

A Survey on Drug-Target Interaction Prediction Methods Analysis of Prediction Mechanisms for Drug Target Discovery

Shyama M Nair¹, Joby George²

^{1,2}Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India.

Abstract: *In Modern drug discovery, identifying drug target interactions is having high significance in terms of suggesting new drug candidates and repositioning old drugs. Prediction of drug-target interactions can be done by experimentally and it is very expensive and time consuming. Therefore, there is a continuous demand for effective and low-cost computational techniques for drug target interaction prediction. Many algorithms have been formulated to infer novel drug – target interactions. In this paper, we make a survey on the recent progress being made on computational methodologies that have been developed to predict drug targets based on different kinds of drug and protein data.*

Keywords: *Drug discovery, Drug repositioning, Drug – target interactions, Computational drug target prediction.*

I. INTRODUCTION

As we know, medical interventions improved human health considerably. But still many diseases remain poorly treated while new ones are emerging. For example cancer and neurodegenerative disorders still lack effective medicines. It is extremely valuable to discover novel drug targets as the source for drug development. However, experimental validation of drug-target interactions is highly expensive and time consuming, Therefore only a small number of drug target interactions have been validated by biological experiments. Thus the interest in computational prediction of potential drug-target interactions has recently grown.

Significant work has been done to predict drug-target interactions by using different types of data individually or together, including gene expression data, protein sequence, protein-protein interactions, mass spectrum, chemical structure of drugs, drug response, metabolic network, drug side effects and so on. Predicting and countering the side effects of a new drug during its developmental phase remain important to the drug's overall commercial success. Side effects are responsible for a significant number of cases where premarket drugs fail during clinical trials. Identifying the underlying mechanisms of side effects is a challenging task, often because of the drugs' pleiotropic effects on a biological system. Most drugs are small compounds that target and interact with proteins to induce perturbations in the proteins network. With the advent of new technologies and open data initiatives, the past decade has witnessed an exponential growth of chemical biology data available in the public databases. It is necessary to collect as much data as possible on drugs, targets, and their interactions, for predicting DTIs. Online biological databases that store and maintain information on already known drugs and drug-target interactions; examples of such databases include KEGG [1], DrugBank [2], ChEMBL [3] and STITCH [4]. In this paper, we present the latest advancement on computational methodologies that have been developed to identify drug targets. Besides, we introduce popular public resources about drug target information, which can significantly facilitate the discovery of new drugs. Note that this survey aims to summarize the most recent development on computational approaches for prediction of drug targets; however, it is by no means comprehensive due to the rapid evolution of the field.

II. BACKGROUND

This section describes the concept of Drug Repositioning, its benefits, and different data resources used in computational drug target prediction methods.

A. Drug Repositioning

Drug repositioning or Drug repurposing has long been a necessary procedure of drug development because it can renew a failed drug or increase the number of indications for a successful one. Potentially, repositioning can minimize the traditional time line of 10-17 years and make drugs available for use in patients in 3-12 years. Computational repositioning is the process of designing and validating automated workflows that can make hypotheses for new indications for a drug candidate. The potential for computational repositioning is high, given that a systematic process can incorporate prioritization information that can accelerate time lines even

further. Typically, repositioning has been accomplished by using the mechanistic knowledge of the target to infer a new disease indication or by observing new clinical phenotypes

Repositioning methods have many features in common with those of standard drug discovery; they identify new links between entities and complete the drug - disease connection using domain expertise. Traditional target-based drug discovery connects a target to a disease, and a screening process is used to find drug like molecules that bind the target. Phenotypic screening skips the target and makes a direct attempt to find a drug-like molecule that will exhibit the correct phenotypic behaviour. Most repositioning methods use similar approaches but connect different concepts. They are

- 1) Transcriptomic methods such as CMap associate the drug with a specific aspect of a disease and then extrapolate that knowledge to the disease itself.
- 2) Side-effect methods associate drugs with new disease indications by representing a disease on the basis of a set of side effects associated with its treatment and then identifying other drugs with similar side effects.
- 3) Genetics-based methods connect the target to a genetic phenotype, which may be the disease itself or a phenotype related to the disease

In each case, it is the systematic combination of experimental data and pre-existing knowledge that completes the connection.

B. Benefits of Drug Repositioning

Drug repositioning offers the following advantages:

- 1) Anticancer drug discovery. Due to high demand for anticancer drugs, identifying new anticancer therapies from existing drugs become highly popular.
- 2) Drug repositioning turns out to be a promising approach for the discovery of anti-infectious drugs that can overcome drug resistance. Because the drug resistance is awful to human beings and can increasingly reduce the drug efficacy.
- 3) It is an alternative therapeutics for orphan and rare diseases.
- 4) Drug repositioning provides a new accessibility to personalized treatment.
- 5) It reduces the development cost for the drugs, because they have already been gone through toxicity and other tests such as clinical trials. They have higher success rate than the original drugs, because of the availability of comprehensive information on their pharmacology, formulation, potential toxicity, safety and adverse drug reaction issues.
- 6) Drug repurposing allows a new drug candidate could be set for clinical trials quickly. Also if approved, their incorporation into health care and thus diminishing their entire processing time.
- 7) Re-profiled drugs can circumvent initial cost and time needed to introduce a drug to the market. As developing a brand-new drug takes massive time, money and effort.

C. Computational Drug Target Prediction – Data Resources

Computational analyses allow the researcher to generate, evaluate, and prioritize data for several drugs and diseases simultaneously. Furthermore, with systematic efforts being continually strengthened by newly available platform data types, the rapidly expanding database in the literature, and improvements in analytical methodology, these methods can be expected to increase in significance. Also the usage of these computational techniques extends beyond drug repositioning; they can also be used to find the initial indications for a drug. This section describes various data resources used for predicting drug target interactions.

- 1) Chemical Similarity: Chemical similarity between two drugs can be defined by various means, based on sub-structural features or physicochemical descriptors. Chemical similarities based on 3D structure can successfully identify off-targets and also 3D-based method facilitates enrichment even for compounds which are not found to be similar in 2D methods.
- 2) Bioactivity Profile: Bioactivity or pharmacological activity describes the beneficial or adverse effects of a drug on living matter in pharmacology. When a drug is a complex chemical mixture, this activity is exerted by the substance's active ingredient or pharmacophore but can be changed by the other constituents. Among the various properties of chemical compounds, pharmacological/biological activity plays a crucial role since it gives the idea about the uses of the compounds in the medical applications. However, chemical compounds may show some adverse and toxic effects which may prevent their use in medical practice. The availability of chemical biology data across multiple assays for a common compound library enables the generation of bioactivity profiles, which can be informative for predicting DTIs.
- 3) Drug Side Effect: Drug side effect is either therapeutic effect or adverse effect. Although the term is predominantly employed to describe adverse effects, it can also be used to describe beneficial, but unintended, consequences of the use of a drug. Also Side effects,

or the adverse effects of drugs, contain critical clinical phenotypic information that may be useful for predicting novel targets of a drug .

- 4) **Therapeutic Effect:** Therapeutic effect deals with the responses(s) after a treatment of any kind, the results of which are judged to be desirable and beneficial. The Anatomical Therapeutic Chemical (ATC) classification system classifies drugs by their therapeutic and chemical characteristics.
- 5) **Drug-Induced Gene Expression:** Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product like protein or RNA. The genetic code stored in DNA is interpreted by gene expression, and the properties of the expression give rise to the organism's phenotype. Regulation of gene expression is thus critical to an organism's development. Gene expression profiles occur from drug treatment can provide insights to DTI prediction.
- 6) **Drug Binding Site :** Binding site is a region on a protein or piece of DNA or RNA to which ligands may form a chemical bond. Binding sites also exhibit chemical specificity, a measure of the types of ligands that will bond, and affinity, which is a measure of the strength of the chemical bond. Ligand-target interactions are mainly identified by the physicochemical properties of the binding sites, which also largely depend on the ligand substructures.
- 7) **Drug-Drug Interaction:** Drug-drug interaction is the modification of the effect of a drug when administered with another drug. It is the change in a drug's effect on the body when the drug is taken together with a second drug. A drug-drug interaction can delay, decrease, or enhance absorption of either drug. This can decrease or increase the action of either or both drugs or cause adverse effects. Drug-drug interaction (DDI) is a promising feature for predicting DTIs
- 8) **Ontology and Semantic Data :** Drug similarity can be measured by ontological terms in a hierarchical classification.
- 9) **Literature and Text Mining:** Hidden drug-target interaction (DTI) in literature can be discovered via text mining based on co-occurrence of drug and target entities.

D. Computational Drug Target Prediction – Methods

Fig 1 shows the latest methods used in drug target prediction. It include integration of heterogeneous data resources , deep learning, matrix factorization methods, machine learning methods and ontology reference models. Overview of these methods is described in section III.

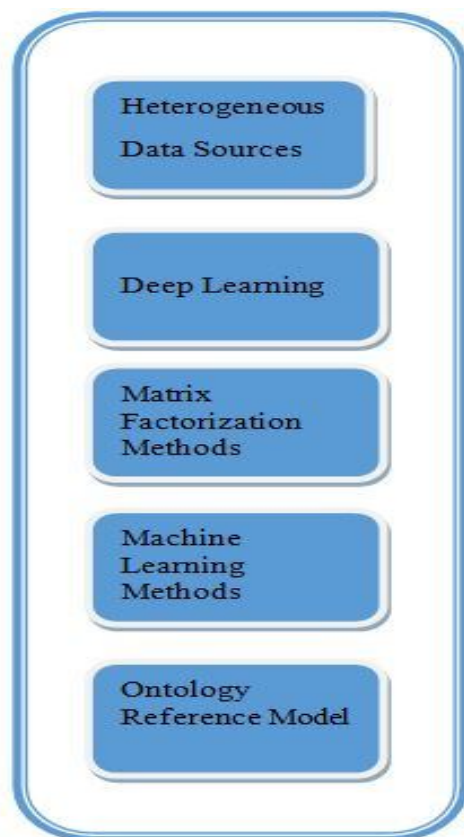


Fig 1 Latest Trends in Computational Drug Target Prediction Approaches

III.OVERVIEW OF DRUG – TARGET INTERACTION PREDICTION MECHANISMS

In this review, we survey latest approaches and applications that were published in the last few years as well as a few earlier pioneering studies. Most studies for DTI prediction were based on the hypothesis that similar targets interact with same drug, and the same target interacts with similar drugs. The similarities among drugs reflect a chemical space while similarities among targets reflect a genomic space. A structure-based maximal affinity model is introduced by Alan C Cheng *et al.*[5] to predict small-molecule druggability. Here a model-based approach using basic biophysical principles yields good prediction of druggability based solely on the crystal structure of the target binding site. The maximal affinity achievable by a drug-like molecule is quantitatively estimated, and showed that these calculated values correlate with drug discovery out-comes. A priori druggability assessment can be used in combination with target validation and feasibility assessments in selecting targets. Results also highlight strategies for difficult druggability targets, including consideration of covalent adduct, metal chelation, prodrug, active transport and allosteric approaches. However, there are some disadvantages to this technique: for docking simulations, Simulation always needs three-dimensional (3D) structures of targets which are often unavailable. Simulation is heavily time consuming. Drug-target interaction prediction by learning from local information and neighbours method is proposed by Jian-Ping Mei *et al.*[6], which discuss about a Bipartite Local Model- Neighbour based Interaction profile Inferring(BLM-NII). BLM-NII is the method presented for predicting drug-target interactions. In this method, the NII procedure is applied for those completely new candidates that have no existing training data at all, and the results are found already good enough to show the usefulness of NII. Since it is quite common that drugs only activate or inhibit a small number of targets and targets are only activated or inhibited by very limited drugs, the NII procedure may be applied to drugs and targets which do not have sufficient training data. More accurate prediction models may be build by using neighbours information to enhance the limited training examples. However, too much emphasis on neighbours tends to eliminate the local characteristics of each drug and target and could cause deterioration in the prediction performance. Predicting drug-target interactions for new drug compounds using a weighted nearest neighbour profile by Twan van Laarhoven *et al.*[7] discuss about Regularised Least Square Weighted Nearest Neighbour method. Here RLS uses RLSkron as its base algorithm and augments it with WNN. A limitation of this approach is that it does not make a difference between an inactive target and a target that has not been measured for a compound. Compounds with a higher mutual chemical similarity also have a higher chance of having the same bioactivity. This information could be considered by WNN by determining directly the weights from the similarity, instead of using the proposed ranking-based decay mechanism. In this way all the compounds with high similarity would be considered with a high weight and all the compounds with low similarity would only have a minor contribution to the final predicted profile. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions method is proposed by Xiaodong Zheng *et al.*[8],which discuss about Matrix factorization techniques that have been used to predict interactions recently. Such techniques decompose the matrix representing the drug-target network into multiple low-rank matrices consisting of latent (or hidden) features that are assumed to govern the drug-target interactions. The key idea of this approach is to use more than one similarity matrices over drugs as well as those over targets, where weights over the multiple similarity matrices are estimated from data to automatically select similarities, which are effective for improving the performance of predicting drug-target interactions. Collaborative Matrix Factorization (CMF), projects drugs and targets into a common low-rank feature space, which is further consistent with weighted similarity matrices over drugs and those over targets. These two low-rank matrices and weights over similarity matrices are estimated by an alternating least squares algorithm. This approach allows to predict drug-target interactions by the two low-rank matrices collaboratively and to detect similarities which are important for predicting drug-target interactions. The main idea of CMF is to project drugs and targets into two low-rank matrices, corresponding to feature spaces of drugs and targets, respectively. In order to distinguish known drug-target pairs from unknown pairs, weighted low-rank approximation is formulated. Mehmet Gönen [9] propose a novel Bayesian formulation that combines kernel-based nonlinear dimensionality reduction , matrix factorization and binary classification for predicting drug–target interaction networks using only chemical similarity between drug compounds and genomic similarity between target proteins. The novelty of this approach comes from the joint Bayesian formulation of projecting drug compounds and target proteins into a unified subspace using the similarities and estimating the interaction network in that subspace. This method is using a variational approximation in order to obtain an efficient inference scheme and give its detailed derivations. Different from previous studies, this method is the first fully probabilistic formulation for drug–target interaction network inference. Ali Ezzat *et. al.*[10] proposed a graph regularized matrix factorization method. The method uses a pre-processing step that adds edges with intermediate interaction likelihood scores to assist with prediction. The authors were interested in predicting ligands for orphan targets, and their experiments have shown that ligands of orphan targets could be predicted with reasonable accuracy whenever close neighbours (with known ligands) of those orphan targets were available, regardless of the kernels used for ligands or targets. A matrix factorization method is used to predict drug-target

interactions. However, unlike CMF and KBMF2K, graph regularization is used to prevent over fitting. In graph regularization, the similarity matrices are sparsified beforehand by keeping only the similarity values to the nearest neighbours for each drug/target. Campillos et al. [11] make use of drug side effects to measure the similarity among drugs, which pave a new way to represent drugs. 20 of unexpected DTIs are discovered by their approach, and 13 of them are validated by in vitro binding assays. Mizutani et al.[12] integrated drug side effects and protein function to predict drug targets. Kutalik et al.[13] proposed a modular Ping Pong approach by integrating gene expression and drug response data in NCI-60 cell lines. Drug-gene associations are discovered by identifying co-modules of drugs and genes, which exhibit similar patterns in some cell lines. Drugs and genes in a co-module are assumed to be associated with each other with a high probability. Chen et al. [14] proposed sparse network-regularized partial least square (SNPLS) method to identify joint modular patterns using large-scale pairwise gene-expression and drug response data, which incorporated a molecular network to the (sparse) partial least square model to improve the module accuracy via a network-based penalty. Ding et al.[15] and Zheng et al.[16] proposed similarity-based methods for predicting drug target interactions. Li et al. [17] have utilized the human metabolic network as a basis for the prediction of novel targets for known anticancer drugs. Emig et al. [18] proposed a network based approach by integrating disease gene expression signatures and a molecular interaction network. Yuan et al. [19] proposed a novel method, Drug E-Rank, to improve the prediction accuracy by effectively integrating the advantages between feature-based and similarity-based methods which are two types of approaches for drug target prediction. Drug E-Rank uses 'Learning To Rank', for which multiple well-known similarity-based methods can be used as components of ensemble learning. Peng-Wei et al. [20] developed a deep learning based approach called multi-scale features deep representations inferring interactions (MFDR). Deep learning has been found to be an appropriate tool for converting high-dimensional features to low-dimensional representations. These deep representations generated from drug-protein pair can serve as training examples for the interaction predictor. Large-scale chemical structure and protein sequence descriptors are extracted so as to machine learning model predict if certain human target protein can interact with a specific drug. MFDR use Auto-Encoders as building blocks of deep network for reconstruct drug and protein features to low-dimensional new representations. Then, support vector machine are used to infer the potential drug-target interaction from deep representations. MFDR is able to predict large-scale drug-target interactions with high accuracy and achieves results better than other feature-based approaches. Limin Li et al.[21] proposed multi view low rank embedding to predict drug target interactions. Here heterogeneous data sources are used to predict drug-target interactions. It has been shown that integrating multi-view representations of drugs and proteins can strengthen the prediction ability. A drug can be represented by its chemical structure, or by its chemical response in different cells. A protein can be represented by its sequence, or by its gene expression values in different cells. The docking of drugs and proteins based on their structure can be considered as one view (structural view), and the chemical performance of them based on gene expression and drug response can be considered as another view (chemical view). Initially, a single-view approach of SLRE based on low rank embedding for an arbitrary view is proposed, and then extend it to a multi-view approach of MLRE, which could integrate both views. Moreover, semantic technologies have facilitated the integration of various data sources and the discovery of new drug indications. For instance, Zhu *et al.* [22] developed an ontology to model FDA-approved breast cancer drugs and their relations with pathways, drugs, genes, SNPs and diseases. New drug-disease pairs were inferred from the ontology-based knowledge base. Chen *et al.* [23] developed a statistical model to assess drug-target associations from a semantic linked network comprised by drugs, chemical compounds, protein targets, diseases, side effects and pathways and their relations. The model considered the topology and semantics of the sub graph between a drug and a target. The similar drug-drug pair from different disease areas may indicate a potential repositioning opportunity.

IV. ANALYSIS OF DRUG – TARGET INTERACTION PREDICTION MECHANISMS

Despite several successful use cases of computational drug repositioning, challenges are still remain. First, the transformation of theoretical computational models into practical use is far from straightforward, due to some inevitable factors like missing data, data bias and technical limitations of computational methods. For example, many in silico repositioning approaches search potential drug-target interactions through chemical structure information. Thus, the lack of high-resolution structural data for targets makes such methods inadequate. Second, the lack of structured gold standard for drug repositioning made it hard to compare and evaluate the performance of computational methods. Third, although computational drug repositioning may merely shorten the process of drug discovery in preclinical and Phase I trials, challenges may still exist after Phase II trials for the repositioned drugs. Integration of heterogeneous data sources, deep learning, matrix factorization and ontology based models are the latest methods in the field of drug target interaction prediction. Each of the aforementioned computational repositioning strategies and approaches has their methodological advantages and limitations. A combination of these methods is often desired for achieving better results. Thus integrative methods showed better performance in both sensitivity and specificity when comparing with individual methods.

V. CONCLUSION

Computational drug repositioning research is of great relevance to improve human health through finding new uses for existing drugs. In fact, a number of studies have already been carried out with various degrees of success. It has great potentials to accelerate drug discovery with interesting opportunities in several particular disease areas like cancer, infectious and orphan diseases. Their applications span almost all stages in the discovery and development pipeline, from target identification to lead discovery, from lead optimization to preclinical or clinical trials. Computational drug target prediction methods increase our ability to efficiently and effectively predict drug targets in high-throughput formats. With huge volume of bio-related data and latest technologies, computational drug target prediction methods will facilitate drug discovery and drug design to address the unmet medical needs.

REFERENCES

- [1] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "Kegg for integration and interpretation of large-scale molecular data sets," vol. 40, no. D1, pp. D109–D114, 201.
- [2] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, and D. S. Wishart, "Drugbank 3.0: a comprehensive resource for 'omics' research on drugs," vol. 39, no. suppl 1, pp. D1035–D1041, 2011.
- [3] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, and J. P. Overington, "ChEMBL: a large-scale bioactivity database for drug discovery," 2011.
- [4] M. Kuhn, D. Szklarczyk, S. Pletscher-Frankild, T. H. Blicher, C. von Mering, L. J. Jensen, and P. Bork, "Stitch 4: integration of protein chemical interactions with user data," vol. 42, no. D1, pp. D401–D407, 2014.
- [5] A. C. Cheng, R. G. Coleman, K. T. Smyth, Q. Cao, P. Souillard, 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.
- [6] J.-P. Mei, C.-K. Kwok, P. Yang, X.-L. Li, and J. Zheng, "Drug-target interaction prediction by learning from local information and neighbors," *Bioinformatics*, vol. 29, no. 2, pp. 238–245, 2013.
- [7] T. van Laarhoven and E. Marchiori, "Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile," *PloS one*, vol. 8, no. 6, p. e66952, 2013.
- [8] X. Zheng, H. Ding, H. Mamitsuka, and S. Zhu, "Collaborative matrix factorization with multiple similarities for predicting drug-target interactions," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1025–1033.
- [9] M. Gonen, "Predicting drug-target interactions from chemical and genomic kernels using bayesian matrix factorization," *Bioinformatics*, vol. 28, no. 18, pp. 2304–2310, 201.
- [10] Ali Ezzat, Peilin Zhao, Min Wu, Xiao-Li Li, and Chee-Keong Kwok, "Drug-Target Interaction Prediction with Graph Regularised Matrix Factorization". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2016.
- [11] Campillos.M, Kuhn.M, Gavin.A.C, Jensen.L.J, and Bork.P. "Drug target identification using side-effect similarity". *Science*, 321(5886):263–6, 2008.
- [12] Mizutani.S, Pauwels.E, Stoven.V, Goto.S, and Yamanishi.Y. "Relating drug protein interaction network with drug side effects." *Bioinformatics*, 28(18):i522–i528, 2011.
- [13] Kutalik.Z, Beckmann.J.S, and Bergmann.S. "A modular approach for integrative analysis of large-scale gene-expression and drugresponse data". *Nature Biotechnology*, 26(5):531–539, 2008.
- [14] Chen.J and Zhang.S. "Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data". *Bioinformatics*, 32(11), 2016.
- [15] Ding.H, Takigawa.I, Mamitsuka.H, and Zhu.S. "Similarity-based machine learning methods for predicting drug-target interactions: a brief review". *Briefings in Bioinformatics*, 15(5):734–747, 2014.
- [16] Zheng.X, Ding.H, Mamitsuka.H, and Zhu.S.F. "Collaborative matrix factorization with multiple similarities for predicting drugtarget interactions". 2013.
- [17] Li.L, Zhou.X, Ching.W.K, and Wang.P. "Predicting enzyme targets for cancer drugs by profiling human metabolic reactions in nci-60 cell lines". *BMC Bioinformatics*, 11(1):501, 2010.
- [18] Emig.D, Ivliev.A, Pustovalova.O, Lancashire.L, Bureeva.S, Nikolsky. Y, and Bessarabova.M. "Drug target prediction and repositioning using an integrated network-based approach". *Plos One*, 8(4):e60618–e60618, 2013.
- [19] Yuan.Q, Gao.J, Wu.D, Zhang.S, Mamitsuka.H, and Zhu.S. "Drugerank: improving drugtarget interaction prediction of new candidate drugs or targets by ensemble learning to rank". *Bioinformatics*, 32(12):i18–i27, 2016.
- [20] Peng-Wei, Keith C.C. Chan, Zhu-Hong You, "Large-Scale Prediction of Drug-Target Interactions from Deep Representations" *International Joint Conference on Neural Networks (IJCNN)* 2016.
- [21] Limin Li; Menglan Cai "Drug target prediction by multi-view low rank embedding," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 201
- [22] Zhu Q, Tao C, Shen F, et al. "Exploring the pharmacogenomics knowledge base (PharmGKB) for repositioning breast cancer drugs by leveraging Web ontology language (OWL) and cheminformatics approaches". *Pac Symp Biocomput* 2014:172–82.
- [23] Chen B, Ding Y, Wild DJ. "Assessing drug target association using semantic linked data". *PLoS Comput Biol* 2012;8(7):e1002574.