **46**(5), 691–704.

Salakhutdinov, R. and Mnih, A. (2008). Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, volume 20.

Schuldiner, M. *et al.* (2005). Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell*, **123**(3), 507–519.

Stein-O'Brien, G. L. *et al.* (2018). Enter the matrix: Factorization uncovers knowledge from omics. *Trends in Genetics*, **34**(10), 790–805.

The Gene Ontology Consortium (2018). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, **47**(D1), D330–D338.

Ulitsky, I. *et al.* (2009). Towards accurate imputation of quantitative genetic interactions. *Genome Biology*, **10**(12), 1–18.

Wong, S. L. *et al.* (2004). Combining biological networks to predict genetic interactions. *Proceedings of the National Academy of Sciences*, **101**(44), 15682–15687.

Wu, M. *et al.* (2014). In silico prediction of synthetic lethality by meta-analysis of genetic interactions, functions, and pathways in yeast and human cancer. *Cancer informatics*, **13**(Suppl 3), 71–80. 25452682[pmid].

Yu, M. *et al.* (2016). Translation of genotype to phenotype by a hierarchy of cell subsystems. *Cell Systems*, **2**(2), 77–88.

Zhou, T. *et al.* (2012). Kernelized probabilistic matrix factorization: Exploiting graphs and side information. In *Proceedings of the 12th SIAM International Conference on Data Mining, SDM 2012*, pages 403–414.

Zitnik, M. and Zupan, B. (2015). Data imputation in epistatic maps by network-guided matrix completion. *Journal of Computational Biology*, **22**(6), 595–608. PMID: 25658751.

Supplementary Material

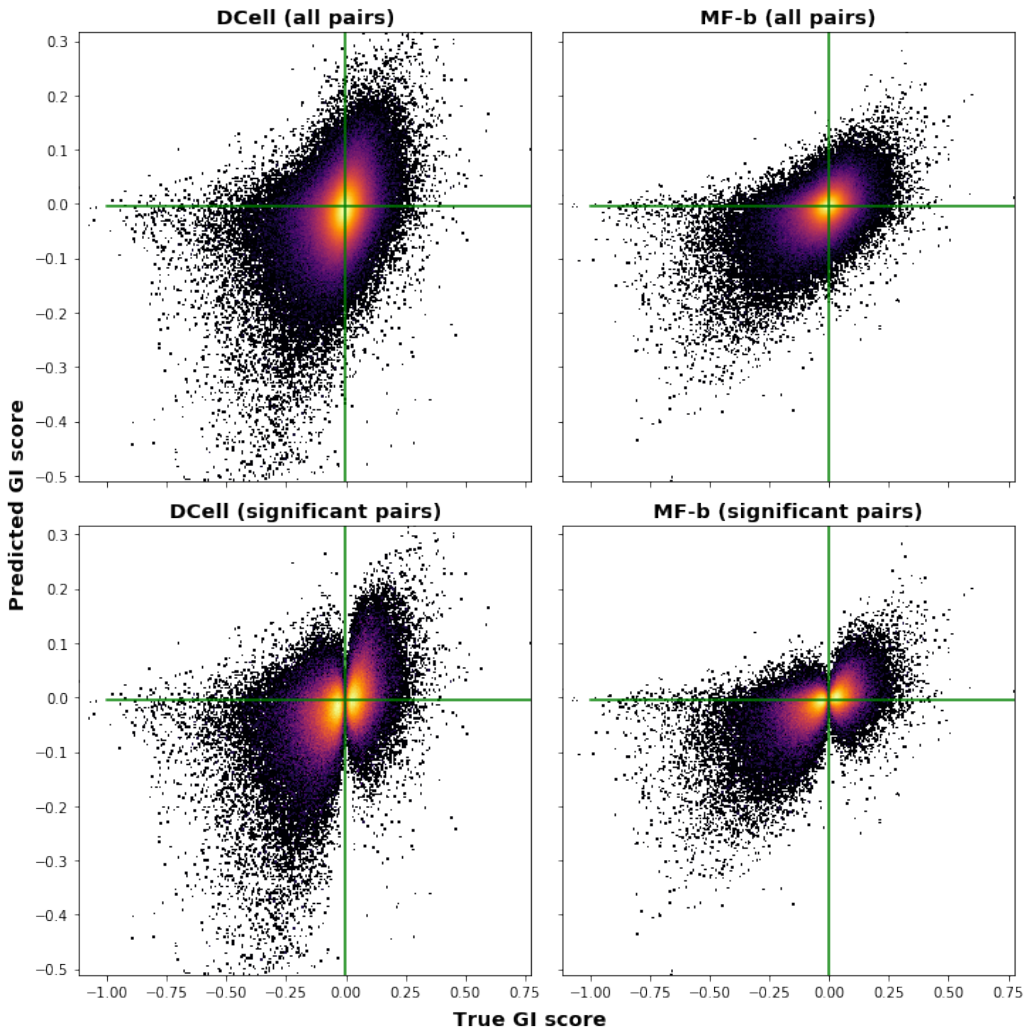

Fig. S1. True (x-axis) versus predicted (y-axis) GI scores by MF-b and DCell (Ma et al., 2018). MF-b is trained with the GI scores for 3.3 million featurizable gene-pairs published by Ma et al. (2018) and Yu et al. (2016). In green, $x = 0$ and $y = 0$ are plotted.

Algorithm	**% of GIs used in training**			
	10%	*25%*	*50%*	*75%*
MF	0.054 ± 0.009	0.178 ± 0.004	0.303 ± 0.005	0.380 ± 0.006
MF-b	0.069 ± 0.005	0.183 ± 0.005	0.308 ± 0.002	0.385 ± 0.006
K-PMF	0.105 ± 0.007	0.215 ± 0.004	0.329 ± 0.005	0.397 ± 0.010
NGMC	0.050 ± 0.011	0.207 ± 0.004	0.304 ± 0.010	0.329 ± 0.028*
K-PMF-b	0.102 ± 0.006	0.218 ± 0.007	0.326 ± 0.007	0.393 ± 0.006
XSMF	0.070 ± 0.004	0.181 ± 0.005	0.304 ± 0.004	0.386 ± 0.007
K-XSMF	0.104 ± 0.005	0.217 ± 0.003	0.327 ± 0.004	$\mathbf{0.399 \pm 0.006}$
K-XSMF-b	$\mathbf{0.116 \pm 0.007}$	$\mathbf{0.225 \pm 0.004}$	$\mathbf{0.330 \pm 0.003}$	0.397 ± 0.005

Table S1. Mean and standard deviation of R^2 scores for imputed versus actual GI scores for chromosome biology dataset in baker's yeast (Collins et al., 2007). Models are evaluated with varying proportions of GI scores used during training. The best performing models are indicated in bold. * Folds that did not converge were excluded from evaluation.

Algorithm	**% of GIs used in training**			
	10%	*25%*	*50%*	*75%*
MF	0.093 ± 0.011	0.220 ± 0.012	0.370 ± 0.005	0.464 ± 0.007
MF-b	0.113 ± 0.005	0.234 ± 0.005	0.371 ± 0.002	0.464 ± 0.006
K-PMF	0.119 ± 0.010	0.266 ± 0.005	0.397 ± 0.005	0.472 ± 0.009
NGMC	0.081 ± 0.008	0.256 ± 0.007	0.396 ± 0.006*	0.479 ± 0.011*
K-PMF-b	0.136 ± 0.010	0.273 ± 0.007	0.391 ± 0.007	0.475 ± 0.010
XSMF	0.106 ± 0.011	0.232 ± 0.008	0.373 ± 0.005	0.466 ± 0.009
K-XSMF	0.142 ± 0.009	**0.278 ± 0.007**	**0.405 ± 0.005**	**0.480 ± 0.009**
K-XSMF-b	**0.155 ± 0.007**	0.270 ± 0.008	0.394 ± 0.007	0.476 ± 0.008

Table S2. Mean and standard deviation of R^2 scores for imputed versus actual GI scores for chromosome biology dataset in fission yeast (Roguev et al., 2008). Models are evaluated with varying proportions of GI scores used during training. The best performing models are indicated in bold. * Folds that did not converge were excluded from evaluation.

Algorithm	**% of GIs used in training**			
	10%	*25%*	*50%*	*75%*
MF	0.049 ± 0.002	0.147 ± 0.001	0.251 ± 0.002	0.316 ± 0.002
K-PMF-b	0.064 ± 0.002	**0.159 ± 0.002**	**0.257 ± 0.001**	0.317 ± 0.002
XSMF	0.062 ± 0.002	0.153 ± 0.002	0.252 ± 0.001	**0.318 ± 0.002**
K-XSMF-b	**0.067 ± 0.002**	0.158 ± 0.002	0.254 ± 0.001	**0.318 ± 0.002**

Table S3. Mean and standard deviation of R^2 scores for imputed versus actual GI scores for EMF models in genome-scale fission yeast dataset (Ryan et al., 2012). The best performing models are indicated in bold.

Algorithm	**% of GIs used in training**			
	10%	*25%*	*50%*	*75%*
MF	0.004 ± 0.001	0.088 ± 0.002	0.180 ± 0.001	**0.267 ± 0.002**
K-PMF-b	**0.026 ± 0.001**	**0.100 ± 0.002**	0.180 ± 0.001	0.238 ± 0.002
XSMF	0.005 ± 0.001	0.084 ± 0.002	0.190 ± 0.001	0.266 ± 0.002
K-XSMF-b	0.019 ± 0.000†	0.085 ± 0.001	**0.200 ± 0.001**	0.250 ± 0.002

Table S4. Mean and standard deviation of R^2 scores for imputed versus actual GI scores for EMF models in genome-scale baker's yeast dataset (Costanzo et al., 2010). Results are restricted to significant GI scores as determined by Costanzo et al. (2010). The best performing models are indicated in bold. † Indicates standard deviation less than 0.0005.

Algorithm	**% of GIs used in training**			
	10%	*25%*	*50%*	*75%*
MF	0.007 ± 0.000†	0.055 ± 0.001	0.133 ± 0.001	**0.189 ± 0.001**
K-PMF-b	0.009 ± 0.000†	**0.061 ± 0.001**	0.126 ± 0.001	0.176 ± 0.001
XSMF	0.006 ± 0.000†	0.059 ± 0.001	**0.134 ± 0.001**	**0.189 ± 0.001**
K-XSMF-b	**0.011 ± 0.000†**	**0.061 ± 0.001**	0.130 ± 0.001	0.182 ± 0.001

Table S5. Mean and standard deviation of R^2 scores for all imputed versus actual GI scores for EMF models in genome-scale baker's yeast dataset (Costanzo et al., 2010). The best performing models are indicated in bold. † Indicate standard deviation less than 0.0005.