Data Fusion by Matrix Completion for Exposome Target Interaction Prediction

Kai Wang[1], Nicole Kim[1], Maryam Bagherian[1,2], Kai Li[1], Elysia Chou[1], Justin A. Colacino[3,4], Dana C. Dolinoy[3,4], Maureen A. Sartor[1,5,*]

[1]Department of Computational Medicine and Bioinformatics, School of Medicine, University of Michigan, Ann Arbor, MI 48109, USA

[2]Michigan Institute for Data Science (MIDAS), University of Michigan, Ann Arbor, MI 48109, USA

[3]Department of Environmental Health Sciences, School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA

[4]Department of Nutritional Sciences, School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA

[5]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA

***Correspondence**:
Maureen A Sartor
sartorma@umich.edu
Room 2017, Palmer Commons
100 Washtenaw Avenue
Ann Arbor, MI 48109-2218

Conflicts of Interest
The authors declare they have nothing to disclose.

Abstract

Background: Exposure to toxic chemicals presents a huge health burden and disease risk. Key to understanding which chemicals cause specific toxic effects is determining the molecular target(s) of these chemicals. Given that over a thousand new chemicals are produced annually, it is infeasible to perform a comprehensive safety assessment for all novel chemicals due to limited resources. Thus, a robust computational method for discovering targets of environmental exposures, which can then be used to prioritize chemicals for further study, is a promising direction for public health research.
Objectives: We implemented a novel matrix completion algorithm named coupled matrix-matrix completion (CMMC) for predicting exposome-target interactions.
Methods: The low rank matrix completion problem can be described as predicting missing values of an incomplete un-observed/under-observed or partially observed matrix. Different from previous matrix completion algorithms based on matrix factorization, the CMMC algorithm incorporates two coupled matrices in form of auxiliary information to help better optimize the distance metric and maintains a fast runtime. The present data fusion model has the potential to automatically reveal shared and unshared components through modeling constraints. The two coupled matrices represent the relationships among the chemical exposures and gene targets, respectively, and help jointly analyze incomplete data sets. We collected and processed relevant data from the Comparative Toxicogenomics Database (CTD) as a benchmark dataset for implementing and testing our method and comparing it to alternative prediction methods.
Results: Our method achieved an AUC of 0.89 on the benchmark dataset generated in this study. Compared to other state-of-the-art methods, CMMC achieved the best performance for predicting environmental chemical-gene target interactions. The case study shows that CMMC can be used to discover potential molecular targets of novel chemicals without any prior bioactivity knowledge.
Discussion: Our CMMC approach is a powerful method for predicting the target genes of environmental exposures.

Introduction

Phenotype can be defined as all observable characteristics of an organism, which are influenced by genotype and the environment. Genomic studies have provided extraordinary genetic knowledge corresponding to different phenotypes, including generating revolutionary knowledge used to investigate the genetic origin of diseases. However, these studies also revealed that genetic variation only partially contributes to the origin of diseases[1]. A great portion of human disease is not caused by genetic variation but by environmental factors, however the effects of environmental factors on phenotypes are not as well studied due to the complexity of environmental exposures[2]. In 2005, as a complement to the genome, the "exposome" was conceived by C. P. Wild[3], defined as the totality of exposures to which the public comes in contact and the corresponding biological responses.

Environmental exposures include but are not limited to environmental chemicals that can enter the human body through routes of exposure including inhalation, ingestion, or dermal. The Global Burden of Disease project[4] estimates that 9 million deaths per year are caused by air, water, and soil pollution alone. This number is likely to be an underestimate due to the fact that many new chemical compounds with uncharacterized toxicological profiles are created and released into our environment annually. In PubChem[5,6], the number of registered chemical compounds increased from 10 million in 2008 to 100 million in 2021. To perform comprehensive toxicity screenings of all these chemicals and study their links to health effects represent a great challenge due to limited time and funds[7]. Traditional animal-based *in vivo* toxicity testing is an expensive and time-consuming task. In addition, the power of using animal models to predict human response to chemicals is limited due to often poor correlation with human

findings[8]. Compared to *in vivo* studies, *in vitro* studies that assess the toxic effects of chemicals on cultured bacteria or mammalian cells can greatly accelerate toxicity testing and reduce animal usage[9]. However, *in vivo* physiological environments cannot be easily mimicked in the *in vitro* setting[10]. Therefore, *in vivo* and *in vitro* methods in toxicity testing can be integrated to perform a comprehensive safety assessment, but unfortunately, performing these safety assessments over millions of chemicals is resource intensive and thus impractical.

Computational methods are widely used in toxicology research to organize and analyze different types of data[11]. Corresponding methods have also been developed and applied to predict chemical toxicity. Previously developed toxicity prediction methods include structure-activity relationship (SAR)[12,13], read-across[14,15], and quantitative structure-activity relationship (QSAR)[16–18]. These methods usually focus on using chemical features to predict a binary label for chemicals, such as whether a chemical is a carcinogen or not, or whether the chemical activates certain marker genes or related biological pathways. Thus, the overall corresponding biological responses caused by a chemical cannot be described through the aforementioned methods. Furthermore, these methods do not exploit the vast amount of data accumulated to date for similar chemicals and molecular targets. Therefore, a method that takes advantage of both chemical features and biological knowledge, as well as the known complex relationships among them, could radically improve and expand the scope of chemical toxicity predictions.

Similar to exposure-target predictions, many computational methods were developed and applied to find drug-target interactions (DTI) in pharmaceutical studies, such as deep learning methods, tree-based methods, network-based methods, etc. Matrix completion is a task to find the missing values in a partially observed matrix. Corresponding algorithms and applications have been widely used in DTI prediction. Compared to other types of methods, a benefit of matrix completion algorithms is they can easily handle big datasets with high sparsity without overfitting[19]. In DTI predictions, a binary value (i.e., 1 or 0) is usually used to indicate the interaction between a chemical drug and its molecular targets, because drugs are typically designed with specific molecular targets in mind. Different from DTI, chemical exposures were not designed to interact with any specific gene targets, they often gave broader ranging effects, and thus a continuous value approach is more appropriate than the typical binary approach used for DTI.

In this study, we take an important step towards this goal by implementing a new matrix completion algorithm[20] called coupled matrix-matrix completion (CMMC) to predict the target genes of environmental chemicals on a broad scale. CMMC algorithm falls under the youngest branch of data fusion that is intermediate/partial integration, i.e., it does not necessarily fuse the input data, nor does it develop separate models for each source. Instead, it preserves the structure of the data sources and merges them at the level of a predictive model[21,22], here CMMC. To our knowledge, this is the first time such a large-scale prediction effort has been attempted for exposure-target interactions. CMMC differs from previous matrix completion algorithms based on matrix factorizations and is subject to a Distance Metric Learning (DML) constraint, i.e., it directly utilizes supporting information (i.e., coupled matrices about similarities among chemicals and genes) in the main iteration to learn an optimal metric for imputing missing values. The method benefits from strong results in group action and representation theory and guarantees a unique solution for the optimization problem. A different implementation of CMMC using a binary matrix was shown to outperform alternative approaches in the related task of predicting drug target interactions[20]. In this study compared with previously published methods, CMMC achieved better performance on the benchmark datasets. To further test our approach, the representative endocrine disrupting chemical, bisphenol A (BPA) and three of structurally related replacements, i.e., bisphenol B (BPB), Bis (4-hydroxyphenyl) methane (BPF), and 4,4'-Sulfonyldiphenol (BPS), were utilized in a case study. Compared to the bioactivity data from the CompTox database[23], over half of the active genes of

these four chemicals in CompTox were predicted in the top 1000 interactive genes in our results. In summary, CMMC is a powerful tool to discover potential chemical-gene interactions and provide vital information for prioritizing further toxicology research.

Methods and Materials

CMMC Algorithm

A low-rank matrix completion for an incomplete unobserved/under-observed matrix **X** can be formulated as:

$$\min_{\mathbf{M}} \|\mathbf{M}\|_{\mu}, \text{ s. t. } \mathbf{X}_{\Omega} = \mathbf{M}_{\Omega} \qquad (1)$$

where Ω denotes an index set and μ is a proper metric to be determined. In other words, given an incomplete unobserved/under-observed matrix **X**, the goal is to find matrix **M** whose rank is minimum, and it has all the known entries in **X**. Most methods for matrix (also tensor) completion rely upon the choice of a fixed metric, for instance the *Euclidean* or *nuclear* norm. If there is a high correlation between the rows/columns in a matrix, then a different metric given by the data itself could be adopted. This is part of a well-known branch in machine learning known as Distance Metric Learning that aims to learn distances from the data[24]. When working with problems including data points in *n*-dimensional space, $\square^{n}$, *Mahalanobis distance*, as the inverse of covariance matrix of the data, learns a metric induced by the data points itself. As proposed in Bagherian M, et al[20], CMMC method development is heavily based on algebraic property of symmetric matrices which can be thought of elements of *General Linear Group* (GL) and *Special Linear Group* (SL). These linear groups belong to a class of groups, called *reductive* groups[25]. Norm minimization problems when put in the form of reductive groups (with a proper group action) can in fact obtain a unique minimum (up to an orbit) using the algebraic results given in *Kempf-Ness theorem*[25]. *Mahalanobis distance* is equivalent to *Euclidean distance* after a linear change of coordinate and normalization with respect to covariance matrix. The *Kempf-Ness theorem* shows that there is essentially a *unique* change of coordinates that is optimal in a certain sense.

Let data matrix $\mathbf{X} \in \square^{n\times m}$ be given, where to any of its modes, n and m, auxiliary information in form of symmetric matrices is present, i.e., adjacent to the mode *n*, there exists an $n \times n$ square symmetric matrix, called $\mathbf{M}_{nn}$ and the same for the mode m and matrix $\mathbf{M}_{mm}$. These two coupled/auxiliary matrices can be thought of regularization terms of the matrix completion problem.

The goal is to find completed observed matrix, $\mathbf{Y} \in \square^{n\times m}$ such that the known entries of **X** and **Y** match. To this end, through the following optimization problem, we first learn two matrices $\mathbf{S}_{nn}$ and $\mathbf{S}_{mm}$ using the given side information, $\mathbf{M}_{nn}$ and $\mathbf{M}_{mm}$, and use them to learn the low rank completed version of **X**, i.e., **Y**. Therefore, the optimization problem to be minimized can be outlined as:

$$\min_{\Gamma} \left\|\mathbf{S}_{nn}\,\mathbf{Y}\,\mathbf{S}_{mm}^{\mathrm{T}}\right\|_{\mathrm{F}}^{2} + \lambda_n\, \mathrm{T}_r\left(\mathbf{S}_{nn}\,\mathbf{M}_{nn}\,\mathbf{S}_{nn}^{\mathrm{T}}\right) + \lambda_m\, \mathrm{T}_r\left(\mathbf{S}_{mm}\,\mathbf{M}_{mm}\,\mathbf{S}_{mm}^{\mathrm{T}}\right), \qquad (2)$$
$$\text{s. t. } \Omega_{nm}(\mathbf{Y}) = v_{nm}$$

over the optimization set $\Gamma = \{\mathbf{S}_{nn} \in \mathrm{SL}_n, \mathbf{S}_{mm} \in \mathrm{SL}_m, \mathbf{Y} \in \square^{n\times m}\}$ and for some regularization parameters λ_n, λ_m. The optimization problem given in Eq. 2 is subject to a vector constraint v_{nm}. If the known entries of **X** (or **Y** since the known entries match) are at positions

$$\Omega_{nm} = \left((i_1, j_1), (i_2, j_2), \ldots, (i_k, j_k)\right) \qquad (3)$$

this constraint can be written as

$$\Omega_{nm}(\mathbf{Y}) = v_{nm} \qquad (4)$$

where $v_{nm} \in \square^{k}$ is some fixed vector consisting of only known then fixed entries. If it happens that the auxiliary matrices $\mathbf{M}_{nn}$ and $\mathbf{M}_{mm}$ are under-observed, the vector constraint may be imposed over these two matrices as a *pre-processing* step. The additional constraints may be written as:

$$\Omega_{nn}(\mathbf{M}_{nn}) = v_{nn} \ \& \ \Omega_{mm}(\mathbf{M}_{mm}) = v_{mm} \quad (5)$$

It is noteworthy that without loss of generality one may perform an extra normalization step with respect to the non-zero determinant of a given symmetric matrix in order for it to belong to the special linear group, SL. Additional constraint, such as positive-definite constraint, or other convex regularization term may also be integrated to the main objective function given in Eq. 2, should they be required.

CMMC was originally implemented in MATLAB for a binary interaction matrix in DTI predictions. To reduce the run time, we re-implemented the CMMC algorithm in C++ (GCC 9.2.0) using Armadillo[26] (v10.1.2) with a OpenBLAS backend and tested the algorithm with nonbinary interaction matrices.

Chemical-gene Interaction Dataset and Normalization

We downloaded chemical-gene interaction data from the Comparative Toxicogenomics Database (CTD v16202)[27]. Human-related interactions were extracted and preprocessed for this study. In total, the CTD reported 871,112 chemical-gene interactions documented in humans, classified into 132 categories. Since this study is focused on environmental exposure chemicals, any drugs listed in the CTD were excluded. Data from Therapeutic Target Database[28] (TTD) and DrugBank[29] were used to filter out any drugs from the human CTD dataset by mapping chemicals to PubChem Compound IDs[6] (CIDs) (Supplementary Figure 1). After the drug filtering, 4,864 chemicals and 22,606 genes with 201,375 known interactions from human studies were included for further data normalization, resulting in a sparsity level of 99.8% for the total chemical-gene matrix. For this study, we calculated an *a priori* interaction value for each known chemical-gene pair and normalized these values in two ways without considering the different interaction categories. The first normalization method (N1) uses the number of occurrences of a chemical-gene pair in the dataset. The second method (N2) uses the number of unique PubMed IDs[30] listed for a chemical-gene pair (same chemical-gene pairs could be reported in different publications). Since approximately 78% of the chemical-gene pairs had only one interaction record in the dataset, the resulting distribution was highly skewed towards zero. A log transformation was therefore applied to both N1 and N2 interaction values to reduce the skewness of their respective distributions. Thus, for a specific chemical-gene pair, there were two values to represent the interactions. Finally, interaction values were scaled to be in the range of 0 to 1 (Supplementary Table 1).

Chemical-chemical Similarities

Two widely used chemical fingerprints[31] were applied to calculate the chemical-chemical similarities, i.e., Morgan fingerprints and Topological Torsion fingerprints. The SMILES[32] (simplified molecular-input line-entry system) string was used for the aforementioned two fingerprint calculations. The CIRpy (v1.0.2) and pubchempy (v1.0.4) Python packages were used to extract the SMILES strings for all chemicals in our dataset from the PubChem database. Chemical-chemical similarities were calculated using the RDKit Python package[33] (v2020.09.1).

Gene-gene Similarities

There are many ways to calculate gene-gene similarities. Pair-wise sequence alignment algorithms are widely used for calculating gene similarities based on either DNA/RNA sequences or protein sequences of genes. However, these methods cannot precisely reflect the biological similarities among genes. Thus, instead of using a sequence alignment-based method for calculating the gene-gene similarities, we used a

Gene Ontology[34] (GO) based method called GOSemSim[35] to measure the gene-gene similarities. Since there are three domains of GO annotation, i.e., cellular component (CC), molecular function (MF), and biological process (BP), three gene similarity matrices were generated for this study.

Generation of Benchmark Datasets

To comprehensively test CMMC on environmental chemical-gene interaction datasets, we generated three different sizes of input matrices with corresponding coupled matrices. The three different sizes are 200 chemicals by 400 genes (200×400), 400 chemicals by 600 genes (400×600), and 600 chemicals by 800 genes (600×800). To make our nonbinary interaction datasets comparable to previously published methods used for binary interaction predictions, we manually created two gold standard datasets to use as true positive and true negative sets for performance testing. As described above, for each chemical-gene pair, two different interaction values, N1 and N2, were calculated. For the true positive dataset, only the most confident chemical-gene pairs, i.e., those that have more than 20 interaction records (not unique PubMed IDs) were selected, which resulted in 462 chemical-gene pairs in the true positive dataset. To generate the true negative dataset, chemicals were sorted by the number genes they interact with to select for the most well-studied chemicals. Then, the top 20 chemicals with the greatest number of interacting genes were selected as chemicals for the negative dataset. In this manner, we generated a chemical-gene pair list with the 20 chemicals and the genes having no interaction evidence with these 20 chemicals as a true negative dataset.

To perform robust performance testing, for all three different sizes of input matrices, 10 different sets of randomly selected chemical-gene pairs from the positive and negative datasets were used to generate 10 different input interaction matrices separately for a 10-fold cross validation. The number of chemical-gene pairs for each fold took up 0.25% of the total size of the input interaction matrix. For instance, 0.25% of a 200×400 interaction matrix equals 200. Thus, 100 chemical-gene pairs were randomly selected from both the positive and negative data sets separately to form a total 200 chemical-gene pairs for one-fold. Since there are only 462 chemical-gene pairs in the positive dataset, all of these interactions were used in each of the ten folds of the 600×800 datasets. However, the 600 chemical-gene pairs in the true negative set were different in each fold for testing. Both types of normalized interaction values (i.e., N1 and N2) were used to generate the benchmark datasets, separately.

Other State-of-the-Art Methods

Similar matrix completion methods are broadly used in the related problem of DTI predictions. Thus, we compared our method to other widely used matrix completion methods to test the performance with respect to the area under the ROC curve (AUC). These methods are CMF[36] (Collaborative Matrix Factorization), GRMF[37] (Graph-Regularized Matrix Factorization), and WGRMF[37,38] (Weighted Graph-Regularized Matrix Factorization). These methods were selected for performance comparison because they also require supporting information to impute missing values in the interaction matrix. WKNKN[37] (Weighted K Nearest Knows Neighbors) is a preprocessing method that can be used to estimate the interaction likelihoods for missing values in an interaction matrix. With and without WKNKN, our method was compared to other six methods, i.e., CMF, CMF+WKNKN, GRMF, GRMF+WKNKN, WGRMF, and WGRMF+WKNKN. The same testing datasets were used for performance comparison testing. Since the aforementioned three methods were developed for imputing missing values in a binary interaction matrix, our nonbinary interaction matrices were converted to binary matrices for these methods by applying three cutoff values separately, i.e., 0.05, 0.1, and 0.2. Any interaction values larger than the cutoff were replaced by 1 while those smaller than the cutoff were replaced by 0. Then, the area under the ROC curve (AUC) values were calculated for performance comparison.

Generation of Case-study Datasets

The ultimate goal of this study is to predict potential gene targets of novel chemicals without any prior knowledge of biological effects. Thus, to test our method on a real-life scenario, we generated a series of datasets for BPA, and three of its structurally related replacements: BPB, BPF, and BPS. To perform a comprehensive case study, we generated 200 different datasets, each with 299 randomly selected chemicals (300 chemicals in total when including the case chemical) and 500 genes from a total 5000 chemicals and 10,000 genes for each case-study chemical, respectively. Since there are 4,864 chemicals after filtering drugs from the CTD dataset, chemicals from US EPA's Toxic Substances Control Act Chemical Substance Inventory (TSCA Inventory) were used to form a chemical list with 5000 chemicals for this case study. To mimic a real-life scenario with missing information, rather than only use a small handful of genes, we used a substantial portion of all genes (10,000) randomly selected from the total 22,606 genes without specifically including all interactive genes of these four chemicals. The complete gene list and chemical list can be found in Supplementary Table 2. In this way, interactions between each case-study chemical and 10,000 genes were imputed 10 times, and the resulting prediction averaged. In all 200 chemical-gene interaction matrices for each case-study chemical, all known interaction values were removed to mimic a real-life scenario. Bioactivity data of BPA, BPB, BPF, and BPS were downloaded from the CompTox Chemical Dashboard for validating our prediction results. The DAVID[39] (Database for Annotation, Visualization and Integrated Discovery) web service was used to identify related GO terms using the genes predicted from CMMC and genes from CTD of the four chemicals in case study, respectively. GO term sets containing more than 500 genes were filtered out for further analysis. An FDR less than 0.05 was used for selecting significant GO terms. The R Bioconductor package org.Hs.eg.db[40] (v3.14.0) was used for corresponding GO annotations.

Results

Overview of the CMMC Method

Matrix completion is an ideal approach for our goal due to the great deal of information already accumulated and available in databases about the genes affected by chemical exposures. CMMC, in particular, is attractive because it not only takes advantage of the known chemical-gene interactions, but also the hidden feature relationships among chemicals and among genes. CMMC was originally developed for predicting DTI with a binary input. As opposed to other coupled matrix completion methods, which use a fixed distance metric, such as Euclidean, CMMC uses DML in order to learn the metric from data itself. The metric is in fact *Mahalanobis* distance for a normalized set of data with mean zero[20]. Thus, while other algorithms make use of coupled matrices, main advantage of CMMC is the theoretical guarantee a unique minimum using the reductive group properties. Learning matrices $\mathbf{S}_{nn}$ an $\mathbf{S}_{mm}$, as given in Eq (2) helps update the coupled matrices with each iteration rather than them being static. In this study, we re-implemented and applied the CMMC algorithm to predict exposome target interactions with continuous input. The inputs of this method are three matrices: matrix M_{EG} represents the chemical-gene interactions, and the two symmetric matrices M_{EE} and M_{GG} represent the chemical similarities and gene similarities, respectively (Figure 1). Given the objective function in Eq. (2), M_{EG} corresponds to the matrix, **Y,** which itself is formed by known values only of initial matrix of gene-chemical interaction. Square symmetric matrices M_{EE} and M_{GG} correspond to $\mathbf{M}_{nn}$ and $\mathbf{M}_{mm}$, respectively. The goal is to learn matrices $\mathbf{S}_{nn}$ and $\mathbf{S}_{mm}$, using DML, through which the missing values of M_{EG} are predicted. It should be noted that M_{EG} is a sparse, partially observed matrix. In M_{EG}, each row represents an environmental chemical while each column represents a gene. The *a priori* interaction values were calculated based on data from the CTD (see *methods* section). M_{EE} and M_{GG} can be thought of as regularization matrices, and two regularization hyperparameters (λ_X and λ_Y as λ_n and λ_m in Eq (2)) are

applied to scale and add weights to the diagonal for preprocessing M_{EE} and M_{GG}, respectively. The preprocessed coupled matrices are M'_{EE} and M'_{GG} (Figure 1). Observed items (i.e., known chemical-gene interactions) in M_{EG} are used as constraints when minimizing the objective function. Two output matrices (B_1 and B_2) are initialized as identity matrices that are the same size as the two coupled input matrices. The preprocessed coupled matrices are directly used in the main loop to find an optimal solution. The comprehensive mathematical description of CMMC can be found in Bagherian M, et al[20].

Datasets Used in This Study

As described above, the data used in this study is from the Comparative Toxicogenomics Database (CTD), a comprehensive, well-maintained publicly available database containing a vast amount of information pertaining to how environmental exposures affects human health, including chemical-gene/protein interactions. Human-related chemical-gene interaction pairs were normalized and adjusted in two ways, first based on the number of occurrences of each chemical-target pair (N1) and the second method is based on the number of unique PubMed IDs listed for each chemical-target pair (N2) (Supplementary Table 1). Two different measurements of chemical similarities and three measurements of gene similarities were calculated and used in this study. For gene similarities, values were calculated from Gene Ontology (GO) cellular components (CC) and molecular functions (MF), which have similar distributions, and from biological processes (BP), whose distribution differed significantly from that of the other two similarity measures. Many gene pairs were assigned a similarity value of 1 based on GOCC and GOMF, while GOBP gene similarities were skewed towards to zero with a peak around 0.2 (Supplementary Figure 2A-C and Supplementary Table 3). Chemical similarity values calculated based on Morgan fingerprints are highly skewed towards zero and there is a clear peak around 0.1, while chemical similarity values calculated using Topological Torsion fingerprints were also skewed towards zero, but with a data peak around 0.02 with a small kurtosis (Supplementary Figure 2D, E and Supplementary Table 3).

To test our method and compare its performance with other state-of-the-art methods used for DTI prediction, we generated three different sizes of input datasets, 200×400, 400×600, and 600×800, and visualized the distribution of gene-gene and chemical-chemical similarities (Figure 2 shows the 200×400 distribution, the others are shown in Supplementary Figure 3). The interaction values from the normalized and adjusted values based on the number of occurrences of each exposure-target pair (N1) are illustrated in Figure 2D. The distributions of the other datasets are similar. Chemicals and genes were randomly selected to generate these three different datasets (see *Methods*). The sparsity of the interaction matrix (M_{EG}) is very high, i.e., 93.44% for the 200×400 dataset, 97.11% for the 400×600 dataset, and 98.23% for the 600×800 dataset (Figure 2D and Supplementary Figure 3). All coupled matrices were also sparse matrices (all over 50% sparsity).

Implementation of CMMC to Chemical Exposure-Gene Target Prediction

To optimize the two regularization hyperparameters for CMMC on chemical exposure-target prediction, the benchmark datasets were used for hyperparameter tuning (see *Methods and Materials*). The previous application[20] of CMMC for DTI prediction (with binary interaction matrix) suggested smaller parameters perform better. A small test on the 200×400 dataset found that the performance of CMMC was not greatly impacted by different iteration numbers (i.e., 1, 2, and 3) with hyperparameters larger than 1 (data not shown). Thus, a grid search with five values (i.e., 0.05, 0.1, 0.3, 0.5, and 0.9) was used to find an optimal hyperparameter pair for the following analysis in this study. To reduce the run time of CMMC on parameter selection, the iteration number was set to 1 for all calculations in this section. Because we generated two different types of chemical similarities, three types of gene similarities, and two different

types of chemical-gene interaction values for this study, there were a total of 12 different scenarios for which we tuned the hyperparameters. In addition, for each of the three sizes of input matrix and corresponding coupled matrices, 10 different testing datasets were generated for a 10-fold cross validation (see *Methods*). Thus, for each size of input dataset with one pair of hyperparameters, AUCs were calculated and averaged from 120 runs (2×3×2×10). The average AUC values increased with large parameters in all three sizes of input data (Supplementary Figure 4). Based on these results, 0.9 was selected for both hyperparameters and used in the following analyses.

After selecting the optimal hyperparameters, we sought to obtain an optimal iteration number for CMMC on exposome target predictions. When applying this algorithm to a binary DTI matrix, this algorithm converged in less than 5 iterations. In this study, the benchmark dataset was used to search for an optimal iteration number from 1 to 6. The average AUC values were calculated using the same procedure described in the previous section. Here, the hyperparameters were set to 0.9. We found that different iteration numbers had minimal impact on the average AUCs (Supplementary Figure 5). Thus, to guarantee a fast calculation, 2 iterations were used in the following analyses.

Performance Comparison

CMMC was compared with CMF[36], GRMF[37], and WGRMF[37,38] with and without WKNKN[37] using the same benchmark datasets for performance testing (see *Other State-of-the-Art Methods*). The two hyperparameters were set to 0.9 and the iteration number was set to 2 in CMMC for this performance comparison. For all other methods, parameters were selected as follows: $k = 100$, $\lambda_l = 2$, $\lambda_d = 0.125$, $\lambda_t = 0.125$, and $K = 5$ if WKNKN is used, where k is the number of latent features, λ_l, λ_d, and λ_t are regularization parameters, and K is the number of nearest neighbors. These parameters were optimized through a test run on our benchmark datasets (200×400) using the built-in function of the GRMF package.

The average AUCs calculated based on the N1 interaction values shows CMMC achieved the highest AUCs for all three sizes of input datasets with a maximum average AUC of 0.893 ± 0.012 using the 400×600 dataset (Figure 3A). However, the AUC values for the different sized matrices were not significantly different. In the following results, 0.1 was used as the cutoff for converting the interaction matrix into a binary matrix. The average AUCs across different combinations of coupled matrices with the 10 different testing datasets for all three sizes of input matrix improved from CMF to GRMF to WGRMF. WKNKN further improved the AUC values for these three methods (Figure 3A and 3B). Results with other binary conversion cutoffs confirmed that CMMC maintained its superior performance in all cases tested (Table 1). It also shows that a small binary conversion cutoff has better performances for CMF, GRMF, and WGRMF. Surprisingly, all other methods besides CMMC showed a consistent decrease in performance moving from smaller to larger input matrices, whereas CMMC demonstrated the most consistent performance. It is possible this result could be explained by the higher sparsity level in the larger matrices. Comparing the results between N1 and N2 for CMMC, N1 resulted in better performance across all three different sizes of input data, with average AUCs near 0.89 compared to 0.84 for N2. Based on these results, the following case studies were performed using CMMC with N1.

Case study

We selected the very well-studied endocrine disrupting chemical BPA and its closely related but lesser studied analogs BPB, BPF, and BPS, for our case study. These were chosen due to the fact that BPB, BPF, and BPS are often used as alternatives for BPA in manufacturing now that BPA has been banned in many products and in some countries[41]. However, given the similarity in BPA's chemical structure compared to the analogs and based on preliminary toxicology and epidemiology studies in the analogs, negative health effects could also be caused by exposure to the alternatives. For each of the four chemicals, we generated

200 different datasets containing randomly selected 299 chemicals and 500 genes from the total 5000 chemicals and 10,000 genes used for this study, respectively. All previously known interactions of the case study chemical in the 200 different datasets were removed from the main matrix (M_{EG}) to mimic a realistic scenario where one does not have any prior bioactivity knowledge of a new chemical. After imputing all 200 datasets for a chemical, the average imputed interaction values across all six runs (two chemical-chemical similarities by three gene-gene similarities) and 200 datasets were calculated and used for further analysis. Bioactivity data for BPA, BPB, BPF, and BPS from CompTox Chemical Dashboard and the corresponding GO terms from CTD were used to validate the predicted genes for each chemical (Supplementary Table 4 and 5). Genes not included in this study were removed from further analysis. A large, imputed interaction value in our results indicates a strong, high-confident interaction. Thus, for each chemical, genes were ordered by the imputed interaction values in descending order. Genes with active bioactivity responses ("active genes") of the four chemicals from the CompTox Chemical Dashboard were compared to our results (Figure 4). For BPA, 20 active genes from CompTox were predicted in the top 100 interactive genes in our results. While for BPB, BPF, and BPS, there are 23, 10, and 7 active genes predicted in the top 100 interactive genes, respectively (Figure 4A). Among these genes, four genes were identified as common target genes across these four chemicals, i.e., *ESR1*, *ESR2*, *AR*, and *NR1I2*. Furthermore, around 50% of the active genes of BPA and BPB and 80% active genes of BPF and BPS are predicted in the top 1,000 genes, respectively (Supplementary Table 4). In the CompTox Chemical Dashboard, seven genes were identified as active genes for all four of these chemicals (*ESR1*, *ESR2*, *AR*, *NR1I2*, *PGR*, *PPARG*, and *PPARD*). Among these genes, all six of them used in our study, i.e., *ESR1*, *ESR2*, *AR*, *NR1I2*, *PGR*, *PPARG*, were predicted to be in the top 1,000 interacting genes for all four chemicals in our results (Figure 4B). Since the 10,000 gene list was randomly selected, PPARD was not included in the total gene list.

To further investigate our results, the top 1,000 interacting genes of each tested chemical were used to identify the related GO terms through gene set enrichment testing. After removing GO terms containing more than 500 genes and filtering out GO terms with an FDR larger than 0.05, we found 401, 394, 343, and 396 significant GO terms for BPA, BPB, BPF, and BPS, respectively (Supplementary Table 5). Strikingly, *Response to Estradiol* (GO:0032355), *Positive regulation of ERK1 and ERK2 cascade* (GO:0070374), and *Cellular response to tumor necrosis factor* (GO:0071356) were ranked in the top 5 significant GO terms for all four chemicals. Since BPA is an endocrine-disrupting chemical[42], hormone related GO terms were extracted from the significant GO term lists (Figure 5A). Most hormone-related GO terms were significantly enriched for all four chemicals, including *Response to Estradiol* (GO: 0032355), *Response to Estrogen* (GO: 0043627), *Estrogen Metabolic Process* (GO: 0008210), and *Estrogen Receptor Binding* (GO:0030331). However, *Insulin receptor signaling pathway* (GO:0008286) and *Estrogen 16-Alpha-Hydroxylase Activity* (GO:0101020) were only significantly enriched for BPB while *Insulin-like growth factor I binding* (GO:0031994) and *Positive Regulation Of Insulin−Like Growth Factor Receptor Signaling Pathway* (GO:0043568) were only significantly enriched for BPA.

For comparison, genes known to interact with these four chemicals were downloaded from CTD and gene set enrichment testing was used to find the significant hormone-related GO terms. Because there are more than 5,000 interacting genes in CTD for BPA and BPS, only the top 1,000 interacting genes were used. For BPB and BPF, all genes in CTD (115 genes for BPB and 144 genes for BPF) were used for gene set enrichment testing. For BPA, 54 hormone related GO terms were significantly enriched from the CTD genes while 23 hormone related GO terms were significantly enriched from genes predicted by CMMC (Supplementary Figure 6). 87% (20/23) of these hormone related GO terms from CMMC results were also enriched in GO terms from CTD genes. This number is 39% (9/23), 65% (13/20), and 59% (13/22) for BPB, BPF, and BPS, respectively (Figure 5B and Supplementary Figure 7-9). Among the significantly

enriched GO terms using genes from CTD and CMMC, six of them were enriched for all four chemicals (Figure 5B).

We visualized the distribution of gene similarities with each of the four chemicals, both for the top 1,000 genes and non-top 1,000 genes of CompTox interacting genes (Supplementary Figure 10). For BPA and BPB, the gene similarities show clear differences in GOCC. For BPF, all three types of gene similarities are different between the top 1,000 genes and non-top 1,000 genes. Because there were only two non-top 1,000 genes for BPS, the non-top 1,000 gene is not shown. This is consistent with our expectation that the supporting information can be utilized by our method to predict potential target genes of a novel chemical.

Runtime and Use of CMMC

Since our implementation of CMMC is in C++ and the other methods are implemented in MATLAB, we felt it would be unfair to compare the run times. However, the elapsed real time of CMMC (excluding any data preparation steps) on the benchmark datasets were very reasonable (Supplementary Figure 11). The average run time for a 200×400 interaction matrix is 3.4 ± 1.7 seconds, for a 400×600 input matrix it is 5.8 ± 3.0 seconds, and for a 600×800 input matrix it is 10.1 ± 5.4 seconds. We executed CMMC on a Red Hat Enterprise Linux Server (release 7.9) with an Intel(R) Xeon(R) Gold 6150 CPU (2.70GHz). No parallel programming was used in the implementation.

Discussion

Data fusion by representing data from multiple sources as a collection of matrices is indeed an effective way to jointly analyze the data together with data in the context or additional constraints. Matrix completion, as an approach in data fusion, is a prediction method designed for finding missing values from a partially observed matrix, where the input is usually a highly sparse matrix. In general, missing values are imputed by using the low-rank matrix of the known values. However, it is extremely difficult to predict the missing values of a new row or column item when there is no relevant prior knowledge. Therefore, matrix completion methods that incorporate supporting information in the form of coupled matrices represent a significant step forward, and have gained traction by showing improved performance in recent years[43,44].

In this study, we created and tested a new implementation of the recently developed CMMC for predicting environmental chemical-gene interactions. The inputs of this algorithm are three matrices, i.e., the main matrix represents the interactions between chemicals and genes, and two coupled symmetric matrices represent the similarities among chemicals and genes, respectively. CMMC was previously applied to predict DTI with a binary interaction matrix, but it was not designed to be limited to binary matrices. In this study, we applied CMMC with a nonbinary interaction a matrix. The environmental chemical-gene pairs used in this study were downloaded from the CTD, and two approaches were used to calculate *a priori* interaction values. Our results show that the total number of interaction values achieved better performance than the number of unique publications reporting an interaction. However, in future studies additional normalization approaches should be tested due to the sensitivity of the performance to the specific pre-processing and normalization used.

We compared the performance of our CMMC implementation for predicting environmental chemical-gene interactions, with three related, published methods (i.e., CMF, GRMF, and WGRMF). Additionally, WKNKN, a preprocessing method for estimating the interaction likelihoods of unknown interactions based on known neighbors, was utilized for the three aforementioned methods in the performance comparison. We chose these three methods because they also permit the use of two coupled supporting matrices as input for prediction. These three methods were originally developed for predicting DTI with a

binary interaction matrix, so we used three different cutoffs (i.e., 0.05, 0.1, and 0.2) to convert the interaction values into binary values for input for other methods. Our original plan was to use three additional cutoffs: 0.3, 0.4, and 0.5. However, because the interaction data is highly skewed towards the left, using 0.3, 0.4, or 0.5 as the binary conversion cutoff, resulted in too few "known neighbors" for WKNKN to estimate the likelihood of unknown interactions which caused an internal error. Thus, we only used the three cutoffs for performance comparisons. The performance of these three methods gradually improved from CMF to WGRMF. Furthermore, WKNKN additionally improved the performance of these three methods. This result is consistent with the previous study that CMF obtained the worst performance[37]. Considering the fact that our nonbinary interaction values are highly skewed towards the left, it is not surprising to see improved performance with smaller binary conversion cutoffs.

BPA and three of its structurally related replacements were selected to perform a case study, namely BPB, BPF, and BPS. We chose BPA because it is a well-studied chemical with abundant bioactivity records for validating our results. BPB, BPF, and BPS are chemicals with similar structures and are used as substitutes of BPA in industrial production. These four chemicals have similar chemical structures which are two phenol groups connected through a linking group. Considering the chemical similarities among these four chemicals (Supplementary Figure 12), we expected similar but not identical prediction results. More than half of the interacting genes of these four chemicals from the CompTox database are identified in our results as top interacting genes. Since CompTox is a data source independent of CTD, our results are independently validated. All known interaction values were removed in the case study for each chemical in the imputation, so this result suggests that our method can be used to effectively find potential target genes for a novel chemical. BPA belongs to the endocrine disrupting class of chemicals, and low dose BPA can impact reproduction, the immune system, and is related to hormone-dependent cancers[42]. Previous studies have shown that some nuclear receptors are target genes of BPA, such as androgen receptor[45] (*AR*), estrogen receptor 1 (*ESR1*), estrogen receptor 2[46] (*ESR2*), and aryl hydrocarbon receptor[47] (*AHR*). In our results, all of the aforementioned four genes were predicted as top 100 interactive genes for BPA. Except for *AHR*, other three genes were also predicted by CMMC as top interactive genes for BPB, BPF, and BPS (Supplementary Table 4). This indicates that the three substitutes chemicals may have similar adverse biological effects as BPA[48]. Furthermore, estrogen and other hormone-related GO terms identified from our gene lists of these four chemicals are very similar to each other and are consistent with the fact that these chemicals are endocrine disruptors and are able to mimic estrogen[49]. The relationship between BPA and type 2 diabetes has been well studied in the last decade[50,51]. A recent longitudinal study provide new evidence that not only the BPA but also BPS are associated with type 2 diabetes[52,53]. In our results, several insulin related GO terms are significantly enriched from our gene list of all four chemicals. Thus, the CMMC results prioritize the biological effects of BPF and BPB on type 2 diabetes for future examination.

The supporting information in the two coupled similarity matrices is critical for any prediction method to find chemical-gene interactions for a novel chemical without any prior knowledge. The input information is very flexible, thus it will be interesting to test alternative similarity metrics, such as gene-gene sequence similarities instead of or in combination with gene-gene similarity based on GO term membership. Using this coupled matrix information with an optimized rather than fixed similarity measure represents a significant strength of CMMC. Furthermore, CMMC is not a black box method. By comparing the similarities of the chemicals being predicted with other known chemicals, the similarities of the known predicted genes for the most similar chemicals with the genes predicted for the new chemicals, and the similarities among the top-ranked genes, we can gain direct insight into the causes for each prediction. For example, G protein-coupled estrogen receptor 1 (*GPER1*) was only predicted in the top 100 target genes of BPB by CMMC in our case study. This is consistent with a previous study that BPB and BPAF

(Bisphenol AF) have an approximate 9-fold higher binding affinity to *GPER1* than BPA[54]. BPAF is an analogue of BPA in which the two methyl groups are replaced with trifluoromethyl groups. Although all of these five chemicals are very similar, BPB and BPAF are more similar to each other in which adding/replacing chemical groups to the linking group of BPA. This can partially explain why only BPB among the four chemicals was predicted to target *GPER1*, that is, it learned from the similarity with BPAF. This information is very useful for post-hoc analyses and follow-up studies and would not be so easily determined with a deep learning approach.

This study has multiple limitations. Currently, we did not differentiate the chemical-gene interaction categories that are provided in the CTD. Although most interactions are labeled as changes in RNA expression changes (52.6%), there are 132 different interaction categories listed. Appropriately incorporating chemical-gene interaction categories into the prediction could be advantageous for future studies. Additionally, only human data were used in this study. Future research is required to apply our method on other species or tissue-specific data for more precise predictions. Identifying the target genes of a novel chemical is key but is just one step in understanding the potential health effects from an environmental exposure. Knowing the target tissues, considering the dosage level, and linking the molecular response to diseases are all downstream steps that are not considered in this paper. However, CMMC combined with our data normalization strategy has great potential to form a prioritization pipeline for discovering potential molecular targets of novel chemicals without any priori bioactivity knowledge. Future studies could potentially incorporate tissue information, dose information, and/or interaction category information as an additional dimension in the interaction matrix to form a partially observed tensor, thus expanding the CMMC to CMTC (coupled matrix-tensor completion) for missing value imputation.

Acknowledgments

We thank Dr. Kayvan Najarian and Dr. Renaid Kim for their professional help of this study, and we also thank the three anonymous reviewers for reviewing this manuscript. Support for this research was provided by NIH/NIEHS grants P30ES017885, R35ES031686, and R01ES028802.

Data Sharing

All data and code are publicly available at https://github.com/sartorlab/CMMC-on-Exposome-Prediction.

References

1. Collins FS, Morgan M, Patrinos A. The Human Genome Project: lessons from large-scale biology. *Science*. 2003;300(5617):286-290.

2. Vermeulen R, Schymanski EL, Barabási AL, Miller GW. The exposome and health: Where chemistry meets biology. *Science*. 2020;367(6476):392-396.

3. Wild CP. The exposome: from concept to utility. *International journal of epidemiology*. 2012;41(1):24-32.

4. Murray CJ, Lopez AD, Mathers CD, Stein C. The Global Burden of Disease 2000 project: aims, methods and data sources. *Geneva: World Health Organization*. 2001;36:1-57.

5. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research*. 2009;37(suppl_2):W623-W633.

6. Kim S, Chen J, Cheng T, et al. PubChem in 2021: new data content and improved web interfaces. *Nucleic acids research*. 2021;49(D1):D1388-D1395.

7. Kavlock RJ, Bahadori T, Barton-Maclaren TS, Gwinn MR, Rasenberg M, Thomas RS. Accelerating the pace of chemical risk assessment. *Chemical research in toxicology*. 2018;31(5):287-290.

8. Van Norman GA. Limitations of animal studies for predicting toxicity in clinical trials: is it time to rethink our current approach? *JACC: Basic to Translational Science*. 2019;4(7):845-854.

9. Jain AK, Singh D, Dubey K, Maurya R, Mittal S, Pandey AK. Models and methods for in vitro toxicity. In: *In Vitro Toxicology*. Elsevier; 2018:45-65.

10. Blaauboer BJ. The contribution of in vitro toxicity data in hazard and risk assessment: current limitations and future perspectives. *Toxicology letters*. 2008;180(2):81-84.

11. Zhang Q, Li J, Middleton A, Bhattacharya S, Conolly RB. Bridging the data gap from in vitro toxicity testing to chemical safety assessment through computational modeling. *Frontiers in public health*. 2018;6:261.

12. Lipnick RL. Structure-activity relationships. In: *Fundamentals of Aquatic Toxicology*. CRC Press; 2020:609-655.

13. Tong W, Welsh WJ, Shi L, Fang H, Perkins R. Structure-activity relationship approaches and applications. *Environmental Toxicology and Chemistry: An International Journal*. 2003;22(8):1680-1695.

14. Wang NCY, Zhao QJ, Wesselkamper SC, Lambert JC, Petersen D, Hess-Wilson JK. Application of computational toxicological approaches in human health risk assessment. I. A tiered surrogate approach. *Regulatory Toxicology and Pharmacology*. 2012;63(1):10-19.

15. Patlewicz G, Ball N, Becker RA, et al. Food for thought: read-across approaches–misconceptions, promises and challenges ahead. *Alternatives to Animal Experimentation: ALTEX*. 2014;31(4):387-396.

16. Cherkasov A, Muratov EN, Fourches D, et al. QSAR modeling: where have you been? Where are you going to? *Journal of medicinal chemistry*. 2014;57(12):4977-5010.

17. Muratov EN, Bajorath J, Sheridan RP, et al. QSAR without borders. *Chemical Society Reviews*. 2020;49(11):3525-3564.

18. Neves BJ, Braga RC, Melo-Filho CC, Moreira-Filho JT, Muratov EN, Andrade CH. QSAR-based virtual screening: advances and applications in drug discovery. *Frontiers in pharmacology*. 2018;9:1275.

19. Bagherian M, Sabeti E, Wang K, Sartor MA, Nikolovska-Coleska Z, Najarian K. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. *Briefings in bioinformatics*. 2021;22(1):247-269.

20. Bagherian M, Kim RB, Jiang C, Sartor MA, Derksen H, Najarian K. Coupled matrix–matrix and coupled tensor–matrix completion methods for predicting drug–target interactions. *Briefings in bioinformatics*. 2021;22(2):2161-2171.

21. Žitnik M, Zupan B. Data fusion by matrix factorization. *IEEE transactions on pattern analysis and machine intelligence*. 2014;37(1):41-53.

22. Pavlidis P, Weston J, Cai J, Noble WS. Learning gene functional classifications from multiple data types. *Journal of computational biology*. 2002;9(2):401-411.

23. Williams AJ, Grulke CM, Edwards J, et al. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *Journal of cheminformatics*. 2017;9(1):1-27.

24. Suárez JL, García S, Herrera F. A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges. *Neurocomputing*. 2021;425:300-322.

25. Jantzen JC. *Representations of Algebraic Groups*. Vol 107. American Mathematical Soc.; 2003.

26. Sanderson C, Curtin R. Armadillo: a template-based C++ library for linear algebra. *Journal of Open Source Software*. 2016;1(2):26.

27. Davis AP, Grondin CJ, Johnson RJ, et al. Comparative toxicogenomics database (CTD): update 2021. *Nucleic acids research*. 2021;49(D1):D1138-D1143.

28. Wang Y, Zhang S, Li F, et al. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic acids research*. 2020;48(D1):D1031-D1041.

29. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*. 2018;46(D1):D1074-D1082.

30. Canese K, Weis S. PubMed: the bibliographic database. *The NCBI handbook*. 2013;2(1).

31. Pattanaik L, Coley CW. Molecular representation: going long on fingerprints. *Chem*. 2020;6(6):1204-1207.

32. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*. 1988;28(1):31-36.

33. Landrum G. *RDKit: A Software Suite for Cheminformatics, Computational Chemistry, and Predictive Modeling*. Academic Press; 2013.

34. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. *Nature genetics*. 2000;25(1):25-29.

35. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*. 2010;26(7):976-978.

36. Zheng X, Ding H, Mamitsuka H, Zhu S. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ; 2013:1025-1033.

37. Ezzat A, Zhao P, Wu M, Li XL, Kwoh CK. Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM transactions on computational biology and bioinformatics*. 2016;14(3):646-656.

38. Guan NN, Zhao Y, Wang CC, Li JQ, Chen X, Piao X. Anticancer drug response prediction in cell lines using weighted graph regularized matrix factorization. *Molecular therapy-nucleic acids*. 2019;17:164-174.

39. Dennis G, Sherman BT, Hosack DA, et al. DAVID: database for annotation, visualization, and integrated discovery. *Genome biology*. 2003;4(9):1-11.

40. Carlson M, Falcon S, Pages H, Li N. org. Hs. eg. db: Genome wide annotation for Human. *R package version*. 2019;3(2):3.

41. Kadasala NR, Narayanan B, Liu Y. International trade regulations on BPA: Global health and economic implications. *Asian Development Policy Review*. 2016;4(4):134-142.

42. Cimmino I, Fiory F, Perruolo G, et al. Potential mechanisms of bisphenol A (BPA) contributing to human disease. *International journal of molecular sciences*. 2020;21(16):5761.

43. Chen Z, Wang S. A review on matrix completion for recommender systems. *Knowledge and Information Systems*. Published online 2022:1-34.

44. Nguyen LT, Kim J, Shim B. Low-rank matrix completion: A contemporary survey. *IEEE Access*. 2019;7:94215-94237.

45. Teng C, Goodwin B, Shockley K, et al. Bisphenol A affects androgen receptor function via multiple mechanisms. *Chemico-biological interactions*. 2013;203(3):556-564.

46. Fischer C, Mamillapalli R, Goetz TG, Jorgenson E, Ilagan Y, Taylor HS. Bisphenol A (BPA) exposure in utero leads to immunoregulatory cytokine dysregulation in the mouse mammary gland: a potential mechanism programming breast cancer risk. *Discover Oncology*. 2016;7(4):241-251.

47. Donini CF, El Helou M, Wierinckx A, et al. Long-term exposure of early-transformed human mammary cells to low doses of benzo [a] pyrene and/or bisphenol A enhances their cancerous phenotype via an AhR/GPR30 interplay. *Frontiers in oncology*. 2020;10:712.

48. Park C, Song H, Choi J, et al. The mixture effects of bisphenol derivatives on estrogen receptor and androgen receptor. *Environmental Pollution*. 2020;260:114036.

49. Moon MK. Concern about the safety of bisphenol A substitutes. *Diabetes & metabolism journal*. 2019;43(1):46-48.

50. Alonso-Magdalena P, Quesada I, Nadal A. Endocrine disruptors in the etiology of type 2 diabetes mellitus. *Nature Reviews Endocrinology*. 2011;7(6):346-353.

51. Duan Y, Yao Y, Wang B, et al. Association of urinary concentrations of bisphenols with type 2 diabetes mellitus: a case-control study. *Environmental pollution*. 2018;243:1719-1726.

52. Seltenrich N. Bisphenol Exposure and Type 2 Diabetes: New Evidence for a Potential Risk Factor. *Environmental Health Perspectives*. 2020;128(7):074006.

53. Rancière F, Botton J, Slama R, et al. Exposure to bisphenol A and bisphenol S and incident type 2 diabetes: a case–cohort study in the French cohort DESIR. *Environmental health perspectives*. 2019;127(10):107013.

54. Cao LY, Ren XM, Li CH, et al. Bisphenol AF and bisphenol B exert higher estrogenic effects than bisphenol A via G protein-coupled estrogen receptor pathway. *Environmental Science & Technology*. 2017;51(19):11423-11430.

Figure 1. Illustration of the coupled matrix-matrix completion algorithm. Three input matrices are M_{EE}, M_{GG}, and M_{EG}. M'_{EE} and M'_{GG} are the preprocessed coupled matrices with the two hyperparameters.

Figure 2. Data distributions for each input matrix for the 200×400 dataset. The sparsity of each matrix is labeled at the right top corner of each histogram. (A-C) gene-gene similarity matrices based on Gene Ontology, (D) the *a priori* known environmental chemical-gene interactions, and (E-F) chemical-chemical similarity matrices based on Morgan and Topological fingerprints, respectively.

Figure 3. Performance comparison of CMMC with alternative methods. (A) Average AUCs of the methods using the N1 interaction values (see Methods). Binary conversion cutoff is 0.1 for CMF, GRMF, and WGRMF. (B) ROC curves of different datasets (with and without WKNKN for the alternative methods).

Figure 4. Results validation with CompTox bioactivity data. (A) Percentage of CompTox interacting genes detected by CMMC. (B) Venn diagram of the top 1000 genes corresponding to (A). All six of the seven interactive genes for all four chemicals from CompTox that were used in the case study were also detected as interactive genes in our results.

Figure 5. (A) Hormone-related significantly enriched GO terms from gene set enrichment test (Top 1000 interactive genes from the CMMC results used as input gene list). Arrows point to the GO terms that differ by chemical. (B) Overlap between GO terms detected by genes from CMMC prediction and genes directly from the CTD.

Table 1. Performance comparisons on benchmark datasets with different binary conversion cutoffs.

Supplementary Figure 1. Illustration of workflow for drug filtering. TTD and DrugBank were used to remove drugs included in the CTD dataset through PubChem IDs.

Supplementary Figure 2. Distributions of overall chemical and gene similarities. (A) Gene similarities calculated based on GOMF. (B) Gene similarities calculated based on GOCC. (C) Gene similarities calculated based on GOBP. (D) Chemical similarities calculated based on Topological Torsion Fingerprints. (E) Chemicals similarities calculated based on Morgan Fingerprints. Results based on 22,606 genes and 4,864 chemicals.

Supplementary Figure 3. Distributions of values in matrices used in the benchmark datasets. The sparsity of each matrix is labeled at the top right corner of each histogram.

Supplementary Figure 4. Grid search results of hyperparameter tunning. Average AUCs across datasets are illustrated by the bar graphs.

Supplementary Figure 5. The average AUCs of different iteration numbers. 0.9 was used as the hyperparameters for these calculations.

Supplementary Figure 6. Hormone-related significantly enriched GO terms from gene set enrichment test for BPA. Top 1000 interactive genes from our result and top 1000 interactive genes from CTD were used as input gene lists.

Supplementary Figure 7. Hormone-related significantly enriched GO terms from gene set enrichment test for BPB. Top 1000 interactive genes from our result and all 115 interactive genes from CTD were used as input gene lists.

Supplementary Figure 8. Hormone-related significantly enriched GO terms from gene set enrichment test for BPF. Top 1000 interactive genes from our result and all 144 interactive genes from CTD were used as input gene lists.

Supplementary Figure 9. Hormone-related significantly enriched GO terms from gene set enrichment test for BPS. Top 1000 interactive genes from our result and top 1000 interactive genes from CTD were used as input gene lists.

Supplementary Figure 10. Gene similarities comparison between top 1000 and non-top 1000 genes of CompTox interacting genes. A moderate shift towards higher similarities is observed for the top 1000 genes, as expected. Only two non-top 1000 genes for BPS and the corresponding gene similarities are: GOCC: 0.449, GOBP: 0.264, and GOMF: 0.516.

Supplementary Figure 11. Run time of CMMC in seconds with different sizes of input data.

Supplementary Figure 12. Chemical similarities of BPA, BPB, BPF, and BPS. (A) Chemical similarities based on Morgan Fingerprints. (B) Chemical similarities based on Topological torsion Fingerprints.

Supplementary Table 1. Normalized and adjusted chemical-gene interactions.

Supplementary Table 2. Total gene list and chemical list used in the case study.

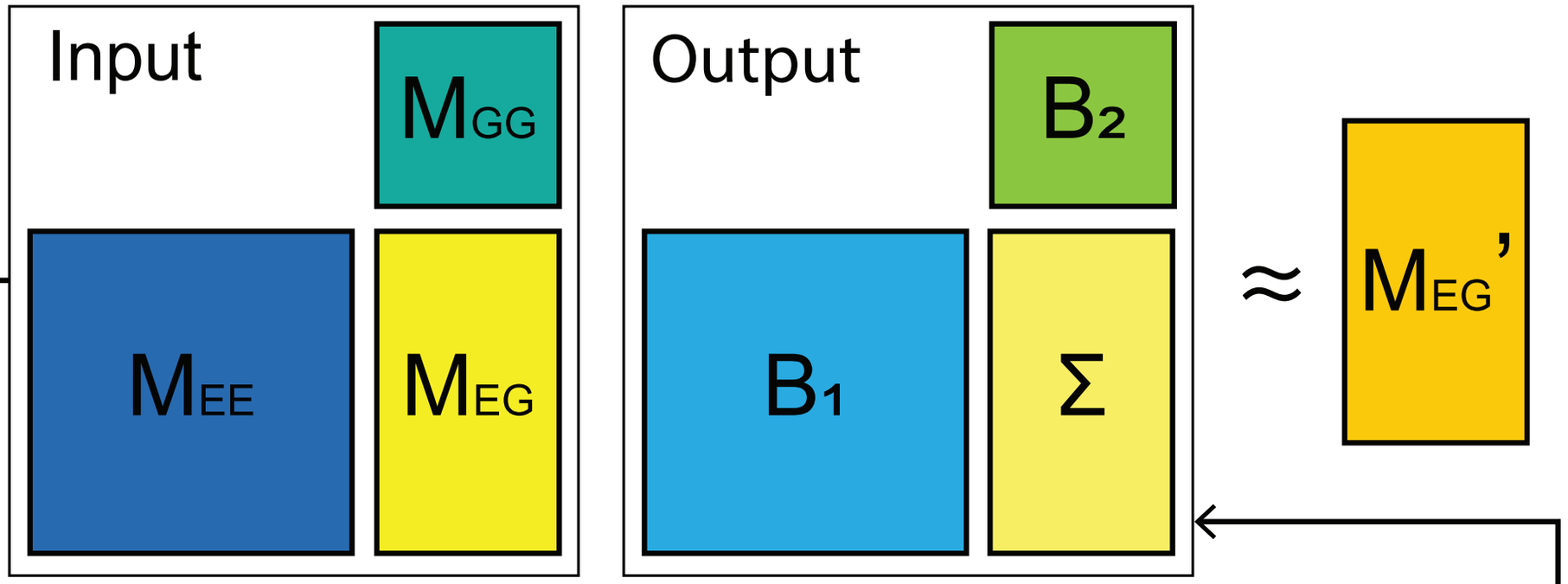
Supplementary Table 3. Data summary of gene-gene similarities and chemical-chemical similarities.

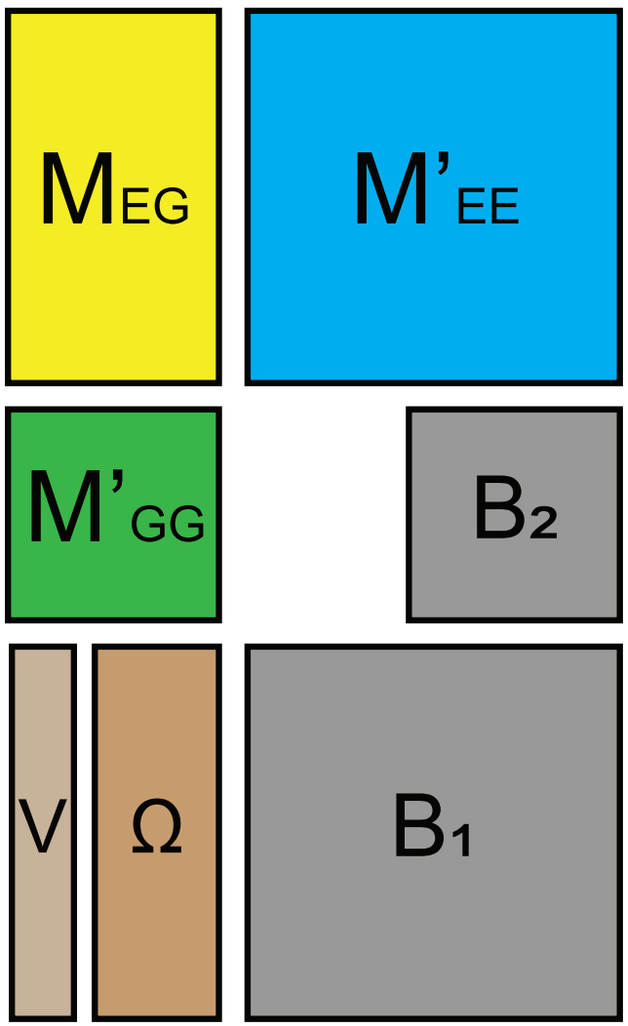
Supplementary Table 4. Imputation results of CMMC for CompTox interactive genes. Separate tabs are used for BPA, BPB, BPF, and BPS.

Supplementary Table 5. Significantly enriched GO terms of BPA, BPB, BPF, and BPS. Input gene lists were the top 1000 ranked interactive genes from CMMC.

Table 1. Performance comparisons on benchmark datasets with different binary conversion cutoffs.

			WKNKN(-)			WKNKN(+)			Binary Conversion Cutoff
		CMMC	CMF	GRMF	WGRMF	CMF	GRMF	WGRMF	
			0.832±0.025	0.844±0.018	0.845±0.021	0.843±0.025	0.854±0.019	0.856±0.021	0.05
	200x400	**0.885±0.020**	0.814±0.026	0.823±0.027	0.828±0.025	0.831±0.028	0.846±0.027	0.848±0.026	0.1
			0.806±0.036	0.796±0.035	0.800±0.036	0.816±0.035	0.833±0.034	0.834±0.036	0.2
			0.824±0.014	0.836±0.017	0.843±0.014	0.842±0.014	0.849±0.016	0.856±0.014	0.05
N1	400x600	**0.893±0.012**	0.817±0.011	0.806±0.016	0.816±0.013	0.834±0.009	0.833±0.012	0.839±0.011	0.1
			0.803±0.014	0.766±0.020	0.778±0.015	0.818±0.013	0.812±0.016	0.818±0.015	0.2
			0.761±0.006	0.806±0.015	0.820±0.005	0.778±0.005	0.818±0.012	0.828±0.005	0.05
	600x800	**0.884±0.004**	0.718±0.006	0.781±0.023	0.803±0.007	0.781±0.005	0.803±0.013	0.814±0.006	0.1
			0.599±0.006	0.660±0.020	0.679±0.004	0.732±0.006	0.746±0.015	0.760±0.006	0.2
			0.815±0.036	0.824±0.031	0.827±0.031	0.824±0.034	0.842±0.031	**0.845±0.032**	0.05
	200x400	0.842±0.028	0.802±0.035	0.807±0.023	0.814±0.023	0.822±0.036	0.836±0.033	0.840±0.032	0.1
			0.785±0.040	0.775±0.026	0.779±0.024	0.799±0.039	0.816±0.032	0.818±0.032	0.2
			0.798±0.018	0.809±0.019	0.815±0.018	0.818±0.014	0.825±0.015	0.832±0.013	0.05
N2	400x600	**0.838±0.012**	0.792±0.016	0.781±0.016	0.793±0.010	0.815±0.014	0.818±0.013	0.826±0.009	0.1
			0.762±0.014	0.753±0.023	0.767±0.016	0.785±0.013	0.784±0.013	0.791±0.011	0.2
			0.738±0.006	0.777±0.012	0.787±0.004	0.768±0.006	0.791±0.012	0.801±0.004	0.05
	600x800	**0.819±0.004**	0.672±0.005	0.749±0.021	0.769±0.005	0.773±0.006	0.783±0.015	0.796±0.004	0.1
			0.588±0.008	0.680±0.026	0.705±0.005	0.678±0.006	0.722±0.015	0.736±0.005	0.2





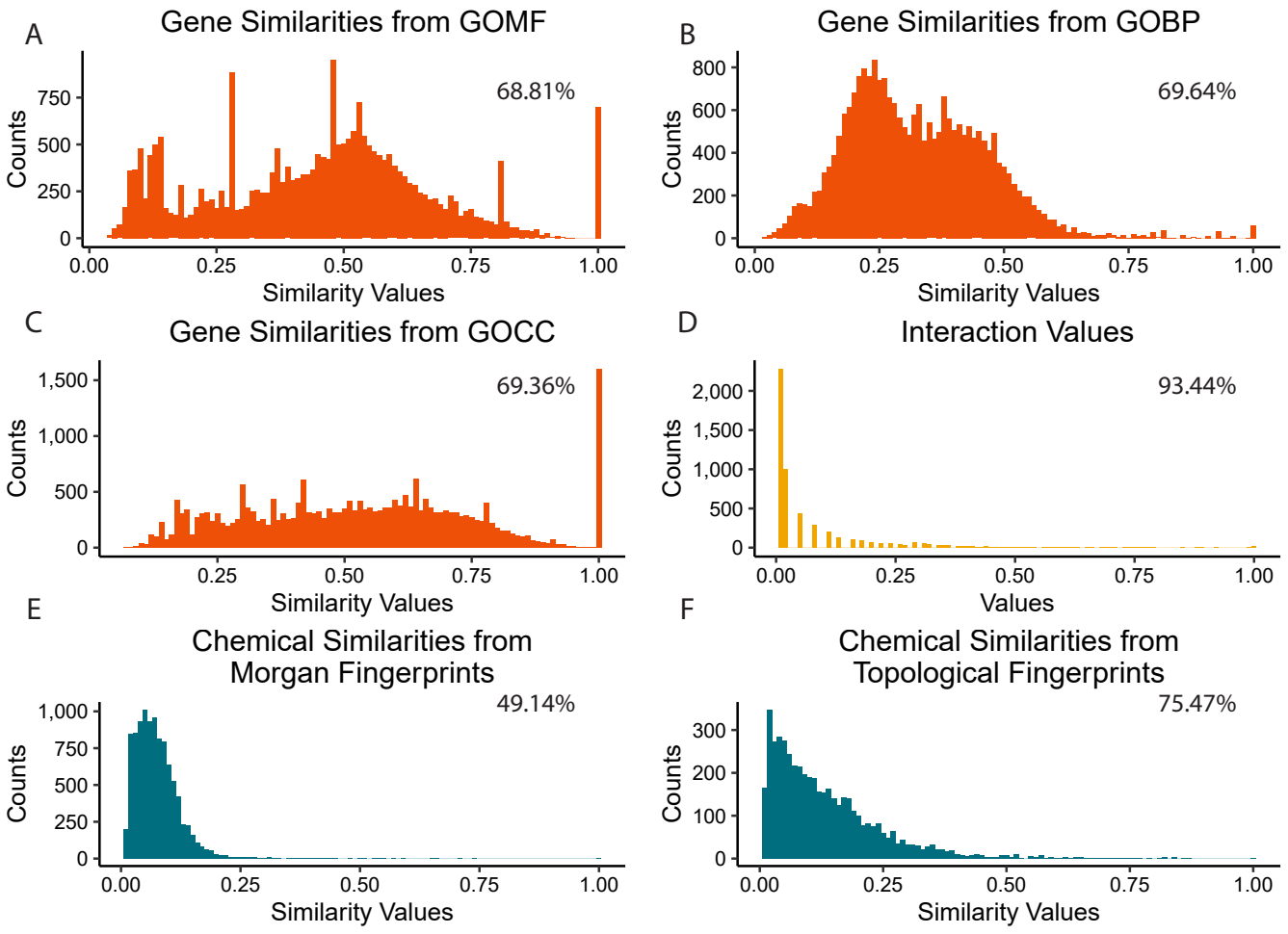
Main Loop

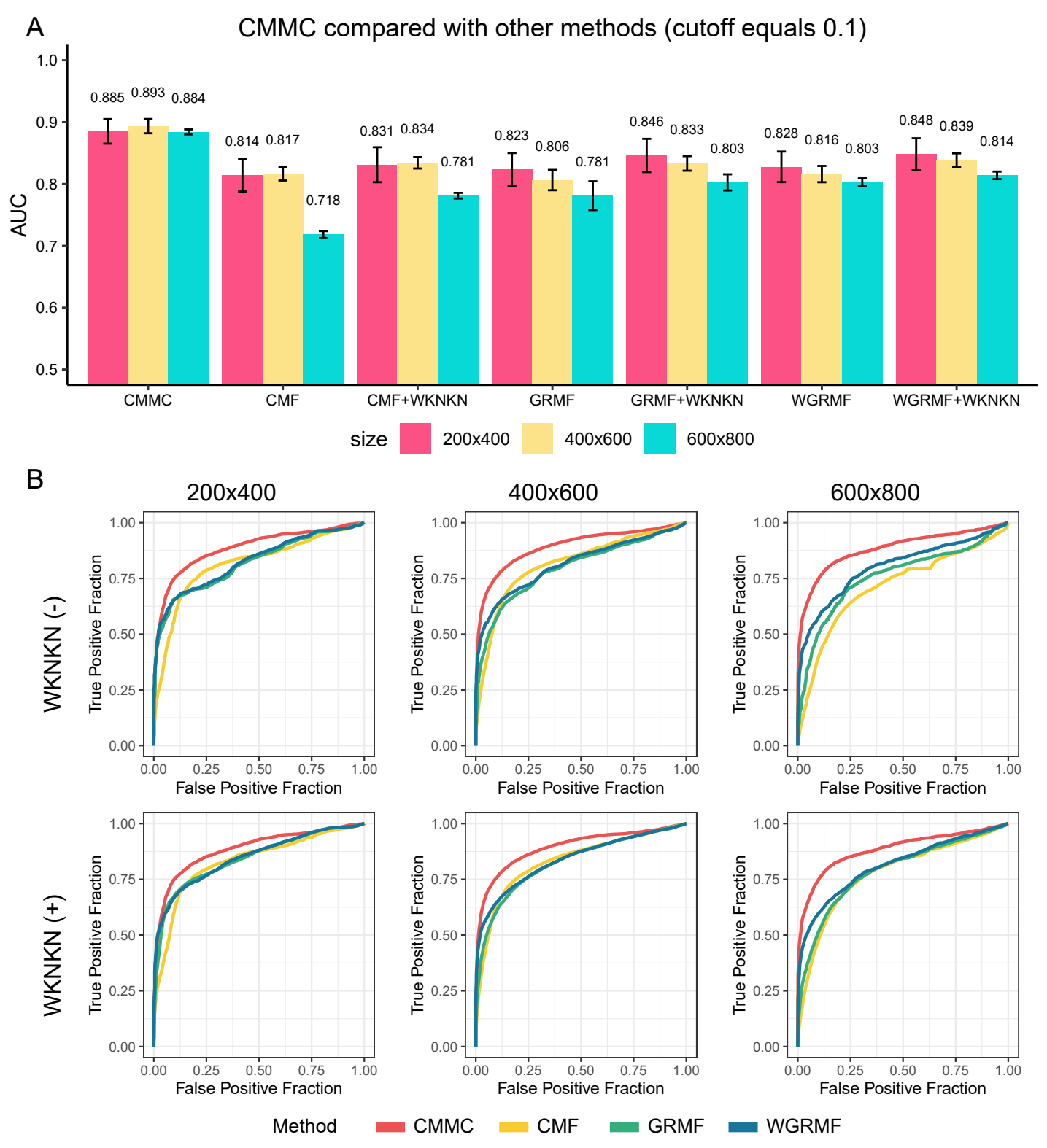
for *i* in *1, 2, ..., iter*:
  choose Σ that minimizes $\|B_1 \Sigma B_2\|_F^2$
  with constraint Ω(Σ) = V

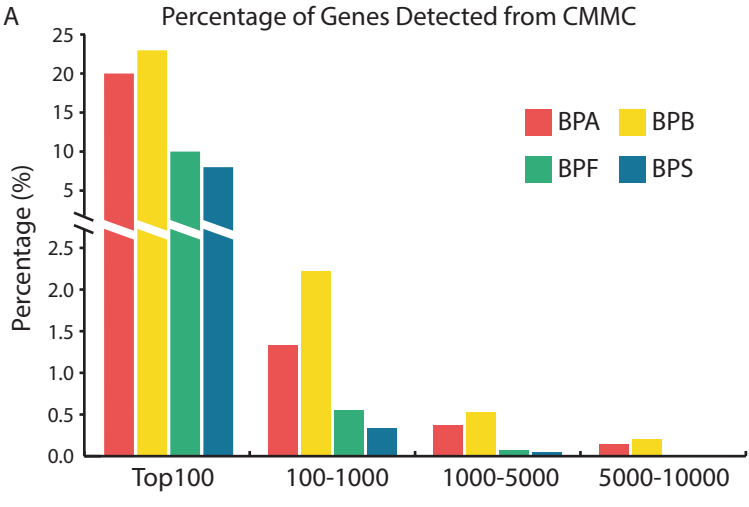
$$B_1 = (M'_{EE} + B_1 M_E B_1)^{-1/2}$$
$$B_1 = det(B_1)^{-1/En} B_1$$
$$B_2 = (M'_{GG} + B_2 M_G B_2)^{-1/2}$$
$$B_2 = det(B_2)^{-1/Gn} B_2$$

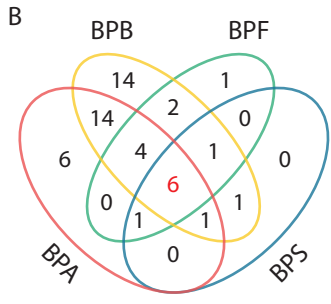
end

V is the vector of observed items in M_{EG}; Ω is the position index of V; B_1 and B_2 are initialized as two identity matrices; M_E and M_G are generated from Ω, B_1, and B_2.


A
Gene Similarities from GOMF
68.81%
Counts
Similarity Values
B
Gene Similarities from GOBP
69.64%
Counts
Similarity Values
C
Gene Similarities from GOCC
69.36%
Counts
Similarity Values
D
Interaction Values
93.44%
Counts
Values
E
Chemical Similarities from Morgan Fingerprints
49.14%
Counts
Similarity Values
F
Chemical Similarities from Topological Fingerprints
75.47%
Counts
Similarity Values


A
CMMC compared with other methods (cutoff equals 0.1)
AUC
1.0
0.9
0.8
0.7
0.6
0.5
0.885
0.893
0.884
0.814
0.817
0.718
0.831
0.834
0.781
0.823
0.806
0.781
0.846
0.833
0.803
0.828
0.816
0.803
0.848
0.839
0.814
CMMC
CMF
CMF+WKNKN
GRMF
GRMF+WKNKN
WGRMF
WGRMF+WKNKN
size
200x400
400x600
600x800
B
200x400
400x600
600x800
WKNKN (-)
WKNKN (+)
True Positive Fraction
False Positive Fraction
0.00
0.25
0.50
0.75
1.00
Method
CMMC
CMF
GRMF
WGRMF





ESR1, *ESR2*, *AR*, *NR1I2*
PGR, *PPARG*, *PPARD*

Six of the seven interactive genes for all four chemicals from CompTox used in the case study were detected as interactive genes by CMMC.

A

Enriched GO Terms

Cellular Response To Estradiol Stimulus
Cellular Response To Insulin−Like Growth Factor Stimulus
Cellular Response To Insulin Stimulus
Cellular Response To Peptide Hormone Stimulus
Estrogen 16−Alpha−Hydroxylase Activity
Estrogen 2−Hydroxylase Activity
Estrogen Metabolic Process
Estrogen Receptor Binding
Insulin−Like Growth Factor I Binding
Insulin−Like Growth Factor Receptor Signaling Pathway
Insulin Receptor Binding
Insulin Receptor Signaling Pathway
Intracellular Steroid Hormone Receptor Signaling Pathway
Positive Regulation Of Insulin−Like Growth Factor Receptor Signaling Pathway
Positive Regulation Of Insulin Receptor Signaling Pathway
Regulation Of Insulin Secretion
Response To Estradiol
Response To Estrogen
Response To Insulin
Response To Peptide Hormone
Steroid Binding
Steroid Hormone Receptor Activity
Steroid Hydroxylase Activity
Steroid Metabolic Process
Thyroid Gland Development

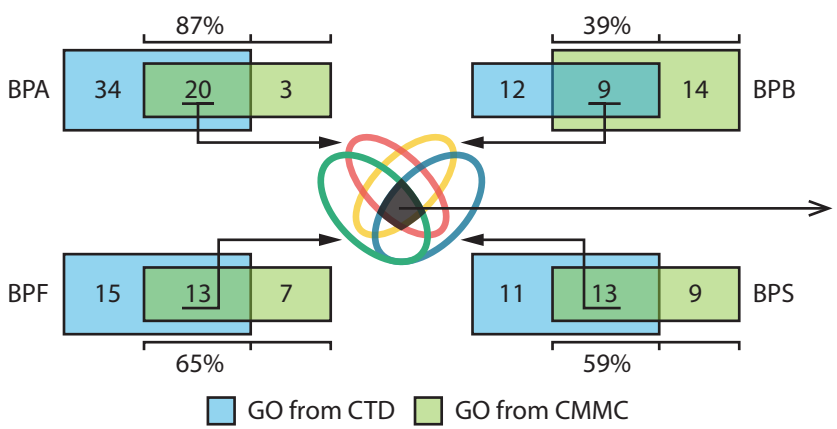
BPA BPB BPF BPS

# of Genes
100
200
300
400

−log10(PValue)
16
12
8
4

Biological Process
Molecular function

B



- Steroid binding
- Response to estrogen
- Response to estradiol
- Steroid hormone receptor activity
- Cellular response to insulin stimulus
- Cellular response to estradiol stimulus
- Intracellular steroid hormone receptor signaling pathway