# Drug Repurposing for Rare Orphan Diseases Using Machine Learning Techniques

**Rajesh Manicavasagam, Prabin B Lamichhane, Prajjwal Kandel, Douglas A. Talbert**
Department of Computer Science, Tennessee Tech University, Cookeville, TN
rmanicava42@tntech.edu, pblamichha42@tntech.edu, prkandel42@students.tntech.edu, dtalbert@tntech.edu

## Abstract

New drug discovery is a time-consuming and costly process. Several drugs have been in clinical trials for a very long period. Finding a new target for existing medications can be an effective strategy to reduce the lengthy and costly drug development cycle. Drug repurposing (or repositioning) is a cost-effective approach for finding drugs that can treat diseases for which those medications are not currently prescribed. Drug repurposing to treat both common and rare diseases is becoming an attractive option because it involves using already approved drugs. Through drug repurposing, we can identify promising drugs for further clinical investigations. This paper presents machine learning techniques for drug repurposing to find existing drugs as an alternate medication for other diseases through drug-drug, drug-genes, drug-enzymes, and drug-targets interactions. We develop a model to find similar drugs that can treat similar diseases. We then use the model to predict potential candidate drugs for rare orphan diseases.

## Introduction

Drug discovery is a time-consuming and costly process. By conservative estimates, a single drug takes about 15 years and more than 800 million dollars to get approved. Because the cost and time for new drug development are so high, it is estimated that for every dollar spent on the research and development of a new drug, less than a dollar of value is returned on average (Berger et al. 2014). Thus in recent years, drug repurposing is becoming more popular. Drug repurposing (or repositioning) is a cost-effective approach for finding drugs that can treat diseases for which they are not currently prescribed. Using drug repurposing as a basis for drug discovery can reduce the search space and expedite the discovery process. There are several advantages of drug repurposing over the traditional drug discovery process. First, since the repurposed drug has already been clinically approved and found to be sufficiently safe in humans, it is unlikely to fail from the safety point of view. Secondly, the cost of repurposing is low with minimum investment. (Pushpakom et al. 2019).

Computational methods are becoming a popular choice for drug repurposing. Computational methods require extensive knowledge about a drug like its chemical composition, gene expression, side-effects, the interaction between a drug and its associated target, relationships between targets, and relationships between those targets and diseases for drug repurposing. Although it is difficult to understand the overall picture because of heterogeneous information sources and data unavailability (March-Vila et al. 2017), computational methods are still helpful for identifying promising drugs for further clinical investigations.

In this work, we use a similarity-based technique with different machine learning algorithms to discover potential drug-target interaction. First, we use different drug information like side effects, gene expression, enzymes, etc., to find similar drugs. Then, based on the notion that similar drugs may treat similar diseases, we focus on a smaller group of similar drugs to experiment with different machine learning models. After that, we select the best machine learning (ML) model based on performance metrics. Finally, we use the best model to predict potential existing drugs that might treat some rare orphan diseases.

The rest of this paper is organized as follows: the next section presents related works on computational methods for drug repurposing. Then, in other sections, we discuss the data sets used for experiments, experimental setup, performance analysis, conclusion, and future works.

## Related Works

Since drug repurposing is an attractive option for drug discovery or for finding drug substitutes, there is prior work in this field. In addition, research involving computational methods are becoming popular because they are less expensive, and the experiments are fast. Therefore, this section discusses some existing research for drug repurposing using computational methods.

There are different computational methods for drug repurposing like *molecular docking*, *network-based mapping*, and *machine learning-based methods* (Pushpakom et al. 2019). *Molecular docking* is a structure-based computational strategy that predicts the binding of a ligand (for example, a drug) with a target protein and exploits the prior knowledge of drug-target interaction to find potential interaction of drug with a particular target. While there has been some work using this method (Dakshanamurthy et al. 2012), there are issues with it. For example, getting 3D structures of protein targets of interest is not easy.

*Network-based methods* involve creating networks of various entities associated with drugs and targets. In these models, the nodes in the network represent either drugs, diseases, targets, side effects, or gene expressions, and edges represent the interactions (or relationships) between them. Among the works based on network-based methods, Chen and Liu (Cheng et al. 2012) computed drug-based similarity, target-based similarity, and network-based similarity to predict drug-target interaction.

*Machine learning-based methods* typically use feature extraction and model fitting and evaluation. Several works using machine learning techniques for drug repurposing are reported in the literature. Machine learning-based methods can also be divided into feature vector-based approaches and similarity-based approaches (Ding et al. 2013). In feature vector-based approaches, feature vectors are generated using information like drugs' chemical descriptors, target sequences, etc. Any standard machine learning model can be used to predict drug-target interaction. Similarity-based methods typically integrate drug-drug similarity and target-target similarity information into features and use standard machine learning algorithms for predicting drug-target interactions. PREDICT (Gottlieb et al. 2011) is a similarity-based machine learning framework that applies logistic regression to features based on drug-drug similarity (using drug-protein interaction, sequence, and gene-ontology) and disease-disease similarity (using disease-phenotype and human phenotype ontology).

## Data

We use the drug and drug-related information from heterogeneous sources for our experiments. For example, we use drug-target interactions, drug-gene associations, drug-enzyme associations, and drug-side effect information from the different sources as mentioned in Table 1.

Table 1: List of sources for data related Drug, Target, Gene, Side Effects and Enzyme

| Type | Source | Data |
|---|---|---|
| Drug-Target | DrugBank | 12,148 |
| Drug-Gene | DGIdb | 22,000 |
| Drug-Side Effects | Stanford | 433,000 |
| Drug-Enzyme | NCBI | 4,300 |

The DrugBank database (Wishart et al. 2017) is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information. We deduce drug-target relationship data from the DrugBank database from 12,148 drug entries, including 2,556 approved small molecule drugs, 1,285 approved biotech (protein/peptide) drugs, 130 nutraceuticals, and over 5,865 experimental drugs.

Drug-gene interaction data were obtained from DGIdb (Coffman et al. 2017). DGIdb provides links between genes and their known or potential drug association. In addition, the Stanford Digital repository (Stanford Digital Repository 2005) was used to obtain drug-side effects data.

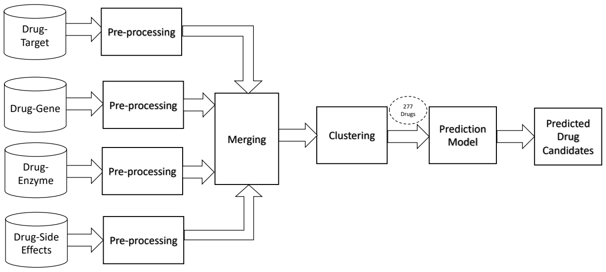

Figure 1: Architecture of Experimental Setup

Drug-enzyme interaction data were obtained from NCBI (National Library of Medicine 1996) (National Center for Biotechnology Information). The National Center for Biotechnology Information provides access to biomedical and genomic information. We also used the ChemBL (Gaulton et al. 2016) database to review genomic data and chemical compounds information.

Orphan data was used to collect data related to different categories of rare diseases and their associations with genes. The Orphanet Rare Disease Ontology (ORDO) (Kibbe et al. 2014) was jointly developed by Orphanet and the European Bioinformatics Institute (EMBL-EBI). It provides information on rare diseases like relationships between diseases, genes, other relevant features, etc. and serves as a useful resource for the computational analysis of rare diseases. Table 2 shows the number of rare diseases that were chosen for testing the trained model.

Table 2: List of rare diseases collected from Orphan data

| Disease class | Data |
|---|---|
| Eye-related | 30 |
| Gene-related | 16,253 |
| Lung-related | 131 |
| Abdomen-related | 193 |
| Skin-related | 766 |

## Experimental Setup

Our experiments trained and compared different machine learning models to see which performed best. Then, we use the best model to predict drugs for rare orphan diseases. The architecture of the experimental setup is shown in Fig. 1.

### Data Pre-processing

The data collected from the different sources had more information than was needed. We had to filter them to get only the data related to drugs, targets, genes, enzymes, and side effects. Some filtered fields included drug synonym names for the Spanish language, dosage information, drug manufacturer information, and PubMed information. We also removed duplicate records. Any drug that has missing data was removed from the training set. DrugBank data was the primary source for matching drugs based on drug names obtained from other resources. The DrugBank ID was used as the primary identifier of any particular drug. After filtering

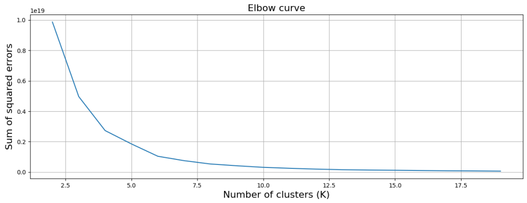

Figure 2: Elbow Plot

the data, we saved data from each source into a separate CSV file.

## Merging

This step merged the data from all the CSV files containing drug-target information, drug-gene associations, drug-enzyme associations, and drug-side effects information into a single CSV. We used a python-based Pandas (McKinney and others 2010) library for this task. At the end of this step, we had a single drug dataset file having targets, genes, enzymes, and side effects as its features.

## Clustering

Based on the notion that similar drugs might treat similar diseases, we sought similar drugs based on associated targets, enzymes, genes, and side effects. To find similar drugs based on those associations, we used clustering. We chose k-means as our clustering algorithm. We used the elbow method to identify the value of $k$ (number of clusters). The elbow plot obtained is shown in Figure 2. We chose eight as the number of clusters. After getting the clusters, we focused on only one of the clusters as each cluster represented collections of similar drugs based on targets, enzymes, genes, and side effects. We chose the cluster with the maximum number (277) of drugs. We used this cluster (277 drugs) to experiment with different machine learning models.

## Prediction

We split the dataset obtained from the clustering into training and test datasets. Then, we trained five different prediction models: Decision Trees, Random Forest, Support Vector Machines, Naive Bayesian classifier, and K-Nearest-Neighbor on target, gene, side effects, and enzyme features to predict the drugs they associate. Next, we used 10-fold cross-validation to assess the predictive performance for each model. Third, we used the test data to calculate the performance score. Fourth, we compared the test results of these models using different evaluation metrics and chose the best model as our prediction model. Last, we used this best model to predict which existing drugs are candidates to treat rare diseases. While predicting, we repeated our experiment many times for each rare disease and kept records of the predicted drugs. Later, we considered the top 5 most frequently predicted drugs as candidate drugs for that rare disease.

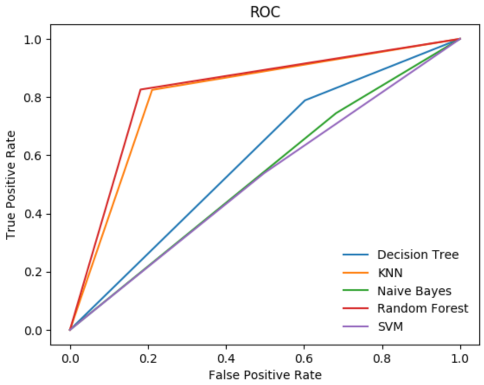

Figure 3: Comparison of AUROC

# Performance Analysis

## Evaluation Metrics

The evaluation metrics typically used in the analysis of ML models are recall, precision, F1-score, and area under the ROC curve. We use the same for our experiments. In general, high values of recall, precision, F1-score, and area under the ROC curve are desired. Similarly, we plot True Positive Rate against False Positive Rate to get the Receiver Operating Characteristic (ROC) curve. The area under the ROC curve (AUROC) is a standard measure of comparison between models.

## Experimental Results

We experimented with different ML models, and the results for all models are shown in Table 3. As shown in Table 3, Random Forest outperformed all other models in all four evaluation metrics.

Table 3: Comparison of classifiers

| **Classifier** | **Precision** | **Recall** | **F1-Score** | **AUROC** |
|---|---|---|---|---|
| *Decision Tree* | 0.610 | 0.593 | 0.579 | 0.593 |
| *Random Forest* | **0.827** | **0.826** | **0.826** | **0.827** |
| *K-NN* | 0.809 | 0.809 | 0.809 | 0.809 |
| *Naive Bayes* | 0.537 | 0.532 | 0.508 | 0.532 |
| *SVM* | 0.520 | 0.520 | 0.520 | 0.520 |

The Receiver Operating Characteristic (ROC) curves for all five models are shown in Figure 3. It can be seen that Random Forest and K-Nearest-Neighbors have higher AUROCs and outperform the other three models, which each have AUROC values near 0.5. Random Forest produced the highest value of AUROC with 0.827.

## Discussion

With Random Forest as the best prediction model from the evaluation metrics, we used it to predict associations between existing drugs and rare orphan diseases. We performed multiple prediction runs for each disease class and kept the records of predicted drugs. In Table 4, we present the five most frequently predicted drugs under each disease class. The drugs that are listed in Table 4 are not used for rare orphan diseases. Instead, our model predicted that these existing drugs might treat rare orphan diseases. For example, the drug ZIPRASIDONE is predicted as a candidate for rare orphan diseases related to the eyes. In today's practice, ZIPRASIDONE is used to treat bipolar disorder. So, the listed drugs are identified as candidate drugs that can be further investigated as potential treatments for rare orphan diseases.

Table 4: Predicted Drugs for Orphan Rare Diseases

| **Disease class** | **Predicted Drugs for Repurposing** |
|---|---|
| *Eye* | ZIPRASIDONE, CABERGOLINE, AMITRIPTYLINE, OLANZAPINE, CLOZAPINE |
| *Gene* | PONATINIB, NINTEDANIB, LENVATINIB, HEPARIN, FOSTAMATINIB |
| *Lung* | SORAFENIB, HEPARIN, REGORAFENIB, PONATINIB, LENVATINIB |
| *Abdomen* | METHYSERGIDE, ROPINIROLE, LISURIDE, KETAMINE, PIPOTIAZINE |
| *Skin* | YOHIMBINE, LOFEXIDINE, NICARDIPINE, CABERGOLINE, NORTRIPTYLINE |

## Conclusion and Future Work

Drug repurposing is an attractive option for drug discovery and for finding substitutes for existing drugs. This paper describes a machine learning-based drug repurposing approach that first fuses heterogeneous information from various sources. Then, based on the notion that similar drugs might treat similar diseases, it constructs drug clusters based on their targets, enzymes, genes, and side effects. The best prediction model is then identified. In our experiments, Random Forest outperformed the other models. Lastly, the best model is used (i.e., Random Forest) to identify potential existing drugs for rare orphan diseases.

This work only considered information from four sources: drug-target, drug-side effects, drug-enzymes, and drug-gene. Integrating data from more sources like protein-protein interactions, 3-D chemical structures of drugs, etc., could improve the accuracy of predictions. Similarly, we only experimented with a small cluster of drugs obtained from k-means clustering. We could, in the future, use different clustering techniques and experiments on a more significant data set as well.

While the results obtained are preliminary, this paper demonstrates how machine learning techniques use for drug repurposing. We anticipate this work will open an approach for drug repurposing by fusing information from heterogeneous information sources.

## References

Berger, A. C.; Olson, S.; Johnson, S. G.; Beachy, S. H.; et al. 2014. *Drug repurposing and repositioning: workshop summary*. National Academies Press.

Cheng, F.; Liu, C.; Jiang, J.; Lu, W.; Li, W.; Liu, G.; Zhou, W.; Huang, J.; and Tang, Y. 2012. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS computational biology* 8(5):e1002503.

Coffman, A. C.; Wollam, A.; Spies, G.; Cotto, K. C.; Spies, N. C.; Kiwala, S.; Feng, Y.-Y.; Wagner, A. H.; Griffith, M.; and Griffith, O. L. 2017. DGIdb 3.0: a redesign and expansion of the drug–gene interaction database. *Nucleic Acids Research* 46(D1):D1068–D1073.

Dakshanamurthy, S.; Issa, N. T.; Assefnia, S.; Seshasayee, A.; Peters, O. J.; Madhavan, S.; Uren, A.; Brown, M. L.; and Byers, S. W. 2012. Predicting new indications for approved drugs using a proteochemometric method. *Journal of medicinal chemistry* 55(15):6832–6848.

Ding, H.; Takigawa, I.; Mamitsuka, H.; and Zhu, S. 2013. Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Briefings in bioinformatics* 15(5):734–747.

Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; et al. 2016. The chembl database in 2017. *Nucleic acids research* 45(D1):D945–D954.

Gottlieb, A.; Stein, G. Y.; Ruppin, E.; and Sharan, R. 2011. Predict: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology* 7(1):496.

Kibbe, W. A.; Arze, C.; Felix, V.; Mitraka, E.; Bolton, E.; Fu, G.; Mungall, C. J.; Binder, J. X.; Malone, J.; Vasant, D.; et al. 2014. Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic acids research* 43(D1):D1071–D1078.

March-Vila, E.; Pinzi, L.; Sturm, N.; Tinivella, A.; Engkvist, O.; Chen, H.; and Rastelli, G. 2017. On the integration of in silico drug design methods for drug repurposing. *Frontiers in pharmacology* 8:298.

McKinney, W., et al. 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, 51–56. Austin, TX.

National Library of Medicine. 1996. National center for biotechnology information. `https://www.ncbi.nlm.nih.gov/`.

Pushpakom, S.; Iorio, F.; Eyers, P. A.; Escott, K. J.; Hopper, S.; Wells, A.; Doig, A.; Guilliams, T.; Latimer, J.; McNamee, C.; et al. 2019. Drug repurposing: progress, challenges and recommendations. *Nature Reviews Drug Discovery* 18(1):41.

Stanford Digital Repository. 2005. Stanford digital repository. `https://library.stanford.edu/research/stanford-digital-repository/`.

Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. 2017. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research* 46(D1):D1074–D1082.