IJCSN - International Journal of Computer Science and Network, Volume 8, Issue 3, June 2019
ISSN (Online) : 2277-5420
www.IJCSN.org
Impact Factor: 1.5

# Semi-Supervised Drug Repositioning Framework based on Drug, Target, and Disease Fingerprints

**Eman Ismail**

Computer Science Department, Faculty of Computers and Information
Helwan University, Cairo, Egypt.

**Abstract -** Drug Repositioning makes a significant contribution to industry and research due to its ability to reduce the time and cost of drug discovery through making use of the existing drugs. At the time of writing this research, many computational methods have been proposed; however, few of them were able to integrate chemical space (drugs) and genomic space (targets) with disease space. In addition, using the feature-based method in Drug Target Interaction (DTI) and Target Disease Interaction (TDI) models are not well-exploited. Hence, developing an efficient approach in order to predict potential DTI and TDI is necessary. In this research, we introduce an integrated computational framework to predict potential interactions of drug-target and target-disease basing on features extracted from drugs, targets, and diseases using various learning methods (e.g., Random Forest, Decision Trees, Logistic Regression).

***Keywords -*** *Drug Repositioning, Random Forest, Feature-based, Drug-Target Interaction, Target-Disease Interaction*

## 1. Introduction

One of the main barriers that confront the pharmaceutical industry is the low productivity of drugs [1]. Over the past decade, the rate of newly approved drugs shipped to the market has been barely increasing (only 55 new molecular entities approved in 2018 [2]). Developing new drug costs more than $1.8 billion, and the average time to reach the market is 12-15 years [3]. The life-cycle of a new therapeutic drug is very long and expensive for which some methods are introduced to reposition the treated disease of an existing drug by acting on multiple targets, which is called drug repositioning. Existing clinical history and toxicology information profoundly aids the repositioning process, which results in investing relatively reduced costs and time [4], [5]. Drugs are small molecules that bind with biomolecular targets to activate or inhibit their functions for treating a specific disease. When a drug interacts with multiple targets, it is called a multi-target drug. Similarly, when a target interacts with multiple drugs, it is called a multi-drug target. Traditional drug discovery relies on screening targets of chemical compounds to identify a small set of hits whose properties studied in further investigations, the succeeding compounds finally approved by the Food and Drug Administration (FDA); furthermore, the relationship is one-drug-one-target [6], that is very limited and does not consider the complexity of relationships.

Computational approaches are one of the crucial solutions to expediting the screening of chemical compounds [3]. Although there exist numerous highly-matured databases for biological classes (i.e., chemical compounds, target proteins, and diseases) and many investments spent yearly on drug design, drug candidates still have a high-failure opportunity in the drug discovery pipeline due to the safety conditions or failed clinical trials. That is why the computational approaches designed for drug repositioning are becoming more attractive. Drug repositioning is an area of interest in drug discovery and design as drug-target interaction and target-disease interaction identification in wet experiments is very time-consuming and costly. Drug repositioning bypasses the tests conducted for drugs in the development cycle since it processes on existing tested drugs that have optimized affinity, efficacy, selectivity, and safety [3]; as a result, it is the potential solution to accelerating the drug development cycle. In spite of de facto repositioned drugs form 30% of newly shipped drugs, and the success fulfilled in recent years, most of the repositioned drugs based on an ad-hoc clinical observation or unfocused screening [7].

A lot of computational approaches introduced in drug repositioning or target prediction, but most of them do not differentiate between target prediction and drug repositioning tasks, also their methods can be used interchangeably. Although the drug-disease association is complex, separating them into drug-target and target-disease can provide the chance to study more intuitive connections [3].

## 2. Background and Related Work

In Silico, prediction of the interactions (mentioned in Section I) is costly and time-consuming due to the need for specific materials experimented with different combinations. Drug repositioning task separated in many research studies; however, integrating target information into drug repositioning resulted in more meaningful predictions [3]. Targets are the link between drugs and diseases, which assumes that a drug's therapeutic effect on the disease is perceived if the drug interacts with the biological target [3]. Drug Repositioning strategies can be divided to:

- Similarity-based methods [8], [9], [10] are the most common strategies in drug repositioning [11]; their prediction highly depends on how close are data spaces to each other, which limit the novel findings [12];
- Network-based is another related approach that exploits data from different biological sources and connections among each source class; it also depends on data in nearby space and the way network constructed [3];
- Machine learning approaches (mainly, supervised learning) work on extracting features with the topmost impact of each biological class and produce an input vector from which correlation among them inferred [13], [14].

Two more strategies, yet not very common are: the Connectivity Map (CM) approach proposed by Lamb et al. [15], which exploits the gene expression signatures to connect new drugs, genes, and diseases; the in-silico method submitted by Cheng et al. [16] that uses the crystal structure of the target binding site to identify the more druggable targets beforehand experiments. Li et al. [17] predicted the new association between drugs and diseases by relying on the "guilt by association" methodology. A network-based approach by Keiser et al. [8] is employed to predict new targets for known drugs using chemical structures of the drugs and their main targets, after which the drug-target network used to predict a new indication for the approved drugs. Yu et al. [10] introduced a similarity-based method which adapts the bipartite graph for applying drug repositioning, showing that better results achieved by combining the chemical structure of the drug and the target profile. Another use of the bipartite graph by Bleakley et al. [18], however, wrapped with supervised learning to predict unknown drug-target interactions through two steps: initially, by predicting target proteins of a given drug, then predicting the drugs interacting with a given protein, these two steps are eventually combined to give a final prediction for each drug-target interaction basing on Support Vector Machine (SVM). Wang et al. [19] proposed a way by using the Restricted Boltzmann Machine which integrates multiple types of drug-target interactions to predict unknown relationships or drug modes of action; impressively, the method incorporates more information about drug-target relationship and drug mode of action which improves the accuracy. The pairwise input neural network (PINN) by Wang et al. [20] was able to show an advantage over the pair-support vector machine that maps targets and ligand features into higher dimensional space, where the PINN does not; additionally, not all parts of the target involved in the target-ligand interaction training process.

### 2.1 Methods of Drug Target Interaction (DTI)

As defined before in Section I, drugs are the substance of molecule compounds which alter the biological behavior of target proteins [21]. On the other hand, targets are a class of molecular structures that could interact with drugs. Known targets so far are only a few hundreds, but there are many predicted targets that might be able to bind to a drug-like molecule, the process of screening all potential targets molecules is a tedious task [19]. The identification of drug-target interactions is a challenging area in drug discovery and design because of the increasing rate of new chemical molecules which have the ability to be drug-like [9]. Many of the drug-target approaches were developed to predict the targets of new chemical entities; the others predict the new target of an already approved drug. Structure-based Drug Design (SBDD) and Ligand-based Drug Design (LBDD) are the most widely used. Structure-based drug design or Molecular Docking is a method that predicts the structure of the intermolecular complex formed between two or more constituent molecules [22], it is powerful and widely used that, disadvantageously, require 3D structures of both proteins and drugs to be available (which is not the case for 40% of drug target), even though it resulted in a good performance [4],[23].

On the contrary, ligand-based drug design can work with the absence of the 3D structures [24], [25]; it compares a candidate ligand with a known ligand of a target protein to predict its binding, using machine learning methods [9]. Performance of LBDD is not satisfying when the number of known ligands of specified target decreases.

Machine-Learning DTI can be similarity-based or feature-based [26] and the instances for them are [25], [27], [28].
In this study, we will use feature-based DTI as it is a more powerful way to detect novel interactions.

### 2.2 Methods of Target-Disease Interaction (TDI)

Finding a link between target and disease is costly and time-consuming, so developing a TDI model is of a great need [29]. The disease is an abnormal condition that changes the function of an organism negatively, observed by its manifestations or symptoms. Many approaches developed in studying the relationship between targets and diseases; examples include [30], [31] who introduced a network-based method that uses the similarity of genes to discover new interaction between genes and diseases. Another network-based method that builds the disease interaction network on gene association data, patient medical history, and diseases, integrating heterogeneous data into a multi-relational network, the result shows that using disease comorbidity can boost the relationship between genes and diseases [32]. Suratanee et al. [33] proposed another network-based method that integrates protein-protein interactions with gene-disease interaction to induce new relation between gene and disease. Zhao et al. [34] is a network-based method that uses gene expression and protein interaction to rank disease genes candidate by Katz-Centrality approach [35].

As discussed, most methods developed are network-based and part of them depend on the similarity scores which are limited due to network-construction method and lack of data.

We propose a feature-based method using disease manifestations and target domains as feature input vector using relatively sufficient data that are reliable to find novel interactions. One of the two close approaches to ours is Wangs' method [3], it proposes a method called Triple Layer Heterogeneous Graph-Based Inference (TL-HGBI) that is an extension to the HGBI method; the method based on a heterogeneous network by integrating drug repositioning and target prediction in the same framework to address the complex relationships between drug, targets, and diseases. More specifically, TL-HGBI calculates the strength between disease and drug iteratively using drug-target information; the similarity between the same nodes calculated to tell whether a pair of nodes could be linked when it exceeds a predefined threshold. The linkage of our work and theirs is that we both integrate information from multiple biological data sources (drug, target, and disease) to identify the complex relationship among them.

On the other hand, Tian et al. [28] work uses deep neural networks to introduce a method called Deep Learning for Compound-Protein (DL-CP), which recognizes the interactions between compounds and proteins; DL-CP was initially trained on a small subset of the data to obtain a set of hyper-parameters which are then applied to gain better results outperforming many other approaches. One of their findings is that using compounds' chemical structures and the targets domain can efficiently express their joint distribution.

## 3. Materials and Methods

In this study, we submit a machine learning framework for drug repositioning. The relationship between drug and disease divided into two principal direct relations: drug-target and target-disease as illustrated by Figure 1. Evoking how the therapeutic effect of a drug and its impact on the disease can be observed using the biological targets, we can infer new drug-disease reciprocities through predicting or using existing relations between diseases and targets which depend on finding new targets for the approved drugs (see Section II).

Considering that a lot of methods developed and many are still being developed; nevertheless, only a few were able to integrate methods from different spaces such as drug, target, and disease spaces. Our goal is to build a feature-based framework that simultaneously predicts how targets interact with drugs (DTI), and also predicting diseases which interact with those targets (TDI); both extensively discussed in Section II.

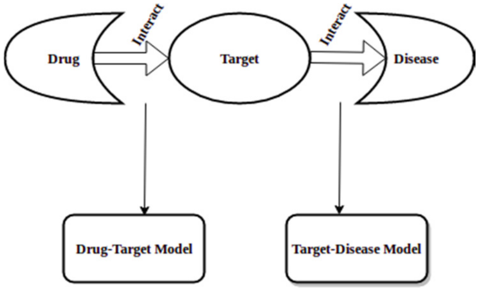

Fig 1. Two main drug repositioning approaches

Finally, the results of both models are a good start to shorten the discovery process cycle and decrease the failure rate. As the relationships between drugs, targets, and diseases are complex and we usually do not have a one-to-one relationship, i.e., a drug can interact with multiple targets, also the target can interact with multiple diseases (see Figure 1). To overcome the limitations of using similarity-based and network-based methods (as earlier shown in Section II) we use feature-based methods for both DTI and TDI models; moreover, integrating data from a different drug, target, and disease spaces showed more meaningful predictions and exploitation of intra-relations of drug-target and target-disease which also boosted the accuracy of the drug repositioning process [3].

### 3.1 One-Class Classification (OCC)Both DTI and TDI

Models bring up a one-class classification problem since only the positive class (e.g., a drug interacts with the target and target interacts with a disease respectively) in each is available. The one class classification is an area in machine learning used in the absence of the negative class. In the traditional binary/multi-class classification problems, observations from all classes are available, allowing you to examine all distributions; availability of only a subset of the classes causes the classifier to be biased.

Hence, OCC works as a class detector to give an unbiased prediction [38], [37]. The main concepts are:

- Positive-Unlabeled learning (PU): is a binary classification that relies on training on positive (P) and unlabeled (U) using a semi-supervised method. This type of OCC used widely in medical diagnosis and knowledge base completion [39], [40], [41].
- Novelty and Outlier Detection: is a one-class classifier that is trained to tell whether an observation should belong to the same existing class or considered as an outlier [42], [43].

Regarding our research, we depend on PU learning due to the existence of a multitude of unlabeled observations. In a nutshell, known drug-target interaction data represent the positive class, and all possible pairs of the drug-target interactions that do not exist within our dataset represent the negative class, and the same goes for the target-disease interaction data. We used Random Forest (RF) [44] and Decision Trees (CART) [45] for both DTI and TDI models. Also, we compared the results of RF and CART against Logistic Regression (LR) for the DTI model. we have a huge bulk of unlabeled data that should introduce more input for the models. Known drug-target and target-disease interactions are our positive data and random pairs that not in positive set are the unlabeled data.

### 3.2 Dataset collection

We initially collected the features of drug, target, and diseases from various data sources; secondly, extracted the unique ones in different ways; finally, get the known interactions data of drug-target and target-disease. Throughout the next subsections, we will be discussing how did we design the methods to do so.

1. Drug data: All drugs were obtained from the DrugBank database [46]. To profile each drug compound, we used Drug chemical structures as features and lastly constructed the corresponding fingerprint. Drug fingerprint is a 881-binary vector where 1 indicates the presence of certain chemical substructure and 0 otherwise. Drug fingerprint fetched from the PubChem database [47].
2. Target Data: For each target, we used its domains as features to represent the target profile. Target domain information was extracted from the Pfam database [48]. Target fingerprint is a 5133-binary vector with value 1 or 0, 1 indicates the presence of domain in feature space and 0 otherwise.
3. Disease Data: All diseases extracted from the DM-PatternUMLS database; it is a large-scale knowledge-base that presented the Disease-Manifestation relation from wide biomedical literature [49]. It is an accurate database from which inference of disease manifestation correlation to targets and drugs is accurate. Disease fingerprint is a 16096-binary vector where 1 indicates that the disease has a certain manifestation and 0 otherwise.
4. Drug-Target Interaction: collected from the DrugBank [46] which have 4 types of targets: proteins, carriers, transporters, and enzymes. However, limited to the associated targets in Pfam [48]. We also used the BindingDB [50] database as an extra database to enlarge interactions data.
5. Target-Disease Interaction: obtained from the DisGeNET data [51]. Interactions collected limited to the associated targets from Pfam [48] too, and the associated diseases from DMPatternUMLS [49].

### 3.3 Dataset Selection

For summary statistics of used datasets (see Table I).

Table 1:Biological Data sources

| *Data/Class* | *Data sources* | # Samples | *Total* |
|---|---|---|---|
| Drug | DrugBank<br>BindingDB | 8709<br>34 | 8743 |
| Target | DrugBank<br>BindingDB | 5161<br>12950 | 18111 |
| Disease | DMPatternUMLS | 11439 | 11439 |
| DTI | DrugBank<br>BindingDB | 22289<br>19718 | 42007 |
| TDI | DisGeNET | 361585 | 361585 |

### 3.4 Feature Extraction

Feature extraction is an essential step in developing a machine learning model, which helps to boost accuracy by extracting the most informative features. In our case, we dropped the zero variance features (i.e., all zeros or ones). The drug is an 881-binary vector, and the target is a 5133-binary vector, we fed both as an input to our DTI, overall the 6014-binary vector, only 2136 features were extracted; for the TDI, the disease is a 10963-binary, and only 11084 were extracted from the 16096 (i.e., the target features along with diseases'); both illustrated in Figure 2.

## 4. Experiments and Results

Purpose of both models is to find intra-relation of drug-target and target-disease that can be used furtherly to find a new drug-disease connection. We selected Stratified Sampling as our data splitting approach for both DTI and TDI models.

For the DTI data, we have 42007 positive samples and millions of possible random pairs labeled as negative samples. We randomly select a portion of the unlabeled pairs in three different ratios: 1,0.5,0.25 of the labeled data size, therefore, having three datasets (see Table II).

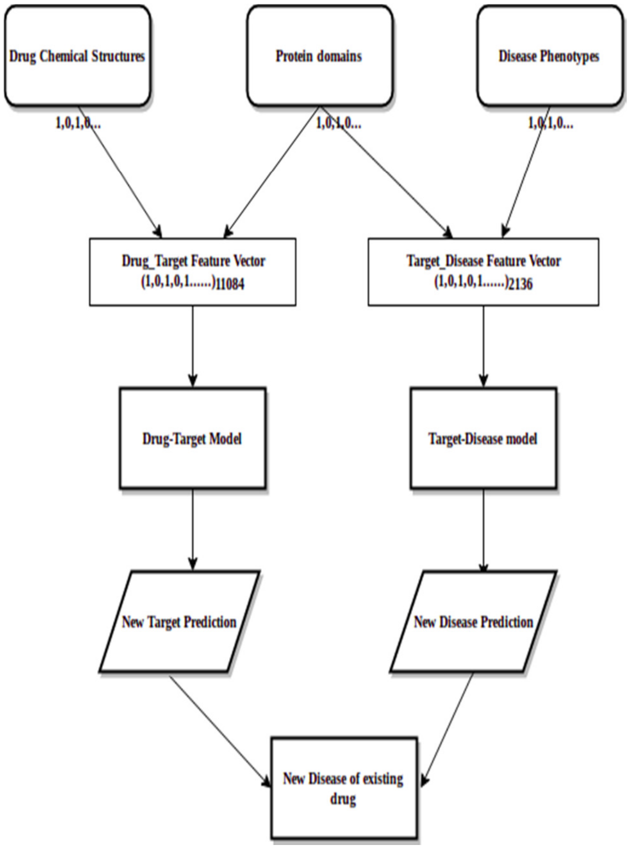

Fig 2. Proposed Method

Table 2:DTI Datasets

| *Dataset* | *DS1* | *DS2* | DS3 |
|---|---|---|---|
| Positive:Unlabel | 1:1 | 1:0.5 | 1:0.25 |
| # Positive Samples | 42007 | 42007 | 42007 |
| # Negative Samples | 42007 | 21003 | 10501 |
| Input Vector | 2136 | 2136 | 2136 |

For the TDI data, we have more than 360,000 positive examples and generated the negative examples the same way as DTI's, however, the ratios are 0.8, 0.4, and 0.2. Taking the same ratios as DTI's was impractical for the resources used.

Table 3:TDI Datasets

| *Dataset* | *DS1* | *DS2* | DS3 |
|---|---|---|---|
| Positive:Unlabel | 0.8:0.8 | 0.8:0.4 | 0.8:0.2 |
| # Positive Samples | 289268 | 289268 | 289268 |
| # Negative Samples | 289268 | 144634 | 72317 |
| Input Vector | 11084 | 11084 | 11084 |

### 4.1 Comparison among binary classifiers

In the One-Class Classification problem (OCC), any binary classifier could be used. We compared among three binary classifiers: Logistic Regression (LR) [52], Decision Trees (CART) [45] and Random Forest (RF) [44] for the DTI. Likewise, we compared between CART [45] and RF [44] for the TDI; LR not included due to the limitations of resources. The experiments showed that RF's accuracy is slightly higher than the CART for both DTI and TDI. LR's performance was poor when employed for the DTI.

For timing benchmarks, RF was dominating with the least training time. The accuracy results of both DTI and TDI models are shown in Figure 3 and 4, The assessment was carried only on the positive set, because unlabeled observations are, indeed, not guaranteed to be negative. Figure 5 and 6 represent average training time of both models.

## 5. Discussion

We have proposed a machine learning approach for drug repositioning depending on the relations between drug-target and target-disease. Basing on this framework, we have introduced two one-class classifiers drug-target and target-disease interactions. Experimental results exposed that RF outperforms other classifiers in accuracy and training time; it builds multiple decision trees which are then merged to get the prediction, that is obtained by traversing the merged trees and selecting the highest vote. Thus, selecting RF classifier was very intuitive due to the fact that data come from multivariate Bernoulli distribution, which represented as a binary decision tree [53].

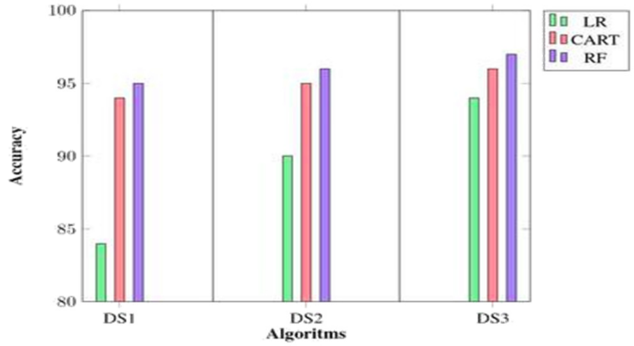

Fig 3. Avg Accuracy of DTI

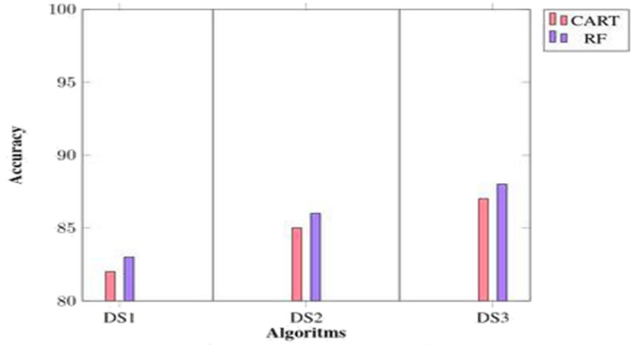

Fig 4. Avg Accuracy of TDI

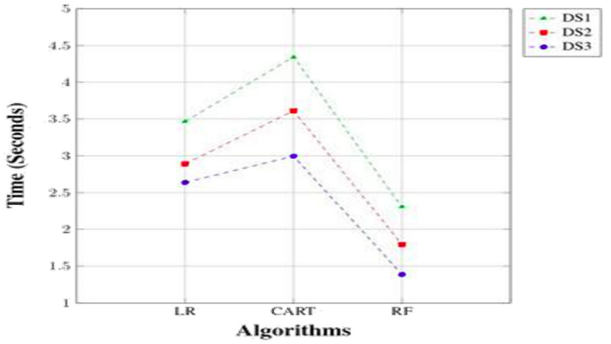

Fig 5. Avg Training Time of 3 datasets DTI

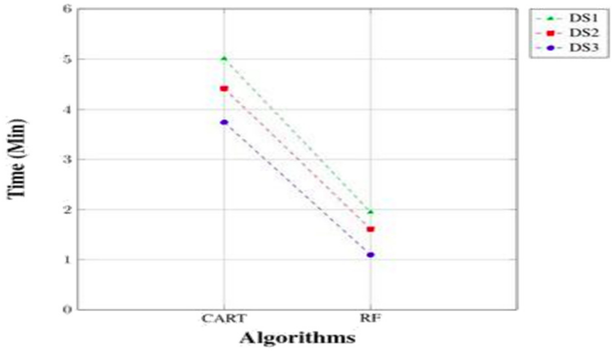

Fig 6. Avg Training Time of 3 datasets TDI

We compared the classifiers against three datasets, one balanced and the rest are imbalanced, to cover as many cases as possible.

To validate the generality of approaches used to infer interactions of drugs-targets and targets-diseases, we simulate two artificial datasets for assessing DTI and TDI models. In the assessment experiments, we depend on Copulas approach [54], [55], [56], [57], by using the simstudy package in R, which implements copulas to simulate data from different distributions. Firstly, we generated the marginal probability and correlation matrix of the dataset for TDI and DTI. Secondly, fed both to the package and tested models against the simulated data. Unfortunately, however, known, the generated data are not accurate since its distribution mapped from continuous to discrete distribution, and the package is limited in the case of correlated binary data, because correlation is biased downwards towards zero [58]. Future work includes applying more binary classifiers for the introduced models, especially to the Target Disease interaction (TDI); because we, originally, have a more rich dataset for the TDI, restricted computations and long training time prevented us from running such experiments. Also, using better assessment approach because we tested our classifiers for both DTI and TDI on only positive data even though the negative class included in the training process. As the copula method does not address the simulation of correlated binary data in an efficient manner, we are currently working on a far more stable package which has already shown promising results.

## 6. Conclusions

For enhancing the drug repositioning problem, we involved the target as a link between drug and disease classes; most of the developed approaches do not address the problem as two relations: drug-target and target-disease. Conducted experiments revealed that involving the target information boosts the performance relatively. The two models we defined, i.e., the Drug Target Interaction (DTI) and Target Disease Interaction (TDI), showed that the target correlated with both the drug and disease. In addition, applying the Positive-Unlabeled (PU) approach to obtain distribution from the unlabeled space caused our models to be unbiased towards the positive predictions. Using the feature-based approach for the DTI and TDI models was an efficient solution to overcome the limitations practiced in the similarity and network approaches. Although Drug Repositioning is an efficient way to shorten the drug discovery process, we still need, not surprisingly, input from expertise in biochemistry to validate our findings.

### Acknowledgments

The author would like to thank Andrew A. Naguib for proofreading of earlier versions of the manuscript and his efforts in conducting the experiments. Also, Dr.Olivier Bodenreider for his contribution to collect the dataset.

## References

[1] D. Cook, D. Brown, R. Alexander, R. March, P. Morgan,G. Satterthwaite, and M. N. Pangalos, “Lessons learned from the fate of AstraZeneca’s drug pipeline: A five-dimensional framework,” Nature Reviews Drug Discovery, vol. 13, no. 6, pp. 419–431, 2014.

[2] K. Sharma, “CDER New Drugs Program: 2018 Update,” p. 22, 2018.

[3] W. Wang, S. Yang, X. Zhang, and J. Li, “Drug repositioning by integrating target information through a heterogeneous network model,” Bioinformatics (Oxford, England), 2014.

[4] W. Baalawi, O. Soufan, M. Essack, P. Kalnis, and V. B. Bajic, “DASPfind: new efficient method to predict drugtarget interactions,” Journal of Cheminformatics, vol. 8, 2016.[Online].

[5] L. Perlman, A. Gottlieb, N. Atias, E. Ruppin, and R. Sharan, “Combining Drug and Gene Similarity Measures for Drug-Target Elucidation.”

[6] J. T. Dudley, E. Schadt, M. Sirota, A. J. Butte, and E. Ashley, “Drug discovery in a multidimensional world: Systems, patterns, and networks,” Journal of Cardiovascular Translational Research, vol. 3, no. 5, pp. 438–447, 2010.

[7] P. Zhang, F. Wang, and J. Hu, “Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity.” AMIA ... Annual Symposium proceedings. AMIA Symposium, vol. 2014, pp. 1258–67, 2014.

[8] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. H. Thomas, D. D. Edwards, B. K. Shoichet, and B. L. Roth, “Predicting new molecular targets for known drugs,” 2009.

[9] Y. Yamanishi, M. Kotera, M. Kanehisa, and S. Goto, “Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework,” Bioinformatics, 2010.

[10] Z. Yu, M. D. Gonciarz, W. I. Sundquist, C. P. Hill, and J. Jensen, “A New Method for Computational Drug Repositioning Using Drug Pairwise Similarity,” vol. 377, no. 2, pp. 364–377, 2012.

[11] 000K. Zhao and H.-C. So, “A machine learning approach to drug repositioning based on drug expression profiles: Applications in psychiatry,” arXiv preprint arXiv:1706.03014, 2017.

[12] R. Hodos, B. Kidd, K. Shameer, B. Readhead, and J. Dudley, “Computational Approaches to Drug Repurposing and Pharmacology,” Wiley interdisciplinary reviews. Systems biology and medicine, vol. 8, no. 3,

[13] G. Wu, J. Liu, and C. Wang, “Semi-supervised graph cut algorithm for drug repositioning by integrating drug, disease and genomic associations,” Proceedings - 2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016, pp. 223–228, 2017.

[14] D. H. Le and D. Nguyen-Ngoc, “Drug repositioning by integrating known disease-gene and drug-target associations in a semi-supervised learning model,” Acta Biotheoretica, vol. 66, no. 4, pp. 315–331, 2018. [Online].

[15] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross et al., “The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease,” science, vol. 313, no. 5795, pp. 1929–1935, 2006.

[16] A. C. Cheng, R. G. Coleman, K. T. Smyth, Q. Cao, P. Soulard, D. R. Caffrey, A. C. Salzberg, and E. S. Huang, “Structure-based maximal affinity model predicts small-molecule druggability,” Nature Biotechnology, vol. 25, no. 1, pp. 71–75, 2007.

[17] Z.-C. Li, M.-H. Huang, W.-Q. Zhong, Z.-Q. Liu, Y. Xie, Z. Dai, and X.-Y. Zou, “Identification of drug–target interaction from interactome network with guilt-by-association principle and topology features,” Bioinformatics, vol. 32, no. 7, pp. 1057–1064, 2015.

[18] K. Bleakley and Y. Yamanishi, “Supervised prediction of drug-target interactions using bipartite local models,” Bioinformatics, 2009.

[19] Y. Wang and J. Zeng, “Predicting drug-target interactions using restricted Boltzmann machines,” in Bioinformatics, 2013.

[20] C. Wang, J. Liu, F. Luo, Y. Tan, Z. Deng, and Q.-N. Hu, “Pairwise input neural network for target-ligand interaction prediction,” in Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on. IEEE, 2014, pp. 67–70.

[21] Z. C. Li, M. H. Huang, W. Q. Zhong, Z. Q. Liu, Y. Xie, Z. Dai, and X. Y. Zou, “Identification of drug-target interaction from interactome network with ’guilt-by-association’ principle and topology features,” Bioinformatics, 2016.

[22] B. Sushma, C. V. Suresh, S. Mary, and E. G. D. Ap, “DOCKING-A Review,” Applicable Chemistry, vol. 1, no. 2, pp. 167–173, 2012.

[23] F. Yang, J. Xu, and J. Zeng, “DRUG-TARGET INTERACTION PREDICTION BY INTEGRATING CHEMICAL, GENOMIC, FUNCTIONAL AND PHARMACOLOGICAL DATA HHS Public Access,” Pac Symp Biocomput, pp. 148–159, 2014. [Online].

[24] L. Xie, T. Evangelidis, L. Xie, and P. E. Bourne, “Drug discovery using chemical systems biology: Weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir,” PLoS Computational Biology, vol. 7, no. 4, 2011.

[25] L. Yang, K. Wang, J. Chen, A. G. Jegga, H. Luo, L. Shi, C. Wan, X. Guo, S. Qin, G. He, G. Feng, and L. He, “Exploring off-targets and off-systems for adverse drug reactions via chemical-protein interactome clozapine-induced agranulocytosis as a case study,” PLoS Computational Biology, vol. 7, no. 3, 2011.

[26] Z. Mousavian and A. Masoudi-Nejad, “Drug–target interaction prediction via chemogenomic space: learning-based methods,” Expert opinion on drug

metabolism & toxicology, vol. 10, no. 9, pp. 1273–1287, 2014.

[27] S. Kim, D. Jin, and H. Lee, "Predicting drug-target interactions using drug-drug interactions," PLoS ONE, vol. 8, no. 11, pp. 1–12, 2013.

[28] K. Tian, M. Shao, Y. Wang, J. Guan, and S. Zhou, "Boosting compound-protein interaction prediction by deep learning," 2016.

[29] Y. Bromberg, "Chapter 15: Disease Gene Prioritization," PLoS Computational Biology, vol. 9, no. 4, 2013.

[30] L. Huang, Y. Wang, Y. Wang, and T. Bai, "Gene-Disease Interaction Retrieval from Multiple Sources : A Network Based Method," vol. 2016, 2016.

[31] K. Wysocki and L. Ritter, "An Approach to Understanding GeneDisease Interactions."

[32] D. A. Davis and N. V. Chawla, "Exploring and Exploiting Disease Interactions from Multi-Relational Gene and Phenotype Networks," vol. 6, no. 7, 2011.

[33] A. Suratanee and K. Plaimas, "Network-based association analysis to infer new disease-gene relationships using large-scale protein interactions," PLoS ONE, vol. 13, no. 6, pp. 1–20, 2018.

[34] J. Zhao, T.-h. Yang, Y. Huang, and P. Holme, "Ranking Candidate Disease Genes from Gene Expression and Protein Interaction : A Katz-Centrality Based Approach," vol. 6, no. 9, 2011.

[35] L. Katz, "a New Status INDEX DERIVED From Sociometric," Psychmetrika, vol. 18, no. 1, pp. 39–43, 1953.

[37] S. S. Khan and M. G. Madden, "One-class classification: taxonomy of study and review of techniques," The Knowledge Engineering Review, vol. 29, no. 3, pp. 345–374, 2014.

[38] P. Juszczak, "Learning to recognise: A study on one-class classification and active learning," 2006.

[39] R. Kiryo, G. Niu, M. C. du Plessis, and M. Sugiyama, "Positive-unlabeled learning with non-negative risk estimator," in Advances in neural information processing systems, 2017, pp. 1675–1685.

[40] E. Sansone, F. G. De Natale, and Z.-H. Zhou, "Efficient training for positive unlabeled learning," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018.

[41] J. Bekker and J. Davis, "Learning from positive and unlabeled data: A survey," arXiv preprint arXiv:1811.04820, 2018.

[42] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," Signal Processing, vol. 99, pp. 215–249, 2014.

[43] S. Marsland, "Novelty detection in learning systems," Neural computing surveys, vol. 3, no. 2, pp. 157–195, 2003.

[44] A. Liaw, M. Wiener et al., "Classification and regression by randomforest," R news, vol. 2, no. 3, pp. 18–22, 2002.

[45] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," IEEE transactions on systems, man, and cybernetics, vol. 21, no. 3, pp. 660–674, 1991.

[46] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. MacIejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, and M. Wilson, "DrugBank 5.0: A major update to the DrugBank database for 2018," Nucleic Acids Research, vol. 46, no. D1, pp. D1074–D1082, 2018.

[47] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, and S. H. Bryant, "PubChem substance and compound databases," Nucleic Acids Research, vol. 44, no. D1, pp. D1202–D1213, 2016.

[48] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman, "The Pfam protein families database: Towards a more sustainable future," Nucleic Acids Research, vol. 44, no. D1, pp. D279–D285, 2016.

[49] R. Xu, L. Li, and Q. Wang, "Towards building a disease-phenotype knowledge base: Extracting disease-manifestation relationship from literature," Bioinformatics, vol. 29, no. 17, pp. 2186–2194, 2013.

[50] M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong, "BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology," Nucleic Acids Research, vol. 44, no. D1, pp. D1045–D1053, 2016.

[51] J. Piñero, Á. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, and L. I. Furlong, "Dis-GeNET: A comprehensive platform integrating information on human disease-associated genes and variants," Nucleic Acids Research, vol. 45, no. D1, pp. D833–D839, 2017.

[52] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, Logistic regression. Springer, 2002.

[53] M. Fishbein and I. Ajzen, Predicting and changing behavior: The reasoned action approach. Psychology Press, 2011.

[54] C. D. Cunha, B. Agard, and A. Kusiak, "quality r P Fo r R w On ly," 2010.

[55] M. Bee, "Simulating copula-based distributions and estimating tail probabilities by means of adaptive imporance sampling," 2010.

[56] P. Trivedi and D. Zimmer, "A Note on Identification of Bivariate Copulas for Discrete Count Data," Econometrics, vol. 5, no. 1, p. 10, 2017.

[57] P. K. Trivedi and D. M. Zimmer, "Copula Modeling: An Introduction for Practitioners," Foundations and Trends R in Econometrics, vol. 1, no. 1, pp. 1–111, 2006.

[58] "simstudy update: improved correlated binary outcomes," 2018.