COMPUTING THERAPY FOR PRECISION MEDICINE: COLLABORATIVE FILTERING INTEGRATES AND PREDICTS MULTI-ENTITY INTERACTIONS

SAM REGENBOGEN
Department of Pharmacology, Baylor College of Medicine
Houston, TX 77030, USA
Email: regenbog@bcm.edu

ANGELA D. WILKINS
Department of Molecular and Human Genetics, Baylor College of Medicine
Houston, TX 77030, USA
Email: aw11@bcm.edu

OLIVIER LICHTARGE
Department of Molecular and Human Genetics, Baylor College of Medicine
Houston, TX 77030, USA
Email: lichtarg@bcm.edu

Biomedicine produces copious information it cannot fully exploit. Specifically, there is considerable need to integrate knowledge from disparate studies to discover connections across domains. Here, we used a Collaborative Filtering approach, inspired by online recommendation algorithms, in which non-negative matrix factorization (NMF) predicts interactions among chemicals, genes, and diseases only from pairwise information about their interactions. Our approach, applied to matrices derived from the Comparative Toxicogenomics Database, successfully recovered Chemical-Disease, Chemical-Gene, and Disease-Gene networks in 10-fold cross-validation experiments. Additionally, we could predict each of these interaction matrices from the other two. Integrating all three CTD interaction matrices with NMF led to good predictions of STRING, an independent, external network of protein-protein interactions. Finally, this approach could integrate the CTD and STRING interaction data to improve Chemical-Gene cross-validation performance significantly, and, in a time-stamped study, it predicted information added to CTD after a given date, using only data prior to that date. We conclude that collaborative filtering can integrate information across multiple types of biological entities, and that as a first step towards precision medicine it can compute drug repurposing hypotheses.

1. Introduction

At the same time as advances in biomedical research have enabled humanity's knowledge to grow far beyond the limits of any one person, that knowledge is being applied on ever-smaller scales. Specialized therapies are benefiting smaller subsets of the population, using all available knowledge to design a therapy for a specific case or to repurpose an existing drug for a novel use.

Online databases that compile this knowledge have become invaluable resources for researchers. Massive interaction networks can be powerful sources for hypothesizing novel relationships between biological entities. However, most of these networks are either focused on one particular type of entity (STRING[1] – genes/proteins) or interaction (DrugBank[2], ChEMBL[3] – drug-gene interactions). A full representation of biomedical knowledge would integrate the interactions among these physical entities and associate them with more abstract entities, such as pathways (KEGG[4], REACTOME[5,6]) and diseases (CTD[7]).

Several approaches to data integration have been explored. One approach is to predict how two classes of entity interact (e.g., drugs and targets) by integrating multiple types of feature data about the entities[8–10], or taking this a step farther, propagating this information to a third entity type[11]. These methods utilize information about the entities themselves, so they are specific to certain classes of entity. We will show an alternative approach, which can predict interactions among chemicals, genes, and diseases utilizing only information about how they connect to one another, and which benefits from the integration of disparate forms of information.

Collaborative filtering (CF) is a computational approach used in online recommendation systems, in which large-scale knowledge of how entities interact is used to predict likely connections[12,13]. Non-negative matrix factorization (NMF) is a popular tool for CF that compresses a matrix into two smaller factors whose product approximates the original[14,15]. NMF has long been used in biomedical science for clustering and classifying microarray data[16], but recent works have used NMF, or related algorithms, in CF strategies to predict drug-target[17,18] or protein-protein[19] interactions. We hypothesized that this basic approach could be pushed farther, to incorporate more than two types of biological entity, improving prediction of novel interactions among them.

Testing this hypothesis required multiple interaction networks, comprising connections between at least three entity types, so we turned to the Comparative Toxicogenomics Database (CTD). CTD is a publicly available resource that employs a team of human "biocurators" to comb the literature, extracting and annotating Chemical-Gene, Chemical-Disease, and Disease-Gene relationships[7]. In this paper, we will demonstrate that NMF can be used to recover hidden interactions in each of these networks individually and that NMF over any two of these networks can predict back the third. To show that this is not an artifact of the data source (CTD), we will demonstrate that NMF over the combined CTD networks recapitulates experimental protein-protein interactions in the STRING database. We will focus in on the CTD Chemical-Gene interaction network, and show that our ability to predict missing connections improves when we perform NMF over a network incorporating Chemical-Gene, Chemical-Disease, and Disease-Gene interactions from CTD and also Protein-Protein interactions from STRING.

2. Methods:

2.1. *Construction of datasets:*

Tables of interactions from CTD were obtained and processed as follows. Unless otherwise noted, all data processing and manipulation was performed in Matlab. Chemical-Gene and Chemical-Disease interactions were downloaded on April 2, 2014[a], each as a single tab-delimited text file. The full Chemical-Gene interactions file was imported into Matlab as a table containing 878,594 rows, each representing one unique curated relationship between one chemical and one gene, or between other relationships. This initial table comprised 10,520 unique chemicals and 32,248 unique genes. Relationships containing nested relationships were removed, as were any relationships whose "Gene Form" was not given as "protein" ("mRNA," for example.) The result of this filtering was a table of direct relationships involving 8,653 unique chemicals and 8,288 unique genes. A binary adjacency matrix was built in which each row and column corresponded to one chemical or gene, respectively, with interacting pairs assigned a value of 1, and all other pairings 0. The resulting sparse 8,653-by-8,288 matrix contains 82,168 unique, binary Chemical-Gene interactions.

The Chemical-Disease interactions file was similarly imported into Matlab, but was filtered to remove all CTD-inferred relationships by deleting any row for which the "Direct Evidence" column was blank. The filtered table was used to build a binary adjacency matrix as described above, which in this case comprised 8,226 chemicals, 3,031 diseases, and 80,433 unique, curated interactions.

The full Disease-Gene interactions file was too large to process in the same way, so CTD's Batch Query tool[b] was used to retrieve only the curated interactions. On April 18, 2014, the CTD Disease Vocabulary file was downloaded, and the Disease IDs were input to the Batch Query tool, which was set to export all Curated Gene Associations for each disease. The output tab-delimited interactions were then imported into Matlab and, as before, used to build a sparse, binary adjacency matrix of 4,907 Diseases by 7,362 Genes, with 23,133 unique interactions.

For construction of a combined Chemical-Gene-Disease (CGD) interaction matrix, the interaction tables used to build the individual matrices were used. A single list of 30,102 unique entities was obtained from the union of the three individual matrices' unique entity lists, comprising 12,119 Chemicals, 6,333 Diseases, and 11,650 Genes. Each of the three interaction tables was then used to populate a matrix in which each of the 30,102 entities was represented as both a row and a column. Thus, for each row in the three tables, the interacting entities' positions in the combined entity list defined two symmetrical pairs of indices in the 30,102-by-30,102 matrix at which to represent the interaction.

For later experiments, we used the STRING network of human protein-protein interactions[c], which we mapped to the CTD CGD matrix. When comparing our predictions to STRING, we focused on 7,604 genes whose IDs we could map between databases, and used the confidence scores

[a] from http://ctdbase.org/downloads - dates noted because CTD updates monthly; previous versions are unavailable
[b] http://ctdbase.org/tools/batchQuery.go
[c] STRING v9.1, now archived at http://string91.embl.de/newstring_cgi/show_download_page.pl

assigned by STRING (ranging from 0 to 999) to define the positive class at various thresholds. To construct a CTD+STRING CGD matrix, we added protein-protein interactions from this STRING[d] network to the Gene-Gene diagonal block of the CTD CGD matrix. Interactions among the 7,604 genes also in CTD were dropped directly into the corresponding cells in the CGD matrix symmetrically. The matrix was extended by 6,699 rows and columns, corresponding to the genes that were not matched to CTD. The final matrix contains 254,929 nonzero Gene-Gene interactions, 66,685 with values of 0.5 or greater, and 1,405 with the maximum value of 0.999.

2.2. *Non-negative Matrix Factorization (NMF):*

NMF describes several closely related algorithms that, given a non-negative matrix **A** with size $m \times n$ and a positive integer $k \ll \min(m,n)$, attempt to find $m \times k$ matrix **W** and $k \times n$ matrix **H** such that **W** and **H** are non-negative, and such that $\mathbf{A} \approx \mathbf{WH}$. This is done by solving the optimization problem

$$\min_{W,H \geq 0} f(\mathbf{W},\mathbf{H}) = \frac{1}{2} \|\mathbf{A} - \mathbf{WH}\|_{\mathrm{F}}^2 \quad (1)$$

Throughout this work, NMF was run using the *nnmf*() function of Matlab's Statistics Toolbox with all input arguments (other than **A** and k) left at default settings. Consequently, the optimization method used was Alternating Least Squares (ALS), in which initial **W** and **H** matrices are randomly generated, and then alternatingly solved for in the following matrix equations, until the minimization function converges or until the maximum number of iterations has been reached:

$$\text{Solve for } \mathbf{H}\text{: } \mathbf{W}^{\mathrm{T}}\mathbf{WH} = \mathbf{W}^{\mathrm{T}}\mathbf{A} \quad (2.1)$$

$$\text{Solve for } \mathbf{W}\text{: } \mathbf{HH}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}} = \mathbf{HA}^{\mathrm{T}} \quad (2.2)$$

In our applications of NMF to datasets of various sizes, we tested multiple k values for each, to find a value that would give optimal performance without overfitting.

NMF is known to converge at solutions that are local, rather than global, minima of the optimization problem, meaning the product **WH** is not unique. We found that calculating the average of **WH** across multiple replicate factorizations increased performance in our experiments; all results we discuss below were obtained by averaging the output of 4 NMF replicates[e].

2.3. *10-fold Cross-validation Experiments:*

In N-fold cross-validation experiments, each point in a dataset is randomly assigned to one of N subsets. Then, one at a time, every subset is removed, and the remaining N-1 subsets are used as training data for the algorithm to be tested. In the end, the algorithm's predicted values for each dropped subset form a test set covering all of the original data. An algorithm's ability to successfully recover data in cross-validation depends not only on the algorithm itself, but also on the internal consistency of the dataset. Entities with only 1 known interaction were not considered, because NMF would have no way to recover that interaction.

[d] inserted as the confidence score divided by 1000 to match the range of the rest of the CGD matrix, which is binary.
[e] Data not shown. We chose 4 replicates to balance diminishing returns in improvements v. computational cost.

2.4. *Performance evaluation for NMF predictions*

The performance of NMF in each experiment was evaluated by calculating the Receiver Operator Characteristic (ROC) curve, comparing predicted scores to an input positive class, and computing the number of correct predictions at varying score thresholds. An ROC curve can be understood as sorting the list of predictions by score and, beginning at the origin, moving up on the y-axis for each true prediction and moving right on the x-axis for each false prediction. The area under an ROC curve (AUC) can serve as a broad measure of performance, representing the probability that a randomly chosen positive (known) interaction will have been assigned a higher score by NMF than a randomly chosen negative (not known) interaction.

3. Results and Discussion

3.1. *10-fold cross-validation for NMF of individual CTD matrices.*

In order to determine whether a CF approach can integrate interactions between multiple classes of biological entity, we first made certain that NMF can be used to recover unknown pairwise interactions among Chemicals, Diseases, and Genes from incomplete interaction data. 10-fold cross-validation was performed on three adjacency matrices constructed from CTD's Chemical-Disease (CD), Chemical-Gene (CG), and Disease-Gene (DG) networks, respectively.

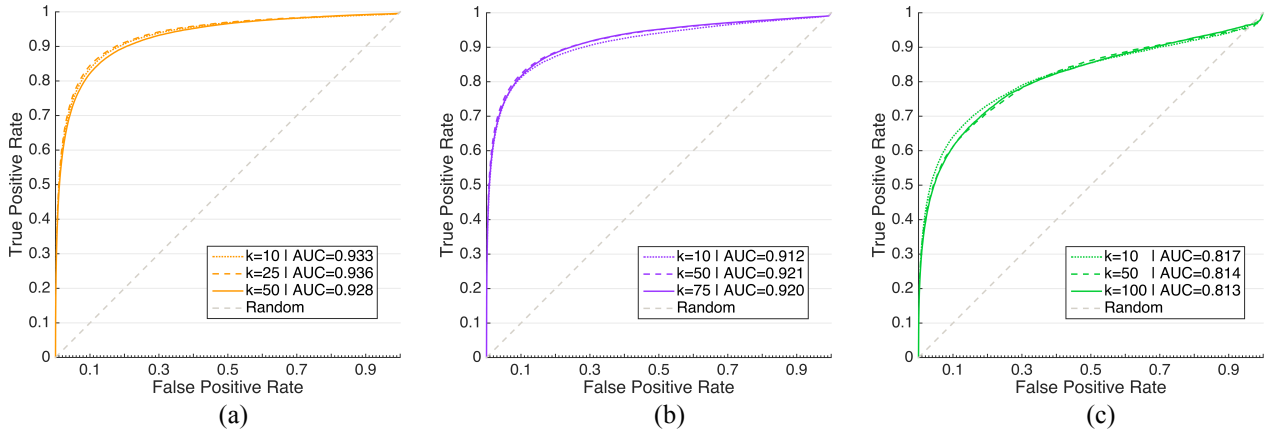

Fig. 1. Receiver Operator Characteristic (ROC) curves of NMF at varying k values in 10-fold cross-validation experiments over individual CTD interaction networks. (a) Chemical-Disease, (b) Chemical-Gene, (c) Disease-Gene

Figure 1 shows NMF performs much better than random guessing in 10-fold cross-validation for the three CTD networks, with performance plotted as Receiver Operator Characteristic (ROC) curves, with k varying over a range to find values that optimize AUC. The best results were AUC of 0.94 (CD), 0.92 (CG), and 0.82 (DG). The results in Figure 1 show these three networks are internally consistent enough to recover missing interactions using NMF, and that the interactions involving Chemicals (CD and CG) are particularly well-suited to prediction by NMF.

3.2. *CTD Chemical-Gene-Disease matrix and leave-one-matrix-out experiments*

Once we verified that the three networks from CTD were, individually, amenable to prediction of missing interactions via NMF, we considered how to utilize this multifaceted data more effectively. The data encompassed three classes – Chemicals, Diseases, and Genes – of biological entity, with information about each category spread across two matrices. When it factorizes an interaction matrix, NMF represents each entity (row/column vector) as a compressed vector that approximates all available information. Therefore, we reasoned that simply combining the asymmetric CD, CG, and DG matrices into one symmetric "all-vs-all" Chemical-Gene-Disease (CGD) matrix would allow NMF access to more information about the relationships between Chemicals, Diseases, and Genes, and thus improve our ability to predict missing ones.

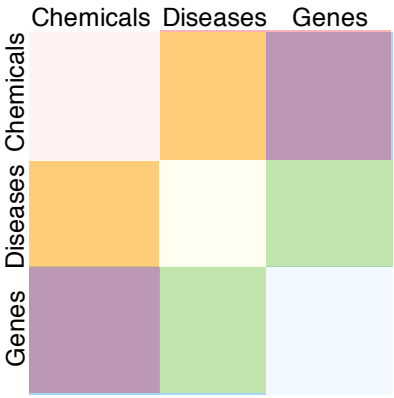

Fig. 2. Illustration of the combined, symmetric CGD matrix. The CTD CD, CG, and DG matrices are orange, purple, and green, respectively. The diagonal blocks are empty before factorization.

In order to test the ability of our CF approach to integrate different types of interaction, we devised a "leave-one-matrix-out" experiment (Fig. 3a-c). From the combined CGD matrix in Figure 2, we removed all interactions of one class (CD, CG, or DG) at a time, and attempted to predict them from only the other two interaction classes. We performed this test, using NMF with various k values, for each of the three interaction types and calculated ROC curves. Fig. 3d shows the AUC for each *k* value used to predict the missing matrices.

Table 1. Amount of data dropped and re-predicted in Leave-One-Matrix-Out Experiments, followed by AUC when NMF was performed over the remaining two interaction matrices. Column headings indicate which interaction matrix was left out.

Dropped Matrix	Chemical-Disease	Chemical-Gene	Gene-Disease
Size	4760x1605	4760x3940	3940x1605
# Interactions	59,766	60,831	15,522
AUC *k*=100	0.801	0.833	0.802
AUC *k*=200	0.810	0.840	0.801
AUC *k*=300	0.813	0.837	0.802
AUC *k*=500	0.817	0.832	0.795

These results show that NMF is able to predict the interactions contained in each of the matrices created from CTD's datasets, given only information contained in the other two matrices, despite the distinctly different biological connections they represent. Put another way, this demonstrates that combining these binary interaction matrices can unlock new layers of information that was not accessible from the individual matrices. Because all three networks share an origin in CTD's manual curation process, however, we need to determine that the latent information tapped by NMF for these predictions provides a meaningful insight to the workings of biology, and not just an insight into the CTD curation pipeline.

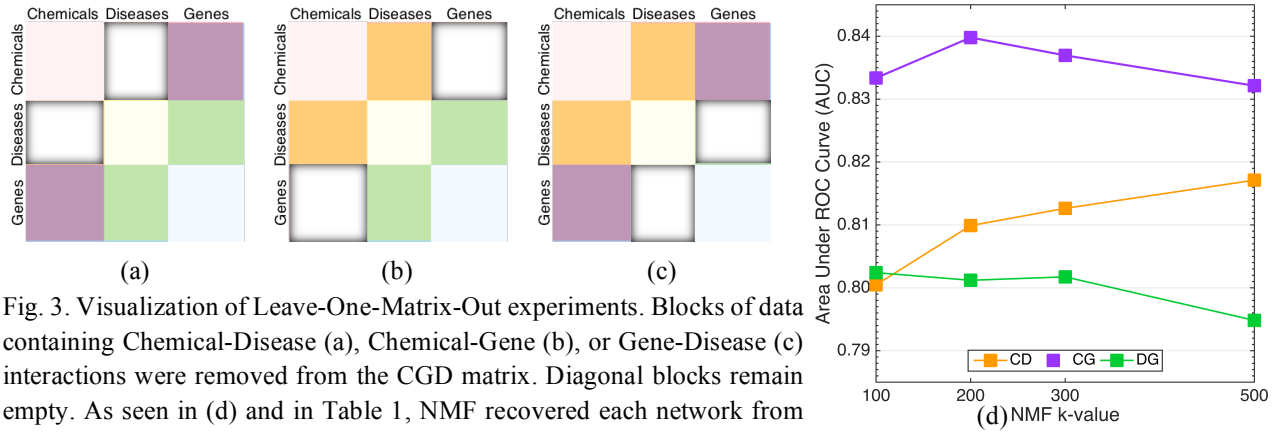

Fig. 3. Visualization of Leave-One-Matrix-Out experiments. Blocks of data containing Chemical-Disease (a), Chemical-Gene (b), or Gene-Disease (c) interactions were removed from the CGD matrix. Diagonal blocks remain empty. As seen in (d) and in Table 1, NMF recovered each network from the remaining two.

3.3. *Prediction of Gene-Gene associations from CTD Chemical-Gene-Disease matrix*

The diagonal blocks of the combined matrix, which would correspond to Chemical-Chemical, Disease-Disease, and Gene-Gene associations, contain no data initially from CTD, but are also filled in when we use NMF. We sought to compare predictions in these regions to an external data source, in order to find out if the values predicted by NMF represent real biological relationships.

Although it is unclear what the disease-disease network might represent, comparing the gene-gene block to existing protein interaction databases was a natural next step. We compared the values from NMF to known protein-protein interactions from the STRING database. 7,604 genes were present in both the combined CTD matrix and the STRING experimental network. Among these 7,604 genes, STRING contained 67,763 experimentally supported protein-protein interactions, of which 38,424 have been assigned confidence scores by STRING of at least 500, and 902 have been assigned the highest confidence score of 999.

As shown in Fig. 4., the values produced by NMF over the CTD CGD matrix predicted these interactions with an ROC AUC of 0.69, which increased to AUC=0.73 for interactions ≥500 confidence score, and to AUC=0.75 when only the highest-confidence STRING interactions (999) were considered.

These results show that the Gene-Gene associations filled into the Chemical-Gene-Disease matrix by NMF correspond to real, experimentally known protein-protein interactions. This result is important because, unlike the Leave-One-Matrix-Out experiments, these predicted edges were never part of CTD, reducing the chance that positive results are due to some inherent bias in the CTD curation process. This also

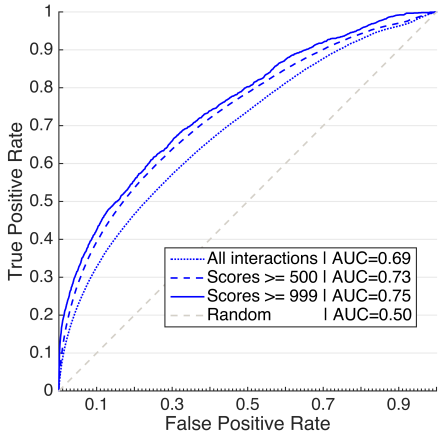

Fig. 4. ROC curves for the prediction of STRING protein-protein interactions using NMF (k=300) on CTD CGD.

suggests that the predictions in the Chemical-Chemical and Disease-Disease blocks may be biologically meaningful, potentially representing drug interactions and disease co-morbidity, for example. At the same time, these results suggest that some of the information contained within the STRING network was not found by NMF in the combined CGD matrix. We created a second Chemical-Gene-Disease matrix containing all the same interactions from CTD, but with protein-protein interactions from STRING added to the Gene-Gene block of the diagonal.

3.4. *10-fold cross-validation of Chemical-Gene edges within combined CGD matrix*

In order to determine if additional data can improve upon the prediction of Chemical-Gene interactions observed in Fig. 1b, we performed an experiment similar to 10-fold cross-validation, which only removed Chemical-Gene edges from the larger matrix. We performed this experiment using the CTD CGD matrix, and also using the CTD+STRING CGD matrix, both with k=200. We also repeated the 10-fold cross-validation using the CG matrix alone, using the best-performing k value, k=50. For this comparison, a single set of randomized cross-validation classes was generated first, and then was used for all three input matrices, to ensure that the only differences in available information were those we were testing.

Table 2. Comparison of NMF performance in 10-fold cross-validation of Chemical-Gene edges without added data, with the addition of CD and DG information from CTD, or with that plus GG information from STRING

Matrix	k	ROC AUC[f]	p-value[f] vs CTD_{CG}	p-value[f] vs CTD_{CGD}
CTD_{CG}	50	0.920	–	–
CTD_{CGD}	200	0.927	$4.6x10^{-109}$	–
$CTD_{CGD}+S_{GG}$	200	0.932	$4.9x10^{-244}$	$5.8x10^{-117}$

As shown in Table 2 and Figure 5a, the Chemical-Gene cross-validation performance after the addition of Chemical-Disease and Disease-Gene interactions yielded AUC=0.927, an increase over the highest-performing k value with Chemical-Gene interactions only (AUC=0.920 at k=50). Moreover, when Gene-Gene interactions from STRING were added to the CGD matrix, performance further improved to AUC=0.932. To measure this improved performance, we used the StAR method[20], which implements an approach based on Mann-Whitney U-statistics[21], to determine if the ROC curves were significantly different. Although the increases in AUC appear small, so many data points were used to calculate the ROC curves that they were found to be highly significant.

AUC of the ROC curve provides an overall indicator of how well a method recovers true interactions. However, practical applications (e.g., drug repurposing,) are likely to focus on relatively few predictions compared to the total interaction space. For this reason, it is often more important that the top predictions have high precision (i.e., few false positives). To be sure the CGD matrices were not only increasing AUC by improving recall of the low-confidence interactions, we calculated precision-recall curves for the cross-validation (Figure 5b). As the inset shows, the

[f] Output by StAR tool, standalone version[20], rounded for table

precision of the highest-scoring 10% of interactions[g] is high for all three test cases, with the CTD+STRING Chemical-Gene-Disease matrix showing precision improvements across the range.

These results show that we can improve the ability of NMF to predict missing Chemical-Gene relationships by incorporating information about how those Chemicals and Genes interact with Diseases, and, further, how those Genes interact with one another.

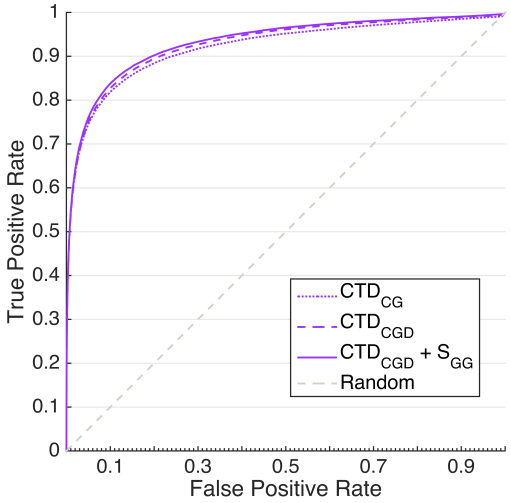

Fig. 6a. ROC curves showing NMF performance for 10-fold cross-validation of Chemical-Gene interactions, improving with more data. AUCs with statistical comparison are in Table 2 above.

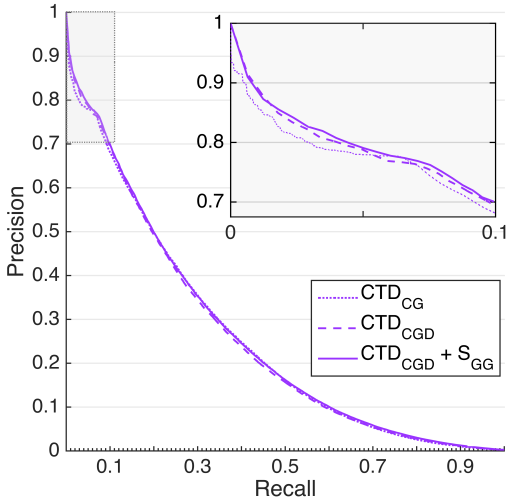

Fig. 6b. Plot of Precision vs. Recall for the same experiments shows precision approaches 1.0 for the top recovered interactions. Focusing on top 10% (inset) shows improvement with more data.

3.5. *Retrospective prediction of new Chemical-Gene interactions*

Finally, to corroborate these results in a more realistic context, we retrospectively predicted Chemical-Gene interactions that had been added to CTD over one year. Following the same process described in Section 2.1, we downloaded the CTD Chemical-Gene network on April 5, 2015, and again built a binary matrix of direct interactions. We mapped this to the 2014 CTD CGD matrix, removing entities that were not present in both versions, resulting in a 2015 matrix of 8,706 Chemicals by 8,304 Genes with 5,879 new interactions.

We calculated an ROC curve (shown in Figure 6) comparing these new interactions to the predictions for the same 8,706 Chemicals and 8,304 Genes that were obtained from NMF (k=200) on our CTD+STRING CGD matrix. The

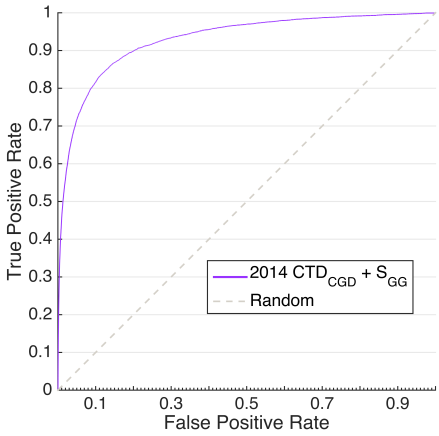

Fig. 6. Retrospective prediction of new CTD Chemical-Gene interactions (added between 4/2014 and 4/2015), using NMF (k=200) on CTD+STRING CGD matrix. AUC=0.930.

[g] The positive class comprised 75,804 interactions, so the inset shows precision for over 7500.

resulting curve, with AUC=0.93, indicates that our approach was able to correctly anticipate missing or undiscovered interactions.

3.6. *Example: prediction of Chemical-Disease interactions for Pancreatic Neoplasms*

One potential application of our approach is to identify unknown or overlooked drugs with connections to a particular disease. In Table 3, we present an example of this involving pancreatic cancer, a disease with high lethality and few effective treatments[22]. Following NMF (k=200) over the CTD+STRING CGD matrix, we inspected the highest[h] values corresponding to new interactions (that is, interactions that have not been curated by CTD at this time) between Chemicals and the disease entity "Pancreatic Neoplasms." Examples were chosen in which the Chemical is a drug[i]; as the primary focus of CTD is toxicology, much of the information therein concerns environmental toxins and disease-causing interactions. As Table 3 shows, literature searches found evidence supporting a connection to pancreatic cancer for 14 of the top 15 drug predictions, over half of which were studied in clinical trials. This shows that, at minimum, our approach generated hypotheses worth testing clinically.

Table 3. Top[h] 15 drugs[i] predicted to interact with Pancreatic Neoplasms by NMF using the CTD+STRING CGD matrix. These interactions were not present in the CTD CD matrix, but 14 are supported by papers or clinical trials in associated PubMed ID (PMID).

Drug Name	Support for Connection to Pancreatic Cancer	Reference
Indomethacin	Pre-clinical cell line study	PMID: 1890839
Carboplatin	Phase II clinical trial	PMID: 15802284
Mitoxantrone	Phase II clinical trial	PMID: 16334117
Simvastatin	Phase II clinical trial	PMID: 24162380
Cytarabine	Phase III clinical trial	PMID: 1833042
Topotecan	Phase II clinical trial	PMID: 11218186
Sorafenib	Phase II clinical trial	PMID: 24574334
Rosiglitazone	Pre-clinical mouse study	PMID: 22864396
Melphalan	Pre-clinical rat study	PMID: 4075299
Methamphetamine	-	-
Thiotepa	Use in other cancers	PMID: 4183076
Thalidomide	Phase I clinical trial	PMID: 15753541
Caffeine	Phase III clinical trial	PMID: 1833042
Sirolimus	Patient Case Report	PMID: 19581741
Gefitinib	Phase II clinical trial	PMID: 19258727

[h] Values above a threshold of 0.425. To provide context for this choice of threshold, the inset in Figure 5b shows cross-validation performance as precision versus recall at varying thresholds; 0.1 Recall in that graph corresponds to a threshold value of 0.425. Thus, we chose predictions whose precision should be at least 0.7.

[i] Approved by the FDA, according to http://www.accessdata.fda.gov/scripts/cder/ob/default.cfm

At the same time, this example highlights key pitfalls. Because we created binary interaction matrices from CTD, we can not say these drugs are predicted to treat pancreatic cancer or to cause it, only that they interact in some way. Indeed, the clinical trial we reference for simvastatin found no significant effect, but suggested further study in specific circumstances that could benefit from it[23]. Incorporating more detail from the interactions in CTD into our CGD matrix will, we believe, help resolve some of the ambiguity in our current predictions. For truly personalized treatments, we foresee a use case in which therapy suggestions are derived from a subset of predicted drug-gene interactions. That subset would be determined by a patient's unique situation; for example, the somatic mutations driving a tumor, or the germ line mutations linked to a disease phenotype (the latter being a possible application for our approach's gene-disease predictions).

4. Conclusions

Taken as a whole, our results show that Collaborative Filtering can integrate biological interaction networks in order to reveal missing connections between diverse entities. This approach depends only on knowledge of connections, so it can be extended to new classes of entity with minimal customization, unlike more specialized methods. Consequentially, our approach is limited to predicting *that* entities interact, rather than *how*. Matrix tri-factorization, which has been used to classify entities by fusing interaction networks with entity feature data[24,25], may enable more detailed predictions. Ultimately, however, we see this as an initial component in a pipeline that will harness the ever-expanding universe of knowledge and focus it on a small point, illuminating a patient's unique situation or highlighting a new use for a drug. This will need to be done rapidly, affordably, and accessibly. Importantly, implementations of NMF have been developed that can efficiently handle matrices with millions of times more entities than we have so far attempted[13,26]. Ultimately, this work may offer a step towards computing therapy.

5. Acknowledgements

This work was funded through DARPA SIMPLEX N66001-15-C-4042, NSF DBI1356569, and NIH NIDCR 1U01DE025181-01.

6. References

1. Franceschini, A, *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 41, D808–15 (2013).
2. Law, V, *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* 42, D1091–7 (2014).
3. Gaulton, A, *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–7 (2012).
4. Kanehisa, M, *et al.* Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Res.* 42, 199–205 (2014).
5. Croft, D, *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res.* 42, 472–7 (2014).

6. Milacic, M, *et al.* Annotating cancer variants and anti-cancer therapeutics in Reactome. *Cancers* 4, 1180–211 (2012).
7. Davis, AP, *et al.* The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Res.* 9880, 1–7 (2014).
8. Xu, T, *et al.* Quantitatively integrating molecular structure and bioactivity profile evidence into drug-target relationship analysis. *BMC Bioinf.* 13, 75 (2012).
9. Huang, L-C, *et al.* A weighted and integrated drug-target interactome: drug repurposing for schizophrenia as a use case. *BMC Syst. Biol.* 9, S2 (2015).
10. Yang, F, *et al.* Drug-target interaction prediction by integrating chemical, genomic, functional and pharmacological data. *Pacific Symp. Biocomp.* 148–59 (2014).
11. Huang, Y-F, *et al.* Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation. *BMC Med. Geno.* 6 Suppl 3, S4 (2013).
12. Zhang, S, *et al.* Learning from Incomplete Ratings Using Non-negative Matrix Factorization. *SDM* 549–53 (2006).
13. Zhou, Y, *et al.* Large-scale parallel collaborative filtering for the netflix prize. *Algo. Asp.* (2008).
14. Paatero, P & Tapper, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* 5, 111–26 (1994).
15. Lee, DD & Seung, HS. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–91 (1999).
16. Kim, H & Park, H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* 23, 1495–502 (2007).
17. Cobanoglu, MC, *et al.* Predicting drug-target interactions using probabilistic matrix factorization. *J. Chem. Inf. Model.* 53, 3399–409 (2013).
18. Zheng, X, *et al.* Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. *Proc. 19th ACM SIGKDD* 1025 (2013).
19. Wang, H, *et al.* Predicting protein-protein interactions from multimodal biological data sources via nonnegative matrix tri-factorization. *J. Comput. Biol.* 20, 344–58 (2013).
20. Vergara, I a, *et al.* StAR: a simple tool for the statistical comparison of ROC curves. *BMC Bioinf.* 9, 265 (2008).
21. DeLong, ER, *et al.* Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44, 837–45 (1988).
22. Ryan, DP, *et al.* Pancreatic Adenocarcinoma. *N. Engl. J. Med.* 371, 1039–49 (2014).
23. Hong, JY, *et al.* Randomized double-blinded, placebo-controlled phase II trial of simvastatin and gemcitabine in advanced pancreatic cancer patients. *Cancer Chemother. Pharmacol.* 73, 125–30 (2014).
24. Zitnik, M & Zupan, B. Matrix factorization-based data fusion for gene function prediction in baker's yeast and slime mold. *Pacific Symp. Biocomp.* (2014).
25. Žitnik, M & Zupan, B. Data fusion by matrix factorization. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 41–53 (2015).
26. Liu, C, *et al.* Distributed nonnegative matrix factorization for web-scale dyadic data analysis on mapreduce. *WWW 2010* (2010).