Non-Negative Matrix Factorization for Drug Repositioning: Experiments with the repoDB Dataset

Gokhan Bakal, M.S.[1], Halil Kilicoglu, Ph.D.[2], Ramakanth Kavuluru, Ph.D.[1]
[1]University of Kentucky, Lexington, KY; [2]National Library of Medicine, Bethesda, MD.

Abstract

Computational methods for drug repositioning are gaining mainstream attention with the availability of experimental gene expression datasets and manually curated relational information in knowledge bases. When building repurposing tools, a fundamental limitation is the lack of gold standard datasets that contain realistic true negative examples of drug–disease pairs that were shown to be non-indications. To address this gap, the repoDB dataset was created in 2017 as a first of its kind realistic resource to benchmark drug repositioning methods — its positive examples are drawn from FDA approved indications and negatives examples are derived from failed clinical trials. In this paper, we present the first effort for repositioning that directly tests against repoDB instances. By using hand-curated drug–disease indications from the UMLS Metathesaurus and automatically extracted relations from the SemMedDB database, we employ non-negative matrix factorization (NMF) methods to recover repoDB positive indications. Among recoverable approved indications, our NMF methods achieve 96% recall with 80% precision providing further evidence that hand-curated knowledge and matrix completion methods can be exploited for hypothesis generation.

1 INTRODUCTION

Repositioning previously approved drugs for new indications has become highly desirable in the biomedical and pharmaceutical research enterprises given expected time/cost reductions in identifying new treatment options. With recent estimates putting new drug development R&D costs over $2.5 billion per drug[1], repurposing has been gaining mainstream attention over the past five years. With previously approved drugs already passing the required safety tests for use in humans, the cost of repositioning is expected to be substantially lower compared with starting from a blank slate. In 2011, the National Library of Medicine (NLM) introduced `drug repositioning` (DR) as a new Medical Subject Heading (MeSH term) and as of now there are 1,161 articles tagged with it dating back to a single article from 2009. Almost 85% of these articles are published in the last five years indicating the sudden and deserved surge of interest in this area. Physicians with deep understanding of both the mechanisms of action for drugs and disease characteristics may be able to recommend off-label use[2] in an *ad hoc* manner. However, this does not constitute FDA approved recommendation for specific new indication(s) of drugs for use in designated groups of patients. As such, DR (via FDA approval) for new indications has significant potential to impact care at a broad scale and might also result in lowered costs for patients.

1.1. Computational drug repositioning. Computational drug repositioning[3] (CDR) is the use of informatics and high performance computing methods to prioritize candidates for new indications. With the simultaneous excitement surrounding biomedical data science and the explosive growth of publicly shared datasets, CDR methods are on the rise in the scientific community. A class of such methods exploits the notion of "similarity" between different entities involved in disease and therapeutic mechanisms. For example, shared traits among drugs including chemical structures, molecular activities, and side effects may be used to define a feature vector to represent a drug. Likewise, similarities can also be established between diseases based on established gene–disease associations or graph based proximity in disease ontologies. Zhang et al.[4] provide a unified framework that exploits these similarities for CDR. Another direction of CDR is exploiting available large-scale genomic data sources. For instance, Dudley et al.[5] utilized drug-gene expression signatures to discover a potential new drug for the inflammatory bowel disease. Topological analyses of drug–target networks and target-involved pathways are another mode of identifying potential new indications[6]. Text mining programs that extract different relations from text using natural language processing (NLP) and literature based discovery approaches that build on such relations are also being employed for CDR[7,8]. A more detailed treatise of CDR methods is available in a recent survey[3].

1.2. CDR method assessment. Although CDR is gaining prominence, evaluating CDR methods can be tricky given the lack of datasets that are tailored for it. Specifically, from our literature review we were able to identify very few

standardized datasets[4,9] that are uniformly used across efforts for benchmarking purposes. Furthermore, the datasets used in prior efforts have a serious shortcoming — they only contain positive drug–disease indication pairs; and hence prior efforts assume that all other combinations are negatives, which is unreasonable and potentially rules out novel repositioning predictions as false cases. Brown and Patel[10] highlight this shortcoming and propose a new gold standard database called *repoDB* for CDR method benchmarking. *repoDB* draws approved indications from the DrugCentral[11] database and failed indications from the American Association of Clinical Trials Database (the 'AACT Database'[12]), which is a structured version of information from NLM's ClinicalTrials.gov service. Given failed indications are part of the dataset, one can directly assess CDR methods with vetted indications and non-indications from `repoDB`. Since its introduction in 2017, however, we are not aware of any CDR efforts evaluating against `repoDB`.

In this paper, we present the first CDR attempt that directly tests against `repoDB` instances. First, a partially observed matrix is built using drug–disease *treatment* relations drawn from the UMLS Metathesaurus[13] and those extracted using automated NLP methods and made available by the NLM as part of the SemMedDB database[14,15]. Next, this matrix is completed by filling unobserved cells via non-negative matrix factorization (NMF) to elicit new indications. Our method uses a small portion of `repoDB` as a validation dataset and uses the bulk of it for testing purposes.

2 MATERIALS AND METHODS

2.1. Datasets. In this section, we describe the data sources from which we derive our training and testing examples. The UMLS[13] and SemMedDB[14,15] are our essential data resources for training instances while `repoDB` is our resource for the test examples. We use the terms "training" and "testing" to emphasize that this is still a (weakly) supervised method where the training instances are simply drawn from external resources both manually curated (UMLS) and automatically extracted (SemMedDB). We will briefly describe each data source in the following subsections.

UMLS Metathesaurus. UMLS is a longstanding terminological resource that integrates over 160 different vocabularies updated every year by the NLM. The Metathesaurus portion of UMLS aggregates equivalent concepts across multiple vocabularies and assigns to each unique concept a *concept unique identifiers* (CUI). Besides synonymous names for each concept, there are also inter-concept relations sourced from the original vocabularies. We obtained *treatment* relations from the MRREL table* in UMLS Metathesaurus[13] version 2017AB. A total of 43,898 such relations are part of our UMLS training dataset.

SemMedDB – Semantic Medline Database. SemMedDB is a repository of (subject, predicate, object) triples called *semantic predications* extracted by a rule-based NLP tool SemRep[16] developed by the NLM. SemMedDB is built by running SemRep over all available PubMed citations (over 27 million) where the subject/object entities are normalized to UMLS CUIs. Likewise, the predicate is mapped to a relation type from the UMLS semantic network[17]. Given a predication can be extracted from multiple sentences, we also have frequency information (number of unique sentences containing it) for each SemMedDB triple. For our experiments, we curated *treatment* predications (triples where predicate = TREATS) in SemMedDB as additional training examples. As SemRep's precision is around 75%[15], we only collected predications which have been extracted at least twice, thrice, and five times to include them in the training set in various configurations (more later). Hence, we were able to obtain three different *treatment* predication sets of 55,349, 34,802 and 19,777 triples for the frequencies of 2, 3, and 5 respectively as long as **they are not occurring** in test sets from `repoDB`.

The `repoDB` database. As indicated in Section 1.2, instances in `repoDB` come from DrugCentral[11] and ClinicalTrials.gov[12] resources. It has a total of 6,677 approved and 3,885 failed drug–disease pairs. After removing the duplicates and the ones which appear in UMLS (given UMLS pairs will be part of the training dataset), we were left with 6,218 approved and 2,852 failed pairs. After removing pairs associated with drugs for which there is not even a single positive pair from UMLS/SemMedDB, we are left with 5,172 approved treatments (ATs) and 2,244 failed indications (FIs). This aligns with the nature of CDR to some extent — if we do not even have a single occurrence of a drug treating some disease, we may not be able to repurpose it for other conditions. This is also an inherent limitation of the matrix completion method we propose to use; if the row corresponding to a drug in the drug–disease matrix is empty, matrix completion methods cannot fill that row and hence it is impossible to come up with new indications for it (more later).

*Specifically, these are the relations where the **RELA** field in MRREL table is equal to one of these four types: *"treats", "may_treat", "treated_by"*, and *"may_be_treated_by"*

Generation of randomly selected negative examples. The `repoDB` examples are vetted ATs and FIs, identified based on clinical trials. We also wanted to build a separate dataset of random indications (RIs) which satisfy domain/range constraints for subjects/objects for *treats* predicate. The purpose is to see if our method would have a relatively easier or harder time when dealing with these when compared with FIs from `repoDB`. In the past we have generated such a dataset[18] for a slightly different task. Basically, these RI examples are created by the following steps.

- Each concept in UMLS has at least one semantic type[19] that represents a class membership. Furthermore, every predicate in the UMLS semantic network has a set of domain/range semantic type constraints defined by the NLM based on domain expert knowledge. Based on the allowable semantic type combination for the *treats* predicate, we randomly select pairs that satisfy the domain/range constraints.
- For the set of pairs selected using the previous step, we simply remove the pairs which appear as treatment relations either in UMLS or SemMedDB. Thus, we ensure that selected pairs do not occur in our training set.

The given steps above pick fairly hard-to-predict *potentially* negative examples because they satisfy the domain/range constraints and are not present in either UMLS or SemMedDB databases. Ultimately, we obtained 3,318 examples to be used as the RI test set for the matrix completion methods.

2.2. Methods. In this section we present the NMF based matrix completion method along with our approach to configure it with different input matrices from external data sources.

Matrix completion through NMF. Matrix completion[20] is the process of filling missing entries in a partially observed matrix. These partially observed matrices arise in many real world scenarios especially in recommender systems where preferences of people are encoded. A matrix with customers as rows and and products (e.g., movies, books) as columns is the typical setup. Given information about their prior ratings or product purchases represented as 1s in the corresponding cells, matrix completion would identify what other cells ought to be 1s — which other products would a customer likely enjoy given what they already liked. In a completely random world, there is no way to guess the new 1s. However, assuming the matrix has a much smaller rank than $\min(m, n)$ for the $m \times n$ matrix, we can use non-negative matrix factorization (NMF[21]) to come up with a low-rank approximation to the original matrix with $[0, 1]$ non-zero entries in blank cells, leading to potential new recommendations. This low-rank assumption is based on the intuition that there are latent themes/traits in user preferences and a typical user's preferences are not distributed truly randomly across the product space. A similar strategy is also employed in information retrieval for latent semantic indexing[22] for computing document similarity through dimensionality reduction.

One can now see that the CDR problem can be modeled similarly where the training treatment relations can be used to partially fill the drug–disease matrix, with NMF filling empty cells with non-zero values pointing to potential new indications. Since this is an approximation process, the new values in empty cells will be non-zero but generally not exactly 1. Thresholding based on a validation dataset can be used to glean indications if a particular cell's value crosses the threshold. The intuition here is also to exploit potential latent themes where groups of drugs sharing certain characteristics (e.g., mechanism of action) may treat clusters of conditions with similar traits (e.g., symptoms). Given we do not know what the myriad latent themes may be, we assume a certain number of them are present — the chosen low rank — and proceed with NMF for matrix completion. Thus, given the partially observed $m \times n$ drug–disease matrix X with m drugs and n diseases, we will approximate it as

$$\underset{m\times n}{\mathrm{X}} \approx \underset{m\times k}{W} \times \underset{k\times n}{H} = \underset{m\times n}{\hat{\mathrm{X}}}, \tag{1}$$

where W and H are the factors with rows of W representing k-dimensional drug vectors and columns of H encoding k-dimensional disease vectors under the assumption that X has rank $k \ll \min(m, n)$. The product $\hat{X} = WH$ approximates X helping us glean new non-zero values hinting at new indications, while the rows of W and columns of H can be used to compute drug and disease similarities respectively. The objective function to find the best approximation is

$$\underset{W,H}{\arg\min} \|X - WH\| + \underbrace{\beta(\|W\|_2 + \|H\|_2)}_{\text{regularization}}, \tag{2}$$

where $W \in \mathbb{R}_+^{m\times k}$ and $H \in \mathbb{R}_+^{k\times n}$ and β is the weight for the regularization penalty term to handle overfitting that corresponds to large norms for W and H. Next, the construction of the input drug–disease matrix X is discussed.

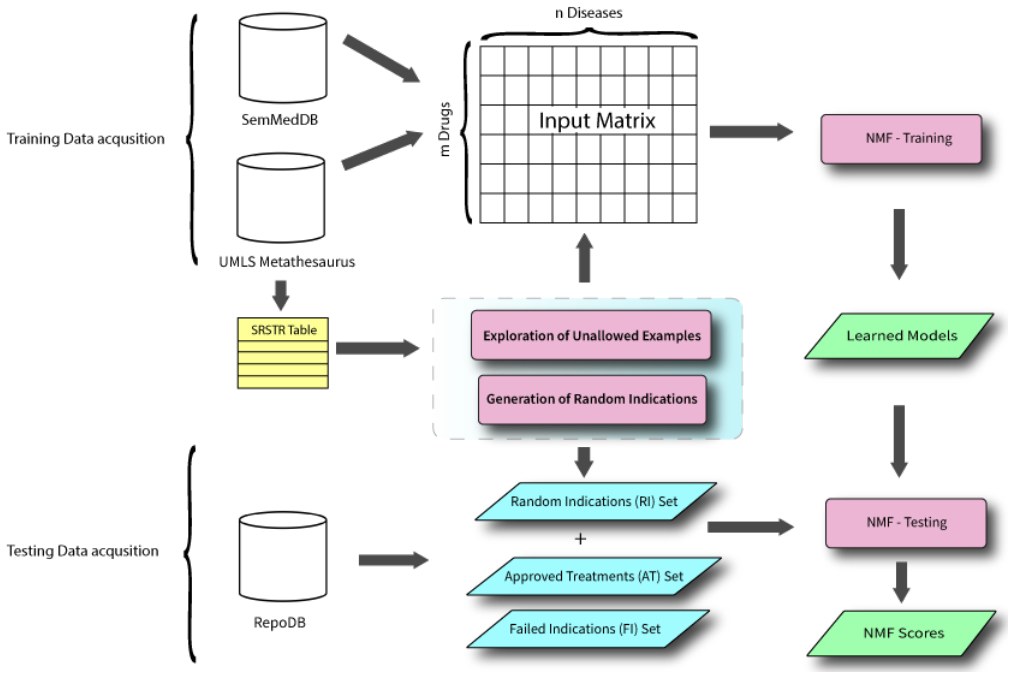

Figure 1: Schematic summary of the proposed NMF based CDR experiments

Building the input training matrix. The input partially observed matrix was constructed based on treatment relations from UMLS and SemMedDB as indicated in Section 2.1. However, we needed to consider a larger matrix to fill compared with drugs and diseases of positive indications from UMLS Metathesaurus and SemMedDB. Otherwise, the test indications in `repoDB` cannot be recovered as part of the completion process. For this, we considered all allowed subject/object semantic type constraints[†] for *treats* predicate (based on SRSTR tables of the UMLS semantic network). Next, we selected all UMLS subject concepts where each has at least one semantic type that belongs to the set of allowed subject types; likewise, we aggregated all UMLS concepts where each has at least one semantic type that is from allowable object types. Based on this, 538,710 subject entities and 314,707 object entities were obtained for the input matrix. However, most subjects do not have a treatment relation with any disease as observed in the Metathesaurus training dataset. Hence, we removed all zero rows (hence the corresponding drugs or treatment agents) and retained only rows that are known to treat ≥ 1 disease in the columns to exploit the shared therapeutic context among the drugs. After this pruning process, we were left with 10,188 drugs (or treatment agents) and 314,707 objects[‡] to build our input matrix to be partially filled (from training datasets) and subsequently completed via NMF.

To populate the input matrix with training relations, we have Metathesaurus treatment relations and three different sets of semantic predications derived from SemMedDB based on extraction frequencies. We have a configuration where training relations are entered in the tables as 1s, which we call the "**binary matrix factorization (BMF)**" model. In addition to BMF, we have another configuration using SemMedDB treatment predications with their extraction frequency counts. We term this as the "**count matrix factorization (CMF)**" model to evaluate the performance when counts are used instead of Boolean indicators.

Finally, as part of the input we have many unallowed input matrix cells (5,531,386 in total) that cannot be 1s because

[†]As an example, (*Pharmacologic Substance*, *Disease or Syndrome*) is a popular semantic type combination allowed for treatment relations. (*Antibiotic*, *Disease or Syndrome*) and (*Therapeutic or Preventive Procedure*, *Congenital Abnormality*) are less common allowed type combinations. Overall, there are 56 different allowed type combinations for treatment relations as per UMLS. There are other additional allowed types that NLM has incorporated as part of the schema design for SemMedDB and those are included for this effort as allowed combinations.

[‡]Note that not all objects are diseases per se, some maybe symptoms and at times different patient groups. For example, the UMLS CUI C4316221 refers to "Patients with a diagnosis or past history of total colectomy or colorectal cancer" and can be an object of some treatment relations. To capture all latent themes we end up using all subjects and objects of treatment relations as part of the input matrix. However, for evaluation purposes, after NMF, we only look at those cells (in $\hat{X}$ from Eq. (1)) corresponding to approved and failed indication pairs in `repoDB`.

the corresponding subject–object pairs do not satisfy the domain/range semantic type constraints. For example, consider this unallowed type combination: (*Drug Delivery Device*, *Patient or Disabled Group*). Although *Drug Delivery Device* is an allowed subject type for some other type combination(s) and *Patient or Disabled Group* is an allowed object type for a different combination, this particular coupling is not allowed. But cells corresponding to this unallowed combination exist in the input matrix given the matrix was built with all allowed subjects and objects as rows and columns, respectively. Thus, these cells corresponding to entity pairs that satisfy this type combination must be designated as unallowed. To avoid such predictions, we assign 0s to the corresponding cells of the input matrix. This is another way to further constrain the factorization process to approximate both positive and unallowed cases while estimating the unobserved cells. We experimented with several input matrices with different numbers of unallowed cases to observe their influence on the prediction task. The levels of these unallowed examples are set to 0% (no unallowed examples), 25%, 50%, 75%, and 100% (all of them) in the input matrix. The overall framework of our approach is demonstrated in Figure 1.

NMF experimental configurations. We note that `repoDB` drugs/diseases for ATs and FIs are already mapped to UMLS CUIs by its creators. Hence, the matrix constructed and completed as described in Section 2.2 naturally suffices for the CDR task. When training, experiments were conducted with singular value decomposition (SVD) to identify a k value (to be used for the dimensionality in Eq. (1)) that minimizes the mean squared error (MSE) for the cells that are already filled in the training matrix. Since there were not any noticeable differences between MSE values with $k = 50, 100, 500$, we chose $k = 50$ for our further matrix completion experiments. Thus results are reported for $k = 50$ for all experiments. The regularization parameter β was left at the default value (0.1) because tuning it did not yield any apparent gains. To carry out the optimization in Eq. (2), the open source MF library LIBMF[23] was used for incomplete matrix approximation. LIBMF is an efficient stochastic gradient descent based software package that runs parallel on multiple cores in a shared-memory environment.

3 RESULTS

We assess NMF results from two different perspectives. First we look the actual NMF scores produced for ATs, FIs, and RIs we created as part of Section 2.1. Subsequently, we observe how these NMF scores can be used to come up with precision, recall, and F-score based on ATs and FIs from `repoDB`.

3.1. NMF scores for `repoDB` pairs. In Table 1 we show mean NMF scores for ATs and FIs in `repoDB` and the RIs we generated. The scores are real-valued numbers with which NMF fills unobserved cells as part of the training process for various configurations. Configurations differ from each other with respect to what is included in the input matrix and the proportions of unallowed pairs included as 0s. Due to the optimization in Eq. (2) and encoding of positive cases as 1s, the higher the value estimated for an unobserved cell, the stronger the plausibility of treatment relationship for the corresponding drug–disease pair. We make the following important observations from Table 1

- Only adding positive training examples to the input matrix and leaving unallowed examples as unobserved (the 0% column) leads to catastrophically bad results where the FIs and RIs are scoring higher. Hence we do not discuss any results going forward where the unallowed examples are left as unobserved. Adding additional unallowed examples (all other columns) as 0s in the input matrix shows more realistic scores following a clear pattern where the ATs score higher than FIs, which fare better than RIs. This confirms a few things: (a). The 0s inserted to account for unallowed cases are providing enough signal to guide the optimization process to distinguish between more plausible indications from random ones (which are mostly useless). (b). NMF based completion is able to score ATs better than FIs in `repoDB` where the mean AT score is 2–3 times higher than that of the FI score, demonstrating its effectiveness. (c). FIs scoring higher than RIs reflects the reality that FIs actually went through the process of clinical trials, which implies researchers felt that those pairs were plausible indications; RIs however are just random pairs of drugs and diseases. (d). Including more unallowed pairs as part of the input quickly decreases the magnitude of the scores (compare the 25% column with the 100% column); however, the relative differences between ATs, FIs, and RIs persist all across the board.
- Adding more training positives from SemMedDB (rows 4–12) increases the absolute values of the scores and also the differences in scores between ATs, FIs, and RIs, but the relative differences are the highest when using just UMLS Methasaurus relations. For example, in the 25% column, the ratio of means of AT and FIs for

Model	Training Data	Test Sets	Portion of included unallowed pairs				
			0%	25%	50%	75%	100%
BMF	UMLS only	Approved treatments	0.958	**0.268**	0.167	0.102	0.071
		Failed indications	0.980	**0.119**	0.072	0.036	0.020
		Random indications	0.995	**0.023**	0.013	0.009	0.005
BMF	UMLS + SemMedDB (MinFreq. 5)	Approved treatments	0.936	**0.564**	0.559	0.556	0.546
		Failed indications	0.957	**0.357**	0.343	0.331	0.325
		Random indications	0.987	**0.102**	0.100	0.098	0.092
BMF	UMLS + SemMedDB (MinFreq. 3)	Approved treatments	0.930	**0.614**	0.611	0.608	0.597
		Failed indications	0.954	**0.383**	0.371	0.359	0.352
		Random indications	0.983	**0.139**	0.135	0.132	0.124
BMF	UMLS + SemMedDB (MinFreq. 2)	Approved treatments	0.927	**0.650**	0.647	0.645	0.636
		Failed indications	0.951	**0.413**	0.399	0.386	0.381
		Random indications	0.976	**0.195**	0.190	0.186	0.174
CMF	UMLS + SemMedDB	Approved treatments	34.758	**7.209**	6.194	5.145	3.595
		Failed indications	30.327	1.678	3.600	0.981	0.895
		Random indications	39.603	**1.977**	0.689	0.186	0.553

Table 1: Mean of the predicted NMF scores of test sets with different configurations

"UMLS only" (0.268/0.119 = 2.25) is higher than the corresponding ratio for "UMLS+SemMedDB (MinFreq. 5)" (0.564/0.357=1.58), which stays at a similar level even as additional relations are added (MinFreq values 3 and 2).

- Count based models (instead of the binary models) where frequencies are included in the input matrix appear not as consistent (last three rows) where for the 25% column, we notice RIs scoring higher than FIs.

We computed 95% confidence intervals that showed that the score differences are statistically significant. Here we disclose the intervals for the three rows of "UMLS+SemMedDB (MinFreq. 2)" (of the 25% unallowed cases column): **0.650** $\pm$ 0.010 (ATs), **0.413** $\pm$ 0.017 (FIs), and **0.195** $\pm$ 0.012 (RIs). The intervals do not overlap further confirming the NMF method's functionality.

3.2. Precision, Recall, and F-score for ATs in `repoDB`. The NMF score ranges in Section 3.1 demonstrate that, *on average*, NMF maps ATs, FIs, and RIs to non-overlapping segments on the real number scale with high confidence. However, we still need a way to make repositioning Yes/No decisions at the instance level based on the score generated for a particular drug–disease pair corresponding to an entry in $\hat{X}$ in Eq. (1). One way to make such a decision is to choose a threshold for the NMF score and assign all pairs with scores above that threshold as new candidates for repositioning. Here we propose to do that by splitting the `repoDB` ATs and FIs into validation and test sets. We considered 20% of ATs and 20% of FIs as comprising the validation set while the rest are left for the final test[§]. We identified a threshold based on grid search over the validation dataset optimized for F-score with a small step size of 0.00001 spanning the range $[\mathcal{T}^v_{min}, \mathcal{F}^v_{max}]$ such that

$$\mathcal{T}^v_{min} = \min(\{\hat{X}_{i,j} : (i,j) \in \mathcal{T}^v\}) \quad \text{and} \quad \mathcal{F}^v_{max} = \max(\{\hat{X}_{i,j} : (i,j) \in \mathcal{F}^v\}),$$

where $\hat{X}$ is the approximation from Eq. (1) and $\mathcal{T}^v$ and $\mathcal{F}^v$ represent the validation datasets for ATs and FIs, respectively. This range was chosen based on the observation on the validation dataset that $\mathcal{T}^v_{min}$ is smaller than $\mathcal{F}^v_{max}$ (so there were some AT scores that were less than other FI scores). Hence choosing $\mathcal{T}^v_{min}$ as the threshold corresponds to 100% recall and selecting $\mathcal{F}^v_{max}$ leads to 100% precision. Thus by limiting the grid search to the threshold range $[\mathcal{T}^v_{min}, \mathcal{F}^v_{max}]$, we are exploring the space of compromise between perfect precision and perfect recall.

[§]This translates to 4138 ATs and 1795 FIs in the test set and 1034 ATs and 449 FIs in the validation dataset — numbers computed based on the original `repoDB` counts from Section 2.1.

		UMLS + SemMedDB(MinFreq. 5)			UMLS + SemMedDB(MinFreq. 3)			UMLS + SemMedDB(MinFreq. 2)			UMLS only			
Unallowed cases	Threshold	P	R	F-score	P	R	F-score	P	R	F-score	Threshold	P	R	F-score
25%	0.00001	0.812	0.942	0.8727	0.808	0.961	0.8787	0.800	0.964	0.8750	0.00003	0.906	0.884	**0.8952**
50%	0.00001	0.811	0.942	0.8723	0.808	0.961	0.8787	0.800	0.964	0.8750	0.00004	0.906	0.883	0.8950
75%	0.00001	0.812	0.941	0.8721	0.808	0.961	0.8787	0.800	0.964	0.8750	0.00002	0.905	0.878	0.8916
100%	0	0.723	0.981	0.8328	0.714	0.996	0.8322	0.709	0.996	0.8287	0	0.840	0.909	0.8736

Table 2: Performance results of BMF models for approved indications in `repoDB`

Once a threshold is chosen to make instance level decisions for test examples, it is straightforward to assess the performance of the method using traditional measures such as precision, recall, and F-score. Thus, in Table 2, we report the performance results for the BMF models for different configurations of the input matrix and different levels of included unallowed examples. The first observation is that the thresholds selected are all very close to zero indicating that boundary case NMF scores were close to zero for ATs across all configurations. The thresholds are identical for configurations with SemMedDB examples but change slightly for UMLS-Only case. We notice that the best F-score of 0.895 (first row, last column) is obtained for UMLS-Only input matrix with 25% unallowed example constraints. This may be explained from the biggest relative difference between mean AT and FI scores for this configuration from Section 3.1. However, adding SemMedDB training instances (with minimum frequencies 2 and 3) seems to lead to a potentially more desirable compromise with recall around 96% and precision over 80%. The results for count based CMF models were disappointing as shown in Table 3. There is no clear pattern as to how the scores are spread with regards to different configurations and overall performance is all across the board inferior when compared to BMF models especially with substantially lower precision values.

Unallowed cases	Threshold	P	R	F-score
25%	0.12162	0.714	0.957	**0.8185**
50%	0.08999	0.698	0.968	0.8119
75%	0.00009	0.692	0.973	0.8093
100%	0.005	0.703	0.963	0.8131

Table 3: Performance results for CMF models over `repoDB`

4 DISCUSSION

As CDR efforts continue to rise, it is critical to have benchmarking datasets that are realistic in terms of representation of both approved indications and failed indications. `repoDB` is first of its kind dataset that creates such an opportunity to conduct comparative evaluations of CDR methods on a publicly available gold standard dataset.

4.1. Main takeaways. Matrix completion through NMF based low-rank approximation is an effective method for CDR based solely on datasets of previously approved drugs and corresponding indications. Actually, in this manuscript, we only use public data sources of treatment relations in the form of hand curated UMLS Metathesaurus relations and those extracted with NLP from PubMed citations (from SemMedDB). As such, these are imperfect resources (especially SemMedDB) and may not necessarily constitute FDA approved drugs. Results still show that among recoverable ATs from `repoDB`, we achieved F-scores close to 90% with the highest F-score achieved with just UMLS relations as input. Using both SemMedDB and UMLS relations helps achieve a better compromise between precision and recall with over 96% recall at 80% precision. The mean NMF score for FIs is at least twice as large as that for RIs, indicating that FIs are indeed much tougher to distinguish from ATs compared with randomly generated pairs. A critical enabler was the encoding of unallowed pairs (derived with incompatible semantic type constraints from UMLS semantic network) as zeros imposing additional structural constraints on the input matrix to be approximated. However, imposing constraints from *all* unallowed pairs could be detrimental by leading to a 3% recall gain with a 10% precision drop. Experiments showed that introducing 25% of the zeroes from unallowed pairs leads to better outcomes and is computationally less expensive. Count based models that consider frequency from SemMedDB substantially underperform compared with simpler binary models. Overall, NMF based methods applied to carefully curated external knowledge

sources constitute a practical approach towards CDR.

Next we discuss some examples of correct predictions made by our approach. In our training dataset we see the drug vincristine treating *malignant neoplasms, follicular lymphoma*, and *Hodgkin disease* and another drug doxorubicin treating the general condition of *malignant neoplasms*. After matrix completion, we saw high values of 0.89 and 0.93 for the entries (doxorubicin, *follicular lymphoma*) and (doxorubicin, *Hodgkin disease*) respectively, which are approved indications in `repoDB` that were never encountered in training data and were blank cells before the training process. Similar new correct predictions are also made for (bleomycin, *follicular lymphoma*) and (bleomycin, *Hodgkin disease*). Next, although, the count based CMF method underperformed overall, there were cases where it lead to correct predications when the binary approach did not. For example, (betamethasone, *berylliosis*) and (bleomycin, *malignant head and neck neoplasm*) are approved indications that were missed by the BMF approach but are recovered by the CMF method. Thus there may be some complementary traits in how the BMF and CMF approaches predict that need further examination toward building an ensemble method.

We set out to explore reasons for errors — false positives (FPs) and false negatives (FNs) — incurred by the NMF models in the context of information available about the corresponding drug–disease pairs. To this end, we examined connectedness of FP and FN pairs in the SemMedDB graph, which essentially conveys the potential shared context between associated entities. In our prior work[24], we identified graph patterns over the SemMedDB graph that are highly indicative of treatment relations using model coefficients of a logistic regression (LR) model[¶]. For FPs of the NMF model, we noticed that there were tens of thousands of highly predictive short paths (length ≤ 3) connecting the corresponding drug and disease CUIs indicating that there are many shared neighbors; some of this neighborhood information is encoded in the input matrix, which could have led to positive predictions. This is also not surprising given FPs are essentially failed cases in `repoDB` but were deemed plausible enough for researchers to launch clinical trials. For FNs, we found relatively fewer and sometimes no such predictive paths in SemMedDB connecting associated entities. For example, for the approved `repoDB` indication (Dexamethasone, Branch retinal vein occlusion with macular edema), the drug and disease were not connected in the SemMedDB graph using LR model's top predictive patterns. Without much shared context, NMF appears to struggle to elicit positive indications for such pairs. We plan to pursue a more detailed error analysis involving physician experts, which may yield additional insights on potential reasons for errors.

4.2. Limitations and future work. This current effort is not without a few limitations, which also point to interesting future research directions for CDR experiments with `repoDB`.

- The method in this paper is clearly not a silver bullet for CDR. `repoDB` does enable excellent benchmarking but in general scientists are often looking at a particular disease that they want to treat. Hence, for disease specific CDR, more sophisticated methods involving gene expression datasets and methods that consider integration of various modalities of information specific to the disease may be needed, as indicated in other prior efforts (e.g., Nagaraj et al.[25] for cancer). However, our method can be an effective initial step in pruning the space of candidates before more sophisticated methods that require more complex modeling and disease specific information can be applied.
- As discussed earlier, the count based CMF models' performance was underwhelming (Table 3) when compared with the binary models even though the counts capture additional information about prior knowledge being incorporated into the input matrix. One reason for this could be that we simply employed raw frequencies of treatment predications in SemMedDB instead of standardizing counts using well-known methods[26] (e.g., mean centering, min-max scaling, log transformation). Using raw frequencies may have lead to potential ill-conditioning that needs to be countered with appropriate pre-processing and/or using more sophisticated methods[27]. These experiments will be part of future extensions of our work.
- Matrix completion methods cannot fill a row that does not have at least one nonzero entry. In our case, this means, a drug for which we do not have at least one known treatment relation cannot be linked to new indications with NMF. However, this can be remedied by moving from matrices to tensors with additional relations between entities connected with other predicates including *prevents*, *diagnoses*, *affects*, and *causes*. Using tensor factorization[28], even

[¶]The LR model, while being effective, is computational prohibitive at times given the explosion of numbers of paths connecting entities in a large graph such as that built from SemMedDB. Our foray into NMF is motivated by these efficiency constraints of the graph pattern based approach.

for a drug with no existing treatment relations, using multi-hop indirect connections, it is possible to elicit a new indication. Similarly, with recent deep learning advances, embedding nodes and edges of the larger SemMedDB graph (including edges arising from other predicates besides *treats*) with graph neural networks can offer a different way for knowledge base completion[29]. We intend to pursue these directions in the immediate future.

4.3. Benchmarking. To enable future comparisons with our results, we provide the validation/test set splits of `repoDB` drug–disease pairs used in this study: `https://github.com/bionlproc/nmf-repoDB-benchmarking`. This will be important for direct comparisons by other researchers using the `repoDB` dataset, especially given we had to resort to using a subset of `repoDB` (owing to issues with lack of training instances for certain drugs without a single human vetted treatment relation).

5 CONCLUSION

With valuable time and cost savings in the offing, CDR efforts are expected to increase in the future. With lack of datasets modeling both positive and failed indications, it is encouraging to notice that datasets such as `repoDB` are being created. However, it is also important to start comparing methods against such datasets for robust assessments of different methods. In this paper, matrix completion through NMF was used to directly predict `repoDB` approved indications by using publicly available treatment relations. F-scores close to 90% were obtained with various training configurations with this method showing its strong potential for practical applications. Validation and test splits of `repoDB` used as part of this effort are made available to facilitate direct comparisons with our results by other researchers in the CDR community. More sophisticated methods such as tensor factorizations and neural graph embeddings may hold the promise of recovering novel indications for drug compounds that have not yet been approved for any known conditions. We believe this is the first attempt to employ `repoDB` for CDR purposes and hope that this will trigger more attempts to pursue this line of work toward rigorous benchmarking.

Acknowledgements

We are grateful for the support of the U.S. National Library of Medicine through NIH grant R21LM012274 and also thankful for partial support offered by the U.S. National Center for Advancing Translational Sciences via grant UL1TR001998. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. We also acknowledge the Ministry of National Education, Republic of Turkey, for providing financial support to Gokhan Bakal with full scholarship for his doctoral studies. HK was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

References

[1] Joseph A DiMasi, Henry G Grabowski, and Ronald W Hansen. Innovation in the pharmaceutical industry: new estimates of R&D costs. *Journal of health economics*, 47:20–33, 2016.

[2] Randall S Stafford. Regulating off-label drug use — rethinking the role of the FDA. *New England Journal of Medicine*, 358(14):1427–1429, 2008.

[3] Jiao Li, Si Zheng, Bin Chen, Atul J Butte, S Joshua Swamidass, and Zhiyong Lu. A survey of current trends in computational drug repositioning. *Briefings in bioinformatics*, 17(1):2–12, 2016.

[4] Ping Zhang, Fei Wang, and Jianying Hu. Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity. In *AMIA Annual Symposium Proceedings*, volume 2014, pages 1258–1267. American Medical Informatics Association, 2014.

[5] Joel T Dudley, Marina Sirota, Mohan Shenoy, Reetesh K Pai, Silke Roedder, Annie P Chiang, Alex A Morgan, Minnie M Sarwal, Pankaj Jay Pasricha, and Atul J Butte. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Science translational medicine*, 3(96):96ra76–96ra76, 2011.

[6] Jiao Li and Zhiyong Lu. Pathway-based drug repositioning using causal inference. *BMC bioinformatics*, 14(16):S3, 2013.

[7] Christos Andronis, Anuj Sharma, Vassilis Virvilis, Spyros Deftereos, and Aris Persidis. Literature mining, ontologies and information visualization for drug repurposing. *Briefings in bioinformatics*, 12(4):357–368, 2011.

[8] Trevor Cohen, Dominic Widdows, Clifford Stephan, Ralph Zinner, Jeri Kim, Thomas Rindflesch, and Peter Davies. Predicting high-throughput screening results with scalable literature-based discovery methods. *CPT:*

pharmacometrics & systems pharmacology, 3(10):1–9, 2014.

[9] Assaf Gottlieb, Gideon Y Stein, Eytan Ruppin, and Roded Sharan. Predict: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology*, 7(1):496, 2011.

[10] Adam S Brown and Chirag J Patel. A standard database for drug repositioning. *Scientific Data*, 4:170029, 2017.

[11] Oleg Ursu, Jayme Holmes, Jeffrey Knockel, Cristian G Bologa, Jeremy J Yang, Stephen L Mathias, Stuart J Nelson, and Tudor I Oprea. Drugcentral: online drug compendium. *Nucleic acids research*, 45(D1):D932–D939, 2017.

[12] Asba Tasneem, Laura Aberle, Hari Ananth, Swati Chakraborty, Karen Chiswell, Brian J McCourt, and Ricardo Pietrobon. The database for aggregate analysis of clinicaltrials. gov (aact) and subsequent regrouping by clinical specialty. *PloS one*, 7(3):e33677, 2012.

[13] National Library of Medicine. Unified Medical Language System Reference Manual. `http://www.ncbi.nlm.nih.gov/books/NBK9676/`, 2009.

[14] National Library of Medicine. Semantic MEDLINE Database. `http://skr3.nlm.nih.gov/SemMedDB/`, 2016.

[15] Halil Kilicoglu, Dongwook Shin, Marcelo Fiszman, Graciela Rosemblat, and Thomas C Rindflesch. Semmeddb: a pubmed-scale repository of biomedical semantic predications. *Bioinformatics*, 28(23):3158–3160, 2012.

[16] National Library of Medicine. SemRep - NLM's Semantic Predication Extraction Program. `http://semrep.nlm.nih.gov`, 2013.

[17] National Library of Medicine. Current Hierarchy of UMLS Predicates. `http://www.nlm.nih.gov/research/umls/META3_current_relations.html`, 2003.

[18] Gokhan Bakal and Ramakanth Kavuluru. Predicting treatment relations with semantic patterns over biomedical knowledge graphs. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 586–596. Springer, 2015.

[19] National Library of Medicine. Current Hierarchy of UMLS Semantic Types. `http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html`, 2003.

[20] Dietmar Jannach, Paul Resnick, Alexander Tuzhilin, and Markus Zanker. Recommender systems — beyond matrix completion. *Communications of the ACM*, 59(11):94–102, 2016.

[21] Yu-Xiong Wang and Yu-Jin Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1336–1353, 2013.

[22] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[23] Wei-Sheng Chin, Yong Zhuang, Yu-Chin Juan, and Chih-Jen Lin. A fast parallel stochastic gradient method for matrix factorization in shared memory systems. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(1):2:1–2:24, 2015.

[24] Gokhan Bakal, Preetham Talari, Elijah V. Kakani, and Ramakanth Kavuluru. Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations. *Journal of biomedical informatics*, 82:189–199, 2018.

[25] AB Nagaraj, QQ Wang, P Joseph, C Zheng, Y Chen, O Kovalenko, S Singh, A Armstrong, K Resnick, K Zanotti, et al. Using a novel computational drug-repositioning approach (drugpredict) to rapidly identify potent drug candidates for cancer treatment. *Oncogene*, 37(3):403, 2018.

[26] Robert A van den Berg, Huub CJ Hoefsloot, Johan A Westerhuis, Age K Smilde, and Mariët J van der Werf. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics*, 7(1):142, 2006.

[27] A Cichocki and R Zdunek. Multilayer nonnegative matrix factorisation. *Electronics Letters*, 42(16):947–948, 2006.

[28] Yuan Luo, Fei Wang, and Peter Szolovits. Tensor factorization toward precision medicine. *Briefings in bioinformatics*, 18(3):511–514, 2016.

[29] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer, 2018.