Using a Recommendation System to Predict Drug Status for Different Indications
A Needle in a Data Haystack Project

Eden Beyna, eden.beyna@mail.huji.ac.il, 209297597, edenbeyna
Aluma Etzion, aluma.etzion@mail.huji.ac.il, 211444179, alumae
Alon Biner, alon.biner@mail.huji.ac.il, 318418118, alonbiner

Problem description

Drug repurposing is a promising approach to identify new therapeutic uses for existing approved or investigational drugs. This approach can potentially save years of research and billions of dollars compared to traditional drug discovery. It is particularly valuable for addressing rare diseases, responding to urgent medical needs (like pandemics), and finding treatments for conditions that lack effective therapies.

Recommendation systems can play a valuable role in drug repurposing efforts by leveraging vast amounts of biomedical data, including drug structures, targets, side effects, and disease pathways to predict potential new indications for existing drugs.

This data-driven approach can significantly narrow down the list of candidates for experimental validation, making the drug repurposing process more efficient and increasing the likelihood of success in finding new treatments for various medical conditions.

Data

We have data about diseases and drugs that went through clinical trials for those diseases. We also have the outcome of the trial: Approved, Withdrawn, Terminated, or Suspended, and the phase in which the trial ended for drugs that weren't approved. In addition to the disease name and drug name we also have for each drug its drugbank ID, and for each disease its NIH concept ID. The dataset has 13558 records, and its size is 1.4MB. It's a .csv file with the following columns: drug_name, drugbank_id, ind_name, ind_id, NCT, status, phase, DetailedStatus. We downloaded it from https://unmtid-shinyapps.net/shiny/repodb/.

Our Solution

In our project, we wanted to use data about drugs and diseases' clinical trial status to build a recommendation system that will be able to predict possible new indications for existing drugs. Our data was very sparse, so we wanted to be able to calculate the similarity between users and the items based on something other than their ratings. Thus, we first created a representation of the drugs and the disease based on publicly available data from the internet. We created two kinds of profiles, one based on medical websites, and the other based on the relevant Wikipedia pages. Then, we used those profiles in our recommendation system. We built our recommendation system using user-user collaborative filtering and item-item collaborative filtering based on the profiles we created, where the diseases are our users and the drugs are our items.

Building Drugs and Disease Profile

We were interested in what kind of data would capture the diseases and the drugs better and would help us create a good representation of them. Thus, we decided to create two kinds of profiles, and compare their ability to capture the essence of the drug or disease.

Medical information Based Profiles

In our database downloaded from repoDB, we had for each drug its drugbank ID and for each disease its NIH concept ID. We used those to get automatically to the relevant URL pages. For the drugs, we took information from drugbank (DrugBank Online), and for diseases we used NCBI's medgen (Home - MedGen - NCBI). When we used web scraping to extract information from the HTML files, we saw that there were big differences within the information found for diseases and within the information found for drugs- some had fields in their HTML files that were missing for others. We thought that having unbalanced profiles, some containing much more information than others, would introduce bias into our results when we will need to calculate the distance between profiles for the recommendation system: well studied drugs or diseases will not have empty values, and thus will be closer to each other than not well studied drugs or diseases that will have a lot of empty values. Thus, we decided to take only the first fields from the drugbank pages, that existed for most drugs. Those fields included Generic Name, Brand Names, DrugBank Accession Number, Summary, Background, Weight and Chemical Formula.

For the diseases, we extracted from the medgen page the disease type, its description, and the names of papers listed on the page for that disease.

For both drugs and diseases, we got fields with a lot of free text that we had to represent mathematically to be able to calculate distances later. We wanted to use tf-idf representation, hoping it would capture the disease's or drug's important aspects. Because our profiles were still sparse, we were worried that this would introduce bias when we calculated distances, and thus we decided to concatenate all of the free-text fields in the profile and run tf-idf on the concatenated information. In addition, we calculated distance based on the other fields that were categorical and numerical- the Type variable for the diseases and the weight of the drugs. This way we created profiles based on both free-text and categorical and numerical information. We were able to create profiles for 1402 diseases out of 1462 diseases in the repoDB dataset, and for 2278 drugs out of 2381 drugs in the dataset.

Wikipedia Based Profiles

We used the wikipedia package and the drug or disease name to retrieve the content of the Wikipedia page about the drug or disease. Then, we saved it to a JSON file, and we created our profiles from the JSON files by running tf-idf on their content. While working on the Wikipedia files, we saw that there is a huge variance between the amount of content that different drugs or diseases had. We think that this variance might cause a bias toward well studied drugs and diseases, which will make it harder to compare the different diseases and drugs. We ran into that problem also when we created the medical websites based profiles, but here we decided to keep all the information we had on each drug or disease. That allowed us to utilize all the data we had, but kept potential bias in our profiles. We wanted to see which approach would do better- removing information that existed only for some of our drugs and diseases, as we did in our medical websites based profiles, or keeping all the data, as we did while building those profiles.

Recommendation System

Recommendation systems are based on ratings. We used the drug statuses as ratings, and we had to find a way to convert them to numerical values. We read about the different stages of clinical trials, and decided on the following conversion:
Approved: 1, Unrated: 0, Withdrawn (Phase 1): -1/2, Withdrawn (Phase 2): -2/6, Withdrawn (Phase 3): -1/6, Suspended (Phase 1): 1/2, Suspended (Phase 2): 4/6, Suspended (Phase 3): 5/6, Terminated (Phase 1): -1, Terminated (Phase 3): -5/6, Terminated (Phase 2): -4/6. After reviewing our initial results, we decided to change the rating of Approved to 2, figuring that there is a major difference between drugs that are suspended at phase 3 and drugs that are approved, which we first underestimated. That change improved our recall, while not significantly changing our precision.
We divided our data to train and test sets, and predicted ratings using item-item and user-user collaborative filtering. For each disease in the test set, we predicted based on the ratings of the k closest items from the train set that the user has rated, or k closest users that had rated the item, based on the method we used. To find the k nearest neighbors, we used the medical based profiles and Wikipedia based profiles. We also predicted based on the rating matrix that we created from the repoDB data, which we considered as baseline. To decide the right value of k, we consulted our milestone visualization, which showed that few drugs treat a lot of diseases, but most drugs treat only a few drugs. Thus, we decided to check our recommendation system using a relatively low number of neighbors, k=5 and k=10. We used the true labels of the test set in our evaluation of the recommendations.

Evaluation

Evaluating the generated Profiles

We wanted to check if our generated profiles captured the essence of the drugs and the diseases, and if drugs that are similar based on our profiles treat similar groups of diseases. Because each drug treats a subset of the diseases, simply clustering and coloring by diseases could not capture the whole picture. Thus, we first created clustered heatmaps of diseases across diseases and drugs across drugs based on the distance between the profiles we generated **(Figure 1)**. This allowed us to see the clusters the diseases and the drugs fell into.

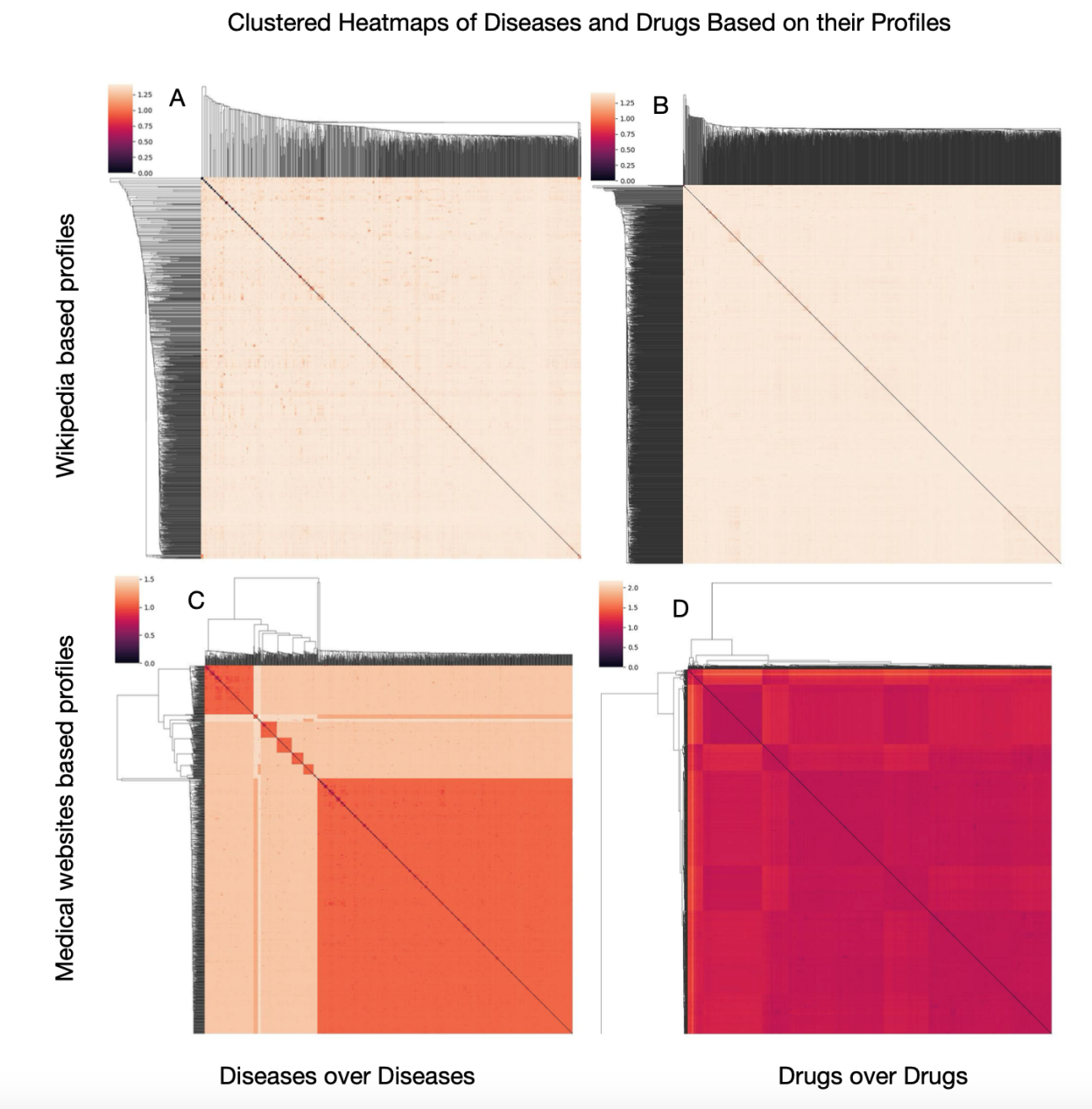

Figure 1. Clustered heatmaps of diseases and drugs based on their profiles

We see that the different profiles led to different results. It seems that clustering by the profiles created using medical data **(C, D)** gave better separation to clusters than clustering by the Wikipedia based profiles **(A, B)**.

Then, we took the order of the diseases and the order of the drugs from the clustered heatmaps and created a new color map plot of drugs across diseases. In the color map we colored each cell by the status from the repoDB database, meaning by the status of the drug regarding the disease in clinical trials. The status could be Approved, Terminated, Suspended, or withdrawn, and the phase in which the trial stopped if the drug wasn't approved.

We chose this visualization because we expected that if our profiles indeed are a good representation of the drugs and the disease, we will see dense square-like areas that will show that similar drugs are used to treat similar diseases.
We did this evaluation for both medical data based and Wikipedia based profiles separately. Our results for both profile types **(Figure 2)** did not show the clear dense square-like area we hoped to see. Some areas are denser and show that similar drugs do treat similar diseases, but it is not as clear as we hoped to see. We could not predict based on our visualization which profile type would lead to better predictions.

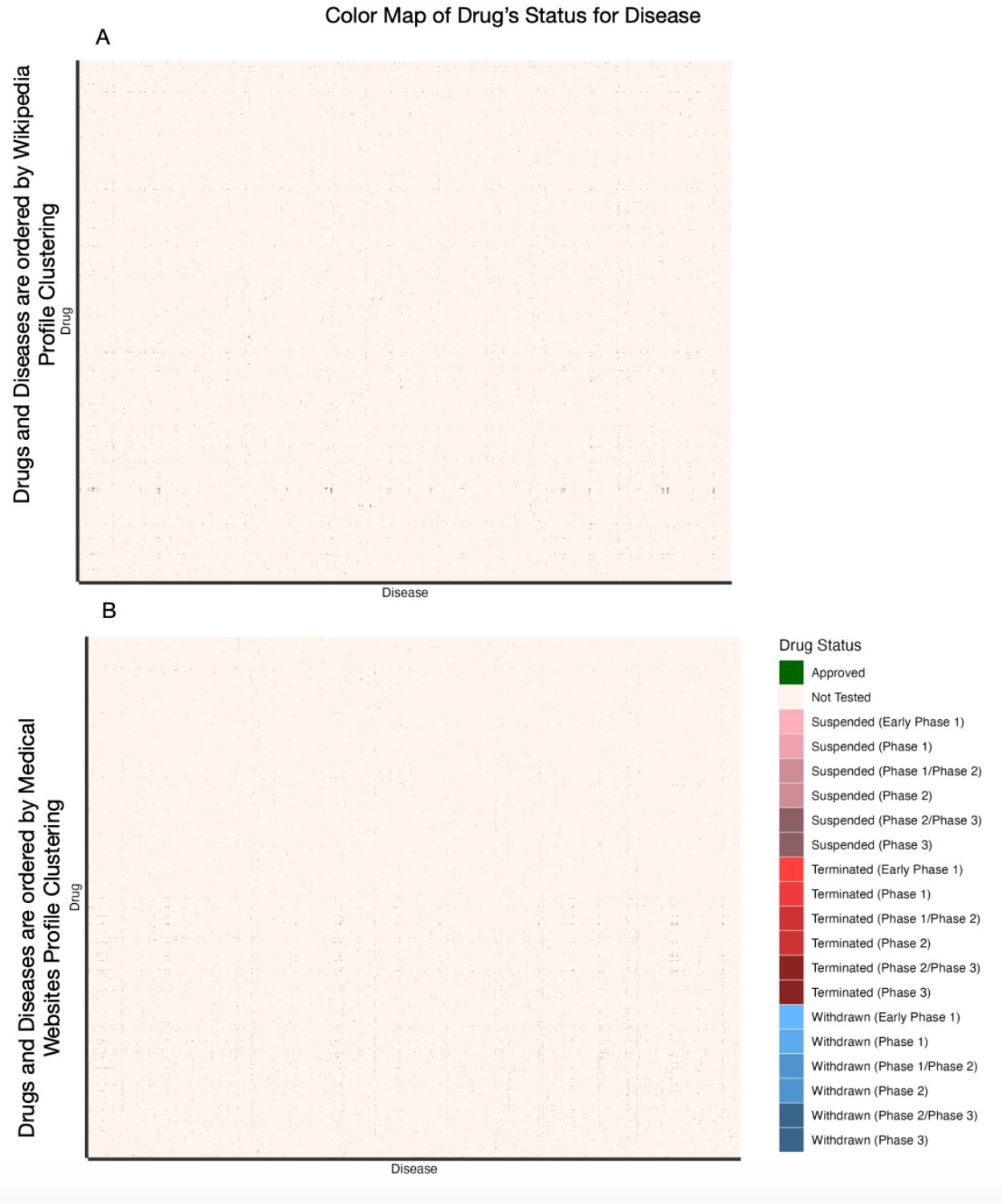

Figure 2. Color map of drug's status for disease

The order of the drugs and the diseases in the color maps is by the order of the clustered drugs and diseases. In **(A)** the clustering was done based on the Wikipedia based profiles, and in **(B)** it was done based on medical websites based profiles.

Evaluating the Performance of Our Recommendation System

When evaluating the performance of our recommendations on the test set, we were interested both in our overall performance, and more significantly in the quality of our predictions for high rated drugs. We are more interested in correctly predicting drugs that will be approved because our goal in this project is to find possible candidates for drug repurposing. Thus, we measured misclassification error to assess our overall success, and also precision and recall to assess our performance for approved drugs. When calculating precision and recall, we regarded the Approved status as positive and all other statuses as negative. In the recall-precision tradeoff, we preferred to improve the recall over precision. That is because our system recommends potential drugs for repurposing, and the recommended drugs would later go through a more thorough inspection before starting clinical trials for them. Thus, not missing good candidates is more important than being very confident in our candidates, and better recall is preferred.

We calculated those metrics for predictions generated by item-item and user-user collaborative filtering, based on the different profiles, and for predictions done based on the rating matrix, that we regarded as baseline **(Figure 3)**.

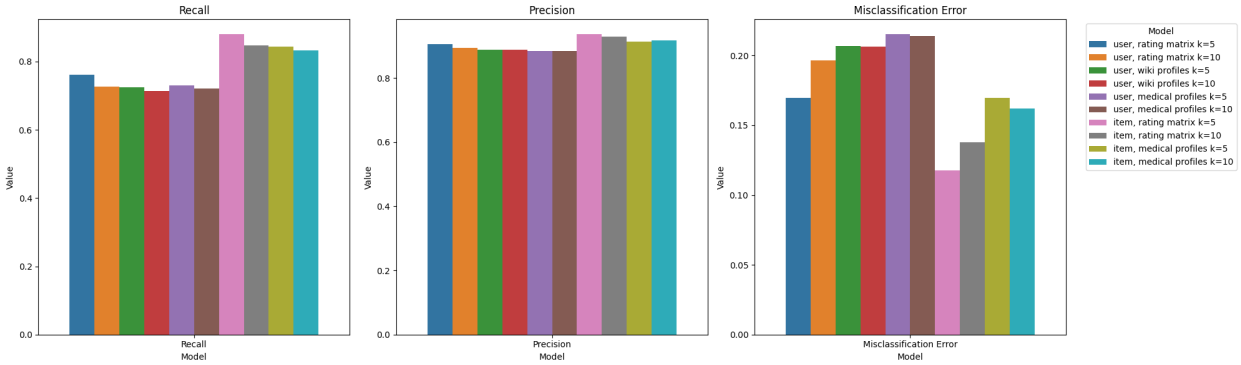

Figure 3. Recall, precision and misclassification error for predictions calculated using different methods

We compared the three metrics for predictions generated using item-item collaborative filtering, and user-user collaborative filtering. We compared predictions calculated based on both types of profiles and on the rating matrix. We also compared predicting with two values of k, k=5 and k=10.

First, we see that regardless of the method used to calculate the distance, item-item collaborative filtering based predictions are better than user-user. That matches what we expected, because the diseases are more complicated than drugs, similar to people and movies in Netflix's recommendation system, making the diseases harder to capture in profiles. Within each method of prediction, the results were similar for both types of profiles and the baseline, and to our surprise, using our profiles gave worse results than the baseline. We suspect that the similarity of the results within a method is due to the sparsity of our data: not only did our users not rate a lot of items, but our items were not rated but a lot of users. That led to a situation where when we looked for the k closest diseases that rated the drug, we found a small number of diseases, and we chose a similar group of diseases regardless of the way that we calculated distance. Similarly,

when looking for the k closest drugs rated by a disease we find a small group of drugs. That can also explain why using k=5 is slightly better than k=10: it allows some differentiation between more similar diseases to not so similar ones. Overall, we did get good results, and we think that using a recommendation system to predict new candidates for drug repurposing can lead to the discovery of new treatments.

Future Work

In our work, we found how complicated medical data is. Many sources contain different kinds of information, such as drug side effects, targets, disease-related genes, and more. Most of that information is not in a Machine Learning compatible format, and needs some processing before it can be used- similar to the free text fields in our profiles. Another problem we encountered with medical data is that we have very different amounts of data for different diseases and drugs, which makes comparing them a lot more difficult. We believe that if medical information from various sources can be combined and transferred to a Machine Learning compatible format, we will be able to create much more comprehensive profiles that will lead to better recommendations and improve our ability to repurpose drugs.

One small example we encountered in our work that poses an interesting question is how we can measure the distance between chemical formulas. When searching for ways to calculate it, we found interesting ideas worth further exploring, for example using the Tanimoto coefficient. We think that better use of medical information could improve our recommendations beyond the baseline.

Brief conclusion

Recommendation systems are a powerful tool, that can help repurpose drugs. Our results here showed that even the data on what drug is currently used to treat what diseases can predict good candidates for clinical trials. We also saw that using data from medical websites and Wikipedia, which is mostly free-text, does not improve the recommendation above the baseline of using the rating matrix as we hoped.