Weighted Semi-Supervised Approaches for Predictive Modeling and Truth Discovery

By

Sai Nivedita Chandrasekaran

Submitted to the Bioengineering Program and the
Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Committee members

Dr. Jun Huan, Chairperson

Dr. Bo Luo

Dr. Jerzy Gryzmala Busse

Dr. Sara Wilson

Dr. Zhou Wang

Date defended: May 3, 2017

The Dissertation Committee for Sai Nivedita Chandrasekaran certifies
that this is the approved version of the following dissertation :

Weighted Semi-Supervised Approaches for Predictive Modeling and Truth Discovery

Dr. Jun Huan, Chairperson

Date approved: May 11, 2017

Abstract

Multi-View Learning (MVL) is a framework which combines data from heterogeneous sources in an efficient manner in which the different views learn from each other, thereby improving the overall prediction of the task. By not combining the data from different views together, we preserve the underlying statistical property of each view thereby learning from data in their original feature space. Additionally, MVL also mitigates the problem of high dimensionality when data from multiple sources are integrated. We have exploited this property of MVL to predict chemical-target and drug-disease associations. Every chemical or drug can be represented in diverse feature spaces that could be viewed as multiple views. Similarly multi-task learning (MTL) frameworks enables the joint learning of related tasks that improves the overall performances of the tasks than learning them individually. This factor allows us to learn related targets and related diseases together. An empirical study has been carried out to study the combined effects of multi-view multi-task learning (MVMTL) to predict chemical-target interactions and drug-disease associations.

The first half of the thesis focuses on two methods that closely resemble MVMTL. We first explain the weighted Multi-View learning (wMVL) framework that systematically learns from heterogeneous data sources by weighting the views in terms of their predictive power. We extend the work to include multi-task learning and formulate the second method called Multi-Task with weighted Multi-View Learning (MTwMVL). The performance of these two methods have been evaluated by cheminformatics data sets.

We change gears for the second part of this thesis towards truth discovery (TD). Truth discovery closely resembles a multi-view setting but the two strongly differ in certain aspects. While the underlying assumption in multi-view learning is that the different views have label consistency, truth finding differs in its setup where the main objective is to find the true value of an object given that different sources might conflict with each other and claim different values for that object. The sources could be considered as views and the primary strategy in truth finding is to estimate the reliability of each source and its contribution to the truth. There are many methods that address various challenges and aspects of truth discovery and we have in this thesis looked at TD in a semi-supervised setting.

As the third contribution to this dissertation, we adopt a semi-supervised truth discovery framework in which we consider the labeled objects and unlabeled objects as two closely related tasks with one task having strong labels while the other task having weak labels. We show that a small set of ground truth helps in achieving better accuracy than the unsupervised methods.

Acknowledgements

I would like to first and foremost thank Dr. Huan for accepting me in his group and for constantly motivating me throughout my study. His positive words have been a very big source of inspiration. I would like to thank Dr. Sara Wilson for serving in my committee and I am very grateful to her for guiding me along the right path when I needed it the most.

I express my gratitude Dr. Jerzy Grzymala-Busse, Dr. Bo Luo and Dr. Zhou Wang for serving in my PhD committee and giving their valuable suggestions.

I would like to thank my lab mates and my seniors who passed out of the lab for helping me in various capacities throughout my study.

My sincere gratitude from the bottom of my heart to my parents Mr.K.Chandrasekaran and Dr.P.V.Geetha without whom this dissertation would not have been possible. Their physical, emotional and financial support is the reason I was able to complete this dissertation.

I would like to thank my husband Mr. Siddharth Gangadhar for being my pillar of support and for always encouraging me to go after my dreams. This dissertation is a dedication to my three year old daughter Sahana Goda who has made her share of sacrifices to help me complete my study here at KU. A special thanks to my parents in-law, Mrs. Radha Gangadhar and Mr. Gangadhar for their immense support.

Lastly I thank all my close friends and family across the globe who have gone out of their way to help me during my toughest of times.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We are at a time in history where data is all around us. Every product that is manufactured are becoming more and more data driven. A lot of this data is open source and available to the public that makes the inference of data a very hot topic. On one hand, it is a great boon to have abundant data but on the other hand, it is difficult to obtain labels for all the data that can be used to build prediction models. Instead of discarding such unlabeled information, there are methods that utilize them in a way that contributes to the learning process. Similarly, a data point could be expressed in terms of heterogeneous features that are extracted from different sources. Each source might have dissimilar statistical property and combining them together would not be a good choice always.

Multi-view learning is a well established concept in machine learning that handles data from diverse sources as well as utilizes unlabeled samples in learning the prediction function. MVL is also indirectly combats the issue of high dimensionality when features from varied contributors are integrated. Although the concept of MVL has been around for quite a few years now, there is still some scope for improvement especially when some applications come with their own bag of challenges.

1.1 Motivation

The motivation of the thesis is based on semi-supervised learning that finds its application in many fields. In our work, we find motivations for it in two directions (i) methodology development (ii) application. In terms of methodology, we have tried to pick the shortcomings of existing methods that could be addressed to build better prediction models. In terms of application, there are many fields that benefit from semi-supervised learning. There are various fields in which obtaining labels can be very expensive and leveraging these unlabeled data could prove to be beneficial.

1.1.1 Cheminformatics

Cheminformatics is a fast growing branch in the field of chemistry where computational tools and techniques have been used to draw insights to address traditional problems such as drug discovery. These in silico methods are used to perform a virtual screening procedure that help in identifying chemicals/compounds of interest. The practice of building prediction models to identify drug candidates from thousands of chemicals has been adopted for many years. In order to fine tune or improve the performance of these models, more information regarding the small molecules could be integrated. Initial methods made use of the fact that compounds with similar structures had similar activities too. Likewise, this notion was extended to other concepts such as, compounds with similar side effects could have interactions with similar targets. While this could be extended to many such similarities, the need to handle such varied feature spaces was needed.

The second aspect of building such activity models is that the cost of getting labeled samples is both time as well as resource consuming. It is hence desirable to use unlabeled samples too to if it means better prediction. Our motivation in this application stems from the fact that compounds can be represented by very high dimensional feature spaces accrued from multiple sources. MVL helps in mitigating the ill effects of both the aforementioned problems.

The third aspect in cheminformatics that drew our attention is the availability of a huge number of related tasks. Multi-task learning is another concept in machine learning which states that, learning multiple related tasks together is beneficial than learning each task separately. The assumption here is that the tasks are related. Learning dissimilar tasks can lead to performance degeneration. The second advantage of MTL is that, a single task with low sample size when put together with a related task leverages its data. Hence MTL is also used to address low sample size. This closely related to the fact that two targets from the same family of proteins have high chances of being similar. For example dopamines D2 and D3 belong to a subclass of a family of proteins called GPCR and are bound to have very similar functional properties and so learning them together would be the right choice to make.

We also study drug-disease associations where drugs could be expressed in terms of their structure, activity with genes, pathways, side effects, etc,. which can be viewed as a multi-view problem. These genes in turn could be associated with a disease thereby letting us form an inferred relationship between a drug and a disease. Similar to the previous example, related diseases can be modelled together as a multi-task problem. In all of these applciations, each view do not have the same potential to predict a given task. These examples in cheminformatics pushed us to postulate two methods called the weighted Multi-view Learning and Multi-task with weighted Multi-view Learning.

1.1.2 Truth Discovery

Due to the information burst in current times and the availability of information regarding an object across platforms, it is a challenge to discover the truth of an object from a list of so called facts that are available across numerous sources. An example of such a scenario would be the location of a person so that customized ads could be recommended. The location could be collected from Facebook, Twitter, LinkedIn, Instagram, personal webpage and so on. Sometimes an unused platform by the user could carry outdated information while some other sources could copy this false

information. The challenge is to account for all these limitations while trying to find the truth about that object.

This aspect of multiple sources was our primary motivation to cast this as a multi-view problem. Current methods are scarce when it comes to addressing truth discovery as a learning problem. Since the truths are not just binary, we also explore the combination of multi-view in a multi-class setup. Currently, there is a huge gap in this area and so we have tried to introduce the concept of semi-supervised learning that could potentially open up many avenues in this direction. A key difference between MVL in a traditional setup and truth dsicovery is that, a fact represented by two or more views might be in conflict with each other.

There are many more shortcomings in the current methods and many applications that need to be enriched with better techniques and this thesis contributes in a small way among the many possibilities.

1.2 Contributions

The contributions of this thesis can again be divided into two categories namely (i) application and (ii) methodology. In terms of application, we have tried to introduce advanced machine learning algorithm such as multi-view and multi-task learning frameworks to predict interactions such as chemical-target and associations like drug-disease relationships. On one hand, the computer science community is churning out sophisticated prediction frameworks and on the other hand, researchers of the bioinformatics and cheminformatics communities still resolve to ML algorithms such as support vector machines, k-nearest neighbors, decision trees and neural networks for prediction purposes. The bridge this gap, we performed empirical studies of using multi-view and multi-task algorithms in cheminformatics.

In terms of methodology we have come up with two frameworks that are based on multi-view learning. The traditional multi-view learning assumes that each view has equal predictive power. Our hypothesis is that, some views might be better than others in terms of their predictive capability. The weighted Multi-View Learning (wMVL) addresses this by automatically learning the weights of the views without any prioir information. MVL learns a function using both labeled and unlabeled samples and the weights are estimated using the labeled samples. We then extend the wMVL framework to combine multi-task learning with the weighted multi-view framework. It has already been shown in the literature that multi-view multi-task learning that combines the advantages of learning from multiple sources and multiple related task is a better framework than each of them separately. This idea led us to extend the wMVL framework to include multi-task learning thereby proposing the Multi-Task with weighted Multi-View Learning framework (MTwMVL). Both these methods were evaluated by using drug-disease and chemical-protein datasets respectively.

The third method that we proposed focuses on the concept of truth discovery that has similarities with multi-view learning. In a multi-view setting, each sample is represented in terms of different set of features in each view whereas in truth discovery, the same set of objects and their facts are collected from multiple sources that might conflict with each other. The end goal of truth discovery is to find the truth of each object from the given set of conflicting facts. One of the main assumptions of MVL is label consistency where the label of a sample is the same across all the views. Truth discovery differs in the fact that facts of an object from two sources might not agree with each other. In order to present a new perspective to truth discovery, we project it as a semi-supervised learning method and borrow some concepts from MTL. The method has been evaluated using real world data sets.

1.3 Organization of the Proposal

The thesis first covers a literature survey of cheminformatics including topics such as drug-target interaction, drug-disease associations and the concept of drug repurposing. The chapter also includes discussions on the machine language algorithms widely used for predicting such interactions and associations following which we concentrate specifically on multi-view and multi-task learning frameworks. The third chapter explores the empirical of multi-view multi-task learning for predicting drug-disease associations and the fourth chapter presents the systematic study of multi-task learning for chemical-protein interaction. In the fifth chapter, we have come up with a weighted multi-view learning framework that estimates the weights of the views thereby presenting the idea that some views might have a better prediction power than the others. The sixth chapter then extends the weighted multi-view framework to integrate multi-task learning. Both these methods are evaluated using datasets from cheminformatics. The focus of this thesis then changes to truth discovery which is similar to multi-view learning where the goal is to find the truth of an object from multiple sources give that the information amongst them has a conflict. We finally present a semi-supervised learning framework for truth discovery. We conclude the thesis by proposing future directions for the semi-supervised truth discovery framework as well as truth discovery methods for cheminformatics that I would work on in the future.

.

Chapter 2

Literature Survey of Machine Learning in Cheminformatics

2.1 Introduction

Cheminformatics has been looked at as a tool used to fasten the process of drug discovery at various stages along the entire process. The strength and the huge development in cheminformatics lies in the ability to have access to a huge wealth of data in the public domain that has over the years increased exponentially. With huge sets of data deposited in databases with open access, the potential of exploiting them increased drastically. The ultimate goal of drug discovery is to zero-in on a chemical/compound that had the capability to interact with a target such as a protein to modulate it to produce the desired effect. Discovering such new associations has been the central theme in drug development. This process involves screening several thousands of small molecules or chemicals as a first stage to short-list candidates that make it to the clinical trials. Since physical testing of every chemical is a long and a time consuming process, researchers resolved to use computational tools to help them speed up this stage of screening the molecules. Virtual screening (VS) as it was called involved searching the library of chemicals to identify the ones to most likely to bind to a target. VS could be broadly classified into two categories namely: ligand based and

structure based methods.

Initial literature points to the development and establishment of Quantitative Structure Active Relationship (QSAR) models which predicted the activity of small molecules with targets based on their structure. The idea behind these models was that molecules with similar structure had similar activity profiles. One of the popular ways of establishing QSAR was by using machine learning techniques like Support Vector Machine (SVM), Neural Networks and Decision Trees. A key factor that has helped the growth of building prediction models can be attributed to the public databases (that are open source) that has a wealth of information in the form of bioassays, curation of the literature, toxicology studies and manually curated relationships. Some of the popular databases are PubChem [72], KEGG [69], Comparative Toxicogenomic Database [30], ChEMBL [51], DrugBank [79] and SIDER [76]. These databases contain chemical-target, chemical-pathway, chemical-gene, annotated drug side effects and other inferred relationships.

Machine learning algorithms were used for building QSAR prediction models both in a supervized as well as unsupervised learning approaches [122][15]. Preliminary empirical studies utilized SVM to build classification models to predict chemical-protein interactions [127][16][17][7]. A key feature in such predictions is the features/descriptors used in characterizing the small molecules. Over the years, researchers have exploited different aspects of drugs to discover these associations. While structural properties were the beginning point, many other characteristics of the compounds such as their genomic similarity, side effect similarity and pharmacological profiles were also utilized [127][17]. These spaces in addition to being used individually were also combined together to achieve better prediction results. Due to wide range of descriptors available to model drugs, empirical studies have been carried out to study their importance and significance. Alexious et al., [74] studied in detail the molecular descriptor space. Although their aim was to benchmark the descriptors, it showed that the different characteristics of a drug represented a whole different perspective and that the descriptors were not redundant. Example of such features are the 2D struc-

tural properties, extended circular fingerprints and pharmocophore descriptors. Beyond the initial application of existing algorithms, they were further fine tuned to highlight the importance of parameter selection and variable selection for QSAR models using SVM [98][131]. Self organizing maps were successfully used to differentiate substrates from inhibitors of P-glucoprotein. Given that both the substrate and inhibitor have the same interaction with the inhibition site, the successful differentiation using SOM was a right step in the direction of using ML for virtual screening [122].

A recent article by Lavecchia discusses the impact of ML approaches in drug discovery [78]. It details the scope and limitations of SVM, Decision Trees, kNN, Neural networks, Naive Bayes and SOM in ligand based virtual screening. Algorithms were not used just to build classification models but were also used for regression [110]. Partial least squares and boosted support vector regression methods have been used to establish QSAR [28][147]. Cross validation and bootstrapping methods were explored to build robust models that achieved better generalization performance. The major setback or hurdle in cheminformatics that might deteriorate the performance of prediction is the high dimensionality. As more spaces are explored, the dimensions can increase drastically. In order to combat this drawback, feature selection methods were adopted. An example of one such work is the stepwise exploration of the features using genetic algorithms and the concept of entropy [43].

Advanced ML frameworks such as active learning and ranking were used to find potential candidates [123][2]. Warmuth et al., showed that the active learning paradigm in machine learning clearly outperformed simpler techniques to choose the active compounds iteratively in each stage of the drug discovery process. Their work combined active learning with SVMs in selecting the active compounds in fewer iterations compared to other linear selection models. A similar work used active learning with SVM to classify cancer genes [89]. The comparison of active and passive learning showed significant difference in the performance of the classifiers which is advantageous

since active learning requires very small set of labeled samples. Experiments on gene classification of lung, colon and prostate cancer show that the labeled samples needed was contrastingly lesser for active learning - 31 samples vs 174 samples. Another aspect of drug discovery is to rank the chemicals to prioritize them instead of using them to build classification or regression models. Agarwal et al., [2] adopted ranking methods used in web retreival applications to perform empirical studies in virtual screening and have shown that the ranking algorithms identified potential candidates better than classification and regression models. A key feature of CPI interactions is that the ratio of the number of positive interaction samples to the number of negative samples is very low.

The challenge of predicting CPI associations using such imbalanced data has been combated by adopting data level as well as algorithmic level strategies [42]. Eitrich et al., addressed the problem of imbalanced data at the data level by sampling the data (either over sampling of the minor class or down sampling of the major class) and moving the threshold value in tandem with feature selection. A similar approach was used to screen drugs in PubChem by Li et al., [83] where they adopted a down sampling method since the ratio of the actives to inactives was as low as 1:377. But most of these works here experimented with SVM as the base classifiers which still leaves a huge gap between the advanced machine learning tools and the lack of their applciation in cheminformatics. Varnek et al., [114] have done an exhaustive work on the methods and trends of machine learning techniques in cheminformatics. They attribute the challenges in this field to incompleteness of molecular descriptors, accounting for multiple species and in-silico design of new molecules. They highlight the drawbacks of ML techniques not performing well on an external dataset since the training and test data might belong to different data domains. The paper explains in detail the statistical inference and modeling level view points of machine learning algorithms that are promising, their achievements in cheminformatics and other approaches that could be useful in improving the prediction accuracy. Yamanishi et al., [129] used multiple regression kernels for predicting drug-side effects relationship using a combined feature space of chemical and biological

spaces. With data science moving in the direction of large-scale analysis, the cheminformatics community is not far behind. Virtual screening has also delved into scalable algorithms that can handle huge amounts of data [124][108][6][64]. Prometheus - a software environment to screen millions of compounds to identify novel drug leads was used to prioritize compounds for biological screening. The authors [124] used Prometheus to dock about a million compounds into the estrogen receptor.

2.2 Drug Repurposing

Due to the long process and resources involved in discovering chemical-target interactions, another novel idea called drug repurposing made inroads in drug discovery [99][100][31]. The drugs go from screening to clinical trials to approval. The idea was to use already approved drugs and find newer targets for them. Initial drug discovery was based on the fact that compounds that had similar structures had similar interactions. This resulted in finding multiple compounds for a target. Drug repruposing on the other hand tried to find multiple targets for a single approved drug. This was beneficial since the intial stages of screening could be bypassed. Drug repositioning can be roughly categorized into categories namely drug based and disease based [41][106]. While the former methods initiate discovering associations from a chemical perspective, the latter is initialized from a pathological point of view or from a clinical perspective. Drug based studies focus on targets having similar binding sites by evaluating chemical-target interactions. Drug-disease based drug repurposing approaches aim at finding chemicals to new indications for which it was not originally approved for [62]. A few of the databases that help in both these types of approaches are MEDLINE [49], SIDER [76] and CTD [30] where a variety of data such as genomic and high throughput screening results are recorded. A rich set of information is also manually curated from the literature to establish inferred relationships [4][31].

PREDICT, a large scale prediction method of drug indications was proposed by utilizing drug-

drug and disease-disease similarities for prediction. The main contribution by the authors showed that by using disease specific genetic signatures, the accuracy of predicting drug indications for new diseases could be improved. This would mean that, drug treatments could be personalized for patients based on their genetic makeup instead of a generic disease signature. Cheng et al., developed three types of inference methods based on network theory to predict drug-target interactions for drug repurposing [24]. While the first two methods were drug-based and target-based similarity inferences, the third method was network-based inference. Each drug and target were represented as the nodes of a bi-partite graph and the nodes were connected based on the presence of an interaction between the drug and target. By propagating the scores in the network, new associations were discovered based on the score of the edges previously not defined. They showed that the network-based inference performed much better than the other two methods. The winning method was further enhanced by defining the edges as a weighted connection instead of an unweighted graph [22]. In all of these works, there has been a constant push to integrate various chemical and disease spaces to build better prediction models [100][25].

2.3 Multi-task Learning

Multi-Task Learning is a powerful concept in machine learning which states that learning multiple related tasks together is better than learning each individual task separately. This concept of learning related tasks drew inspiration from real life examples were learning similar tasks like riding a bicycle and a bike together is efficient than learning each of them separately. MTL has for this reason been used in a wide range of applications such as [115][19][46]. Over the years, MTL framework has evolved in various aspects to accommodate different improvisations to make the general framework better. Some examples are feature selection for similar tasks [5], automatic inference of task relationships [112][45] and structured input structured output [48]. With its wide reach to a lot of applications, MTL also found a strong foot in cheminformatics. In the paper titled "Machine Learning Methods for Property Prediction in Cheminformatics", the authors have sum-

marized ML techniques that have shown promise in cheminformatics for building structure-activity models, structure-property models and accounting for multiple molecular species among the many other uses [114]. The authors have elucidated the advantages of MTL over single task models in cases where obtaining positive labels is expensive and how a small set of labeled samples across multiple related tasks can help each other when learnt jointly. To further substantiate on this point, Geppert et al., showed that MTL was helpful in what is called "Orphan" screening in which some molecules do not have any ligand information [52].

Ning et al.,proposed an MTL framework for virtual screening using back-propagation neural networks [97]. The framework was developed to study structure-selectivity relationship (SSR) and the performance of their SSR-mt method performed substantially better than other baseline SSR models. Hughes et al., worked on a Deep Learning MTL network to accurately detect the binding sites and the probability of reactivity for small molecules with glutathione, cyanide, protein and DNA [61]. Apart from multi-task learning, MTL matrix completion was unearthed by Kshirsagar et al., for jointly learning protein interactions across related diseases namely Hepatitis C, Ebola and Influenza A [75]. The model learns a common low-dimensional subspace for task sharing in addition to task specific sturucture. In addition to the few studies listed above, MTL has been applied to improve prediction of cancer drug sensitivity and predict genetic traits [58][135]. The application of MTL specific to chemical-protein interaction has been explained in Chapter 3.

2.4 Multi-view Learning

A lot of applications that build predictive models have samples/observations that can be defined by more than one feature space. Each feature space is characterized by its statistical properties which might differ from other feature spaces. For example, in image classification, there might be a caption under the image that gives a short description and a main text that explains the image. Each of these text represent a different feature space. Combining the two feature spaces together

to build a model might not be effective in all cases since it might distort the characteristics of the feature spaces. Instead multi-view learning (MVL) treats each of these feature spaces as views and formulates the views to teach other than combining them into a single view. The basic assumptions of multi-view learning are as follows: (i) the views are conditionally independent (ii) each view is sufficient to build a predictive model and (iii) The views agree on the labels. MVL is a semi-supervised model where there there is a small set of labeled samples and a set of unlabeled samples for each view. The model learns a function for each view on the labeled samples and the view iteratively teach each other the labels on the unlabeled samples. The two popular strategies for MVL are (i) Co-training and (ii) Co-regularization.

In applications such as predicting drug-target interactions, protein-protein interactions and other such association predictions, it is common to represent a sample by multiple feature spaces such as chemical, biological and genomic spaces. A common strategy to learn from heterogeneous spaces was to combine them into a single feature space. This strategy has two main disadvantages. The first drawback is that, the statistical property of the spaces might not be preserved. The second disadvantage is that by combining more spaces for relatively the same number of samples, the dimension of the data might increase drastically. High dimensionality is a well known problem in the area of machine learning. Hence MVL combats both these drawbacks effectively and improves the performance of the predictive model. Although such strategies are relatively new to the chem-informatics community, Kang et al., used MVL for virtual screening [70]. The different views were integrated using rank aggregation and the experimental results showed that it is desirable to combine the views than learn a model on each feature space separately. A more detailed work on MVL has been explained in Chapter 3.

Another branch that stems from both multi-target and multi-view learning is multi-task multi-view learning that combines the properties of both the frameworks that learns related multiple tasks, each of which draws its data from different feature spaces. MVL and MTL have had very

limited studies in cheminformatics and our preliminary study of applying MTMVL to predict drug-disease association shows the promise of such frameworks. This dissertation aims at improvising the MTMVL framework by weighting the views based on its predictive power. We have come up with the weighted framework for two categories of problems. The first category of problems are those that need to build predictive models in a multi-view setting where the views agree with each other and the tasks are related to each other. The second category of problems is the application of weighted multi-view multi-task framework for truth discovery in a semi-supervised setting. Here the views might contradict each other and the challenge is to learn the view weights under the disagreement constraint. We also extend it as a multi-task problem where one task has known ground truth and the other task learns from weak truths. The details of the framework are discussed in detail in Chapter 8.

Chapter 3

Preliminary Study I: Investigating Multi-view and Multi-task Learning for Predicting Drug-Disease Associations

3.1 Introduction

Drugs exhibit their therapeutic effects by interacting and modulating one or multiple protein targets simultaneously; hence deeper insights of complex drug-disease-targets associations is of paramount importance in drug discovery. During the last decades the idea of drug re-purposing has been explored, where an old drug is utilized to treat a new disease. Unveiling potentially new and interesting drug-disease-target associations is a challenging task, considering the complexity of biological systems. Hence novel computational approaches are of high demand to analyze, disseminate and predict new interesting interactions, utilizing the growing body of data from different domains, which could then act as starting point for therapy development. Recently, the idea of data integration has been gaining momentum, where information from heterogeneous sources is combined with the expectation to provide additional information regarding the underlying links between drugs-diseases-targets, that otherwise would be difficult to study. In this study, we investi-

gate a new direction for studying drug-disease associations by utilizing and combining multi-view and mutli-task learning through integration of heterogeneous source of data. Our results show the advantages of exploring these methods to more effectively combine information from varied sources. We use multiple data sets to show the consistency of the results obtained.

3.2 Related Work

Drug discovery is a time and resource intensive process, hence novel approaches are required to study drug-protein interactions in a more efficient manner [86][40]. Drug-target deconvolution is a crucial step in drug discovery process since it provides valuable insights of drugs mode of action and leads to development of safer and improved therapeutic agents. Computational approaches can provide valuable inputs during decision making stages complementary to well established but expensive in-vitro and in-vivo methods for identifying and prioritizing promising candidates for further investigation. During the last decade availability of data regarding diseases, disease associated genes, biological pathways and drugs side-effects has been growing at an exponential rate in public repositories such as BindingDB [87], REACTOME [67] and KEGG[68]. Due to the long timeline and high costs involved in drug discovery, researchers turned to computational tools to speed up the process in the pipeline. Although computational tools are not meant to completely replace physical testing in drug discovery, they greatly help in reducing the time taken to zero in on the potential candidates by shortlisting them from a large pool of drugs. With an increase in the volume of such interaction information, new horizons opened up in research for the use of computational tools that helped in drawing conclusions regarding the interaction associations [26].

In addition to finding new drugs, there has been a lot of advancements in using indications of existing drugs for a newer disease for which it was not originally approved for [13][4][99][130]. Drug repurposing has been looked at from different angles in terms of how to amalgamate available data to find new associations or how to predict adverse effects of novel drugs [31]. Most of the

computer aided methods that were used for drug reuse were based on graph network analysis. A majority of the methods involved the construction of a network of known drugs, diseases, genes and targets whose interactions and associations were quantified using a scoring function [24][29]. Other than network analysis, machine learning techniques have also been used to build models for predicting drug-target interactions and for drug repurposing. For example, Yang et al. used the probability matrix factorization (PMF) method for drug repurposing that combined heterogeneous data to form a drug-disease association. The chain was defined by multiple factors namely, drug-target-pathway-gene-disease associations and a multi-level scoring system was used to predict the drug as being therapeutic, marker/ mechanism or both. A similar work of using PMF was studied by Cobanoglu et al. for predicting drug-target interactions with the assumption that a low rank subspace could capture large interaction networks [27].

3.2.1 Multitarget and Multi-task Learning

Drug repurposing is based very closely on the fact that a single chemical can have multiple targets. Promiscuity of drugs that was initially considered harmful was then exploited to model the multi-target property of drugs. Combination drugs also impacted multiple targets simultaneously and so algorithms that could predict and tap into this information gained popularity. One of the ways to tackle the multitarget scenario has been to utilize multi-label learning where the output learned is a binary vector. Afzal et al., [1] used a multi-class multi-label approach to model the target-ligand interactions in ChEMBL where each instance could have more than one label associated with it. They adopted a Naive Bayes framework in which a binary classifier was constructed for each (l,¬l). Similarly Cheng et al., followed a multi-label approach to model chemical-protein interactions by constructing binary classifiers on each label [24] . Multi-task Learning (MTL) has also been used in cheminformatics in terms of the multi-target (mt) problem where a drug has an interaction with multiple targets [44][148].

We can roughly divide the MTL work in cheminformatics into the following two broad cat-

egories namely (i) use of kernel methods and (ii) other methods with the former dealing more with the use of kernel functions and the latter encompassing a wide variety of MTL methods. We summarize a few methods here although the literature is not limited to these approaches. Ning et al., [96] proposed a method to capture the dependencies of related targets and ligands through target/compound specific kernels that were used simultaneously during the SVM learning process. Bickel et al., [11] was used for HIV therapy screening by using a drug feature kernel and virus mutation kernel as prior information in estimating the joint distribution of the data for each task. The prior information is based on the assumption that different drug combinations can have similar activity. ProdiGe is one of the popular methods for prioritizing disease genes [93] where heterogeneous information like phenotype similarity is shared across diseases in a positive and unlabeled learning setting. The amount of information that is shared is controlled by a kernel function using multi-task learning as the backbone.

Cheng et al., modelled the property of a drug's interaction with multiple targets as a multi-label problem by using a one-versus-the-rest classifying approach. This however did not exploit the actual multi-task learning capability. Similar to HIV, Alzheimer's is another disease that has been gaining popularity in terms of using multi-target techniques in finding potential targets [14][47][33]. Fang et al., also defined the mt chemical-protein interaction of ligands against Alzheimer's as a multi-label problem. Zhang et al., [139] used task regularized and boosted MTL for protein-chemical interaction prediction. In this paper, we use a regularized MTL framework for predicting drug-disease association.

3.2.2 Heterogeneous Data Integration and Multi-view Learning

The most crucial and important part of studying interactions of biomolecules for drug repurposing or to discover new associations is the quality of data at hand. With abundant data in the public domain, there has been a need to extract less noisy data and also to integrate the data from different sources. On one hand we have databases that are very specific to one kind of data and on the other

hand have databases that have a wealth of information that encompass a wide range of data characteristics. For example, SIDER [76]is a database that exclusively lists drugs and its side effects whereas CTD includes data regarding pathways, genes, drugs and their associations. One of the challenges in drug discovery that has been gaining importance in the past decade is to combat the problem of combining all these data in a way that is useful for gaining insights [142][109][128]. There has been a growing body of work where researchers found the usefulness of accommodating diverse data [101][63] and its potential to reveal exciting, new observations. Waller et al, [117] studied the techniques available at a systems level in terms of how large scale data are integrated, stored and retrieved and the cost associated with managing such a system. We try to focus on a different aspect of how to learn from this wealth of information to unearth unknown relationships.

Keeping this in mind, we focus on Multi-view Learning that addresses the issue of learning from multiple sources. Multi-view learning (MVL) takes into account the heterogeneous sources from which data is extracted and considers each source as a view. The basic assumption in MVL is that each view is adequate for building a predictive model and that is conditionally independent. The advantage of using MVL is that each feature space might have an underlying statistical property and concatenating them together might not be meaningful always. Yu et al., [134] proposed a multi-view setting for mining biomedical text records for gene prioritization. Different vocabularies are each considered as a view and the authors intended to show the promise of using MVL in a scenario when we do not for sure know which vocabulary is the best to use. They show that MVL is better than each single view learning. Virtual screening can involve multiple factors like ligand or target based screening or use interaction data. Kang et al., [70] showed that MVL can enhance the learning performance by the views teaching each other rather than learning form a single view. Similar studies were performed for gene clustering from microarray data [111] and for ligand based screening for drug discovery [143]. The latter paper discusses about how each platform can have different measurements and how MVL can exploit these heterogeneous data sources.

3.2.3 Multi-view Multi-task Learning

The third algorithm - multi-view multi-task learning (MVMTL) combines the advantage of both MVL and MTL by learning multiple views across multiple tasks. The driving motivation behind this learning is the fact that many real world data sets might involve multiple views as well as multiple tasks from which it can learn. For example, proteins can have features from multiple views such as their 3D structure and their sequence. They can also be characterized by their activity with protein families other than it's own and also by their functional activities. By learning from a diverse set of features, it is possible to unearth some unknown properties or activities of the proteins. Likewise, similar proteins when learned together act as extra information for each other which makes these type of algorithms very relevant for applications that can exploit information from varied sources to the maximum benefit of the task being learned. A graph based framework was proposed for MVMTL by [59] where each task could have their own specific view as well as shared view across different tasks. So for the example explained above, some proteins might have additional data from a source that other proteins might not have. This data could therefore be viewed as a task specific view whereas features such as the protein sequence could be a shared view assuming that the sequence for the proteins are known. Another framework for MVMTL was proposed by Zhang et al., [138] which differed from He's method in the fundamental basis that the former is a transductive learning method and the latter has an inductive learning setup. Zhang further showed the promise of this framework by utilizing it to predict adverse drug reactions [137]. Another significant work in this direction is by Jin et al., [66] who went a step further to propose a setting where tasks with multiple views and other related tasks need not have the same type of labels. In this study we have utilized the framework proposed by Zhang et al., due to its appropriateness to our application.

In this empirical study, we aim at building statistical models that are capable of predicting drug-disease associations and in particular cardiovascular and diabetes diseases. Our main motivation to work on these two diseases is due to the fact that they constitute for most of the diseases both in

developed and developing nations [105][136]. Studies have also shown a strong link between the two diseases in terms of how one disease might influence the other. For example, The Framingham study that was carried out for around 20 years showed the risk of diabetic patients developing cardiovascular diseases to be twice as much as people without diabetes in men and thrice as much in women [71]. Articles from the American Heart Association have stressed on the importance of addressing both the diseases together as an act of prevention [55]. Although diabetes acts only as an independent factor for heart diseases, the relatedness of the two diseases and the possibility of finding interesting and new observations between them make them the natural choice for our study.

3.3 Learning Methods

This study investigates three algorithms, namely the co-regularized MVL, regularized MTL and regularized MVMTL. We adopt the multi-view multi-task learning framework developed by Zhang et al. Equation 3.1 denotes the predictive function for each task which is nothing but the average prediction across all the views. For example, in our experiments we have three views and so $V = 3$ and for each task T, $f(\mathbf{X}_t)$ is the average prediction of the three views. Equation 3.2 represents the objective function of MVMTL framework where the first term is the least square loss function and the second term is the L2 regularization of the coefficients. The third term factors in the multi-view component of the algorithm. μ is the coupling parameter that regularizes the disagreement between different views. The fourth term accounts for the multi-task learning of the algorithm where γ is the parameter that regularizes the mapping function between different tasks.

$$f_t(\mathbf{X}_t) = \frac{1}{V}\sum_{v=1}^{V} f^v(x^v) = \frac{1}{V}\sum_{v=1}^{V} \mathbf{X}_t^v \mathbf{w}_t^v = \frac{\mathbf{X}_t \mathbf{w}_t}{V} \tag{3.1}$$

$$\min_{w_t^v} \sum_{t=1}^{T} \frac{1}{2}\|\mathbf{y}_t - \mathbf{X}_t\mathbf{w}_t\|^2 + \frac{\lambda}{2}\sum_{v=1}^{V}\|\mathbf{w}_t^v\|^2 + \frac{\mu}{2}\sum_{v\neq v'}^{V}\left\|\mathbf{U}_t^v\mathbf{w}_t^v - \mathbf{U}_t^{v'}\mathbf{w}_t^{v'}\right\|^2 + \frac{\gamma}{2}\sum_{t\neq t'}^{T}\left\|\mathbf{w}_t^v - \mathbf{w}_{t'}^v\right\|^2 \tag{3.2}$$

We derive the MVL and MTL algorithms from the above equation. By setting $\mu = 0$, Equation 3.2 reduces to a regularized MTL framework. Similarly by setting $\gamma = 0$, the equation reduces to a co-regularized MVL framework. We compare the performances of the three algorithms with k-Nearest Neighbors and Random Forest.

3.4 Experimental Study

This section explains in detail the data sets, methods and the results obtained for multi-view and multi-task learning algorithms and the comparison of its performance with other baseline methods.

3.4.1 Data Sets

We collected data from the Comparative Toxicogenomics Database (CTD) - a publicly available database that aims to interpret the effects of environmental hazards on human health [30]. The database has data that are manually curated as well as inferred chemical-gene, gene-disease, chemical-diseases and gene-pathway associations to name a few. The curated associations are extracted from published literature and the inferred associations are ascertained from these curated associations. For example, as shown in Figure 7.1 if there is a curated interaction between gene A and disease B and a curated interaction between gene A and chemical C, then CTD establishes an inferred association between chemical C and disease B.

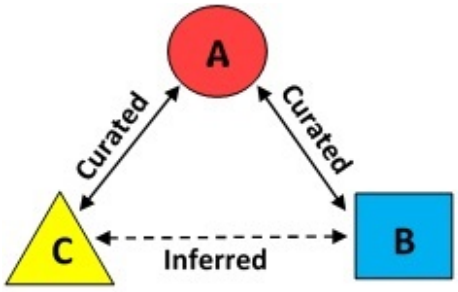


Figure 3.1: A pictorial represented of curated and inferred associations in CTD

The first data set consisted of five cardiovascular diseases namely Aortic Diseases, Heart Failure, Heart Valve Diseases, Arrhythmias and Heart Arrest. The second data set consists of four diabetic diseases namely Diabetic Angiopathies, Diabetic Cardiomyopathies, Diabetic Neuropathies and Diabetes Mellitus. For each data set the diseases represented the different tasks for MTL. We performed the following steps for each of the data sets to obtain the data-feature matrix and labels.

(i) We used the search tool in CTD to obtain information regarding the chemical and genes associated with it.

(ii) From the search results, we filtered out only the records that have experimental validation with direct evidence marked as therapeutic, marker/mechanism or both.

(iii) We pooled in this information from all the diseases in that particular data set to obtain a list of unique chemicals and unique genes.

- Two other data files : gene-pathway associations and drug-enriched pathway associations were downloaded from CTD.

(iv) The gene-pathway associations for the unique genes and the drug enriched pathway associations for the unique drugs involved in step (iii) were filtered out.

(v) If a drug and gene had an association in step (iii) and a gene and pathway had an association in step (iv), we then related the drug and pathway to have an association.

(vi) For each of the diseases in a data set we represented its activity (label) with the unique drugs with a +1 and an absence of activity with a -1.

Figure 3.2 and Figure 3.3 show the correlation of the diseases within the data set. For example, if we consider Figure 3.2, the diagonal of the matrix plot represents the histogram of each of the five diseases. The off-diagonals have the scatter plot of the (observation,value) pair and the correlation coefficient which is the slope of the least squares fit.

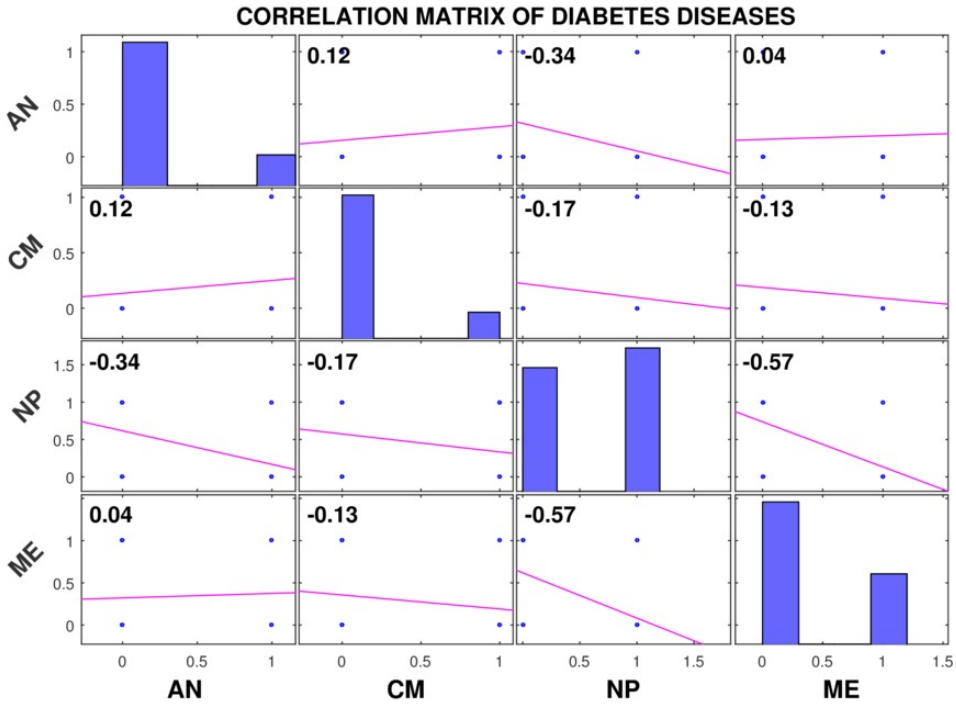


Figure 3.2: Label correlation of the cardiovascular diseases. AO - Aortic diseases, HF - Heart Failure, VF - Valve Defects, AR - Arrhythmias, HA - Heart Arrest

For both data sets, a total of three views was constructed. The first view consists of chemical-gene associations. The second view was an inferred association of drug-pathway obtained from known gene-pathway associations and the third view consists of drug-enriched pathway associations. The features of all the three views were 1 for an association and 0 otherwise. The details of the feature space and the total number of unique chemicals and the number of active compounds for each disease has been tabulated in Table 3.1 and Table 4.1 respectively. For example, Aortic and Heart Failure diseases have 20 and 198 active drugs respectively with a presence or absence of

Table 3.1: Characteristics of the features that form the three views for the data sets

Data set	Drug Gene Associations	Gene Inferred Pathway	Enriched Pathway
Cardiovascular	114	161	278
Diabetes	62	140	265

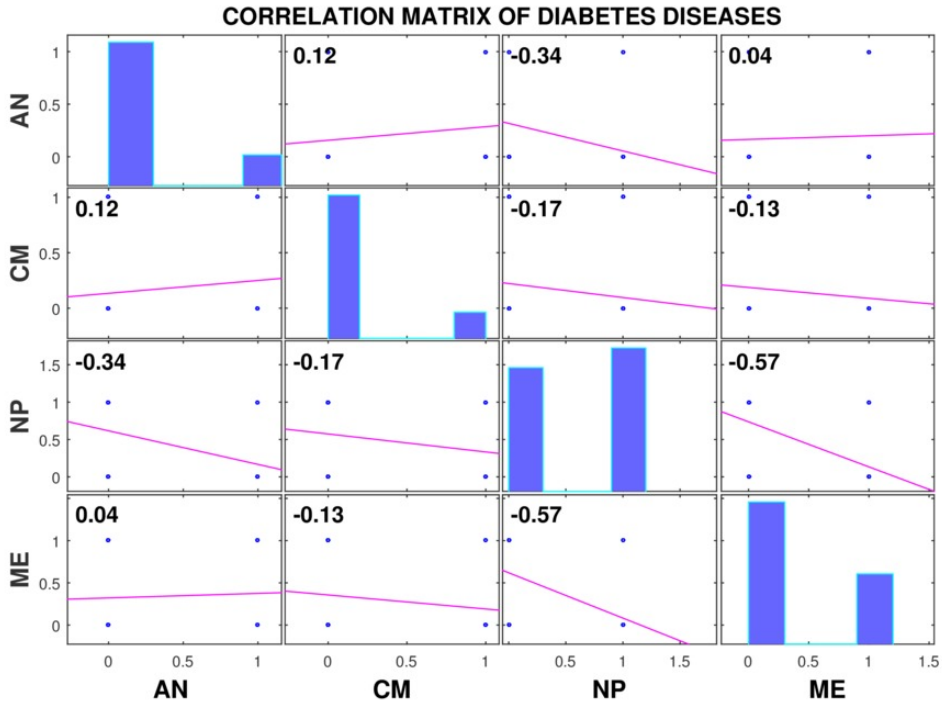


Figure 3.3: Label correlation of the diabetes diseases. AN - Angiopathies, CM - Cardiomyopathies, NP - Neuropathies, ME - Mellitus

association with 114 unique genes, 161 inferred pathways and 278 enriched pathways.

3.4.2 Model Construction and Evaluation

N samples are chosen at random from both the classes of the data set to form the labeled samples with roughly the same number of samples from each class. We then use five fold cross validation on the labeled data set to generate training and test data to achieve better generalization performance. The training data is further subjected to a five fold cross validation to choose the optimal model parameters based on the F1 score. We use the same data across all the algorithms. For MVL, we additionally sample some data for unlabeled data. Approximately N*3 samples are chosen to form the unlabeled set. The above procedure is repeated 10 times and the average F1 score has been reported.

Table 3.2: Disease Data Characteristics where the no.of drugs represent the total number of drugs and the third column represents the active drugs for each disease

Disease Name	No. of Drugs	No. of Actives
Aortic Diseases	408	20
Heart Failure		198
Heart Valve Diseases		89
Arrhythmias		241
Heart Arrest		19
Diabetic Angiopathies	136	24
Diabetic Cardiomyopathies		21
Diabetic Neuropathies		73
Diabetic Mellitus		45

We use the F1 score to measure the performance of the classifiers.

$$P = \frac{tp}{tp + fp} \tag{3.3}$$

$$R = \frac{tp}{tp + fn} \tag{3.4}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{3.5}$$

where P and R represent precision and recall respectively. tp, tn, fp and fn specify the true positives, true negatives, false positives and false negatives respectively.

3.4.3 Performance comparison

A typical method of handling heterogeneous data or different views has been to form a feature set with cardinality equal to the total number of features from all the views which in our case we would combine the three feature spaces given by Equation 3.6. We call this the combined feature space (CFS).

Table 3.3: The average F1 score of the five learning methods on the cardiovascular data set

Methods	Diseases				
	Aortic	**Heart Failure**	**Valve Defects**	**Arrhythmias**	**Heart Arrest**
MTL	0.798±0.010	0.765±0.011	0.698±0.010	0.682±0.013	0.751±0.013
MVL	0.751±0.005	0.739±0.009	0.672±0.010	0.651±0.010	0.736±0.012
MVMTL	**0.828±0.012**	**0.787±0.009**	**0.713±0.010**	**0.704±0.010**	**0.801±0.010**
kNN	0.725±0.004	0.709±0.005	0.623±0.005	0.613±0.005	0.705±0.006
RF	0.728±0.012	0.713±0.005	0.634±0.004	0.634±0.005	0.712±0.005

$$D_i = ds_i \bigoplus dg_i \bigoplus dp_i \tag{3.6}$$

In order to compare the performance of the MVL and MTL algorithms we choose two baseline algorithms that have been commonly used to build prediction models in cheminformatics namely k-Nearest Neighbors (kNN) and Random Forests (RF). Except MVL and MVMTL algorithms, the rest of the methods used the CFS.

3.5 Results

In this section, we present the results of our experiments performed on the two data sets. Table 3.3 shows the F1 score of the five cardiovascular diseases for the five learning methods explained as before. For both our data sets, the number of views V was equal to 3 and the number of tasks T was 5 for cardio diseases and 4 for diabetes respectively. The rows represent the learning methods and the columns of the table represent each disease in the data set. The method with the highest F1 score is printed in bold. For all the five methods, we have sampled the same number of positive and negative samples since we are not concentrating on the data imbalance problem.

As a first observation, we can see that the three MT and ML methods perform better than kNN and RF. Secondly, the trends across all the five methods are very similar. The higher or lower F1

Table 3.4: The average F1 score of the five learning methods on the diabetes data set

Methods	Diseases			
	Angiopathies	**Cardiomyopathies**	**Neuropathies**	**Mellitus**
MTL	0.628±0.008	0.623±0.006	0.753±0.008	0.652±0.007
MVL	0.618±0.004	0.577±0.002	0.721±0.010	0.637±0.006
MVMTL	**0.652±0.006**	**0.639±0.007**	**0.784±0.010**	**0.658±0.003**
kNN	0.592±0.005	0.559±0.005	0.660±0.004	0.570±0.005
RF	0.603±0.003	0.556±0.003	0.705±0.004	0.590±0.004

scores for each disease is due to the nature of data set itself and is not due to the learning methods used. If we consider aortic disease, MVL performs better than kNN and RF. But this particular task greatly benefits from MTL since the problem of low sample size can be overcome by learning multiple tasks.

Similar to the previous plot, Table 3.4 represents the F1 scores of the four diabetes diseases. Here again we observe consistent and better performance of the three methods in comparison with RF and kNN. Typically in drug reurposing, the number of approved drugs are only a handful (low positive sample) and hence multi-task learning techniques will help in alleviating this problem. As we saw in the previous section, the different drug activity profiles among the diseases act as additional data points when we adopt the MTL framework. This "extra" information helps in improving the performance of a classifier.

As a general rule of thumb, heterogeneous features are generally combined to form a single feature matrix. This was the method incorporated for all the methods except the multi-view learning framework. Out of the other four methods, two of them had the multi-task component and the rest adopt different classifying strategies. In order to compare the performance of MVL with the combined features space (CFS) within a similar framework we consider Equation 3.2 by setting $\gamma = 0$. For the CFS method, the unlabeled samples will not have a significance since we have only

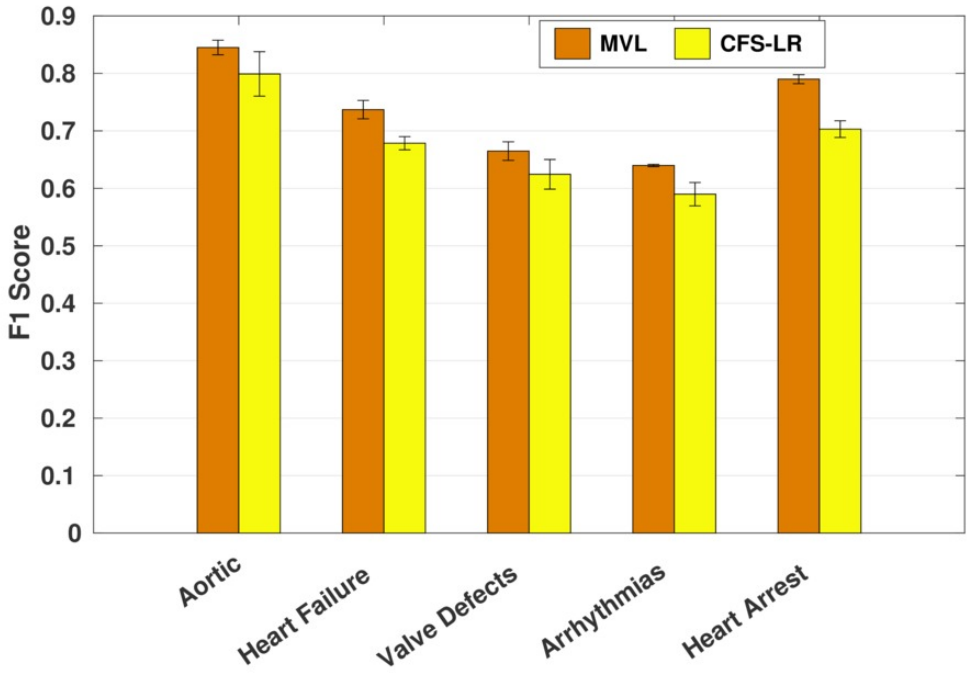


Figure 3.4: Comparison of MVL vs CFS for cardiovascular diseases

one view and for MVL we consider the three different views.

The CFS scenario basically reduces the objective function to a ridge regression framework. Figures 3.4 and 3.5 represent the F1 scores of MVL versus CFS for cardiovascular and diabetes diseases which are represented by the orange and yellow bars respectively. We can clearly see that MVL outperforms CFS although the margin is pretty close for some diseases like diabetes cardiomyopathies and mellitus. For the majority of the diseases, MVL exhibits an improved performance over the other method.

In order to show the advantage of using features from heterogeneous sources, the third set of plots (Figures 3.6 and 3.7) show the comparison of F1 scores between MVL versus using each feature space individually. The red bar denotes MVL and the yellow, green and blue color bars represent the gene features, gene inferred pathway features and enriched pathway features respectively. We observe that for a majority of the diseases, MVL method performs better than the

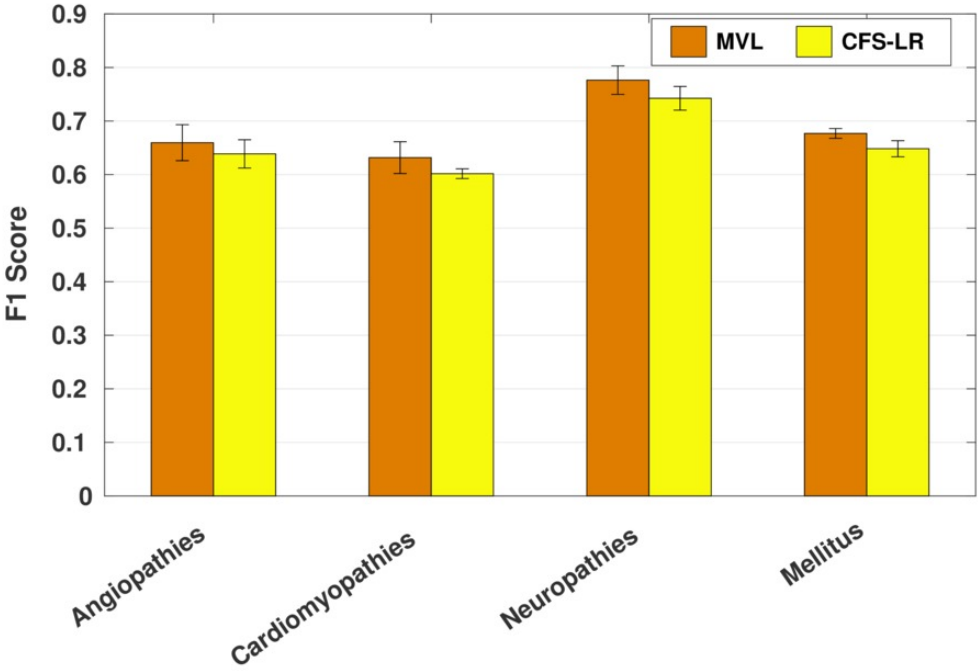


Figure 3.5: Comparison of MVL vs CFS for diabetes diseases

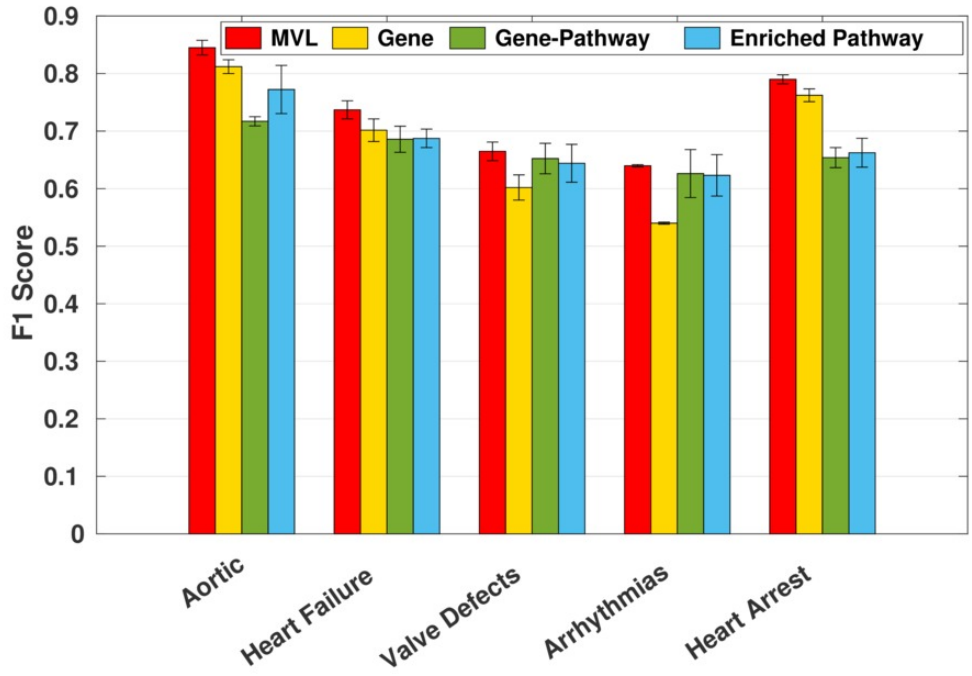


Figure 3.6: Comparison of MVL vs learning on individual features for cardiovascular diseases

individual features. We also notice the inconsistency of F1 scores among the three feature spaces. Some features work better than the rest for each disease and we do not get a clear conclusion regarding which is the best. From the F1 scores we see that including additional information only helps the classifier to predict better in all the cases.

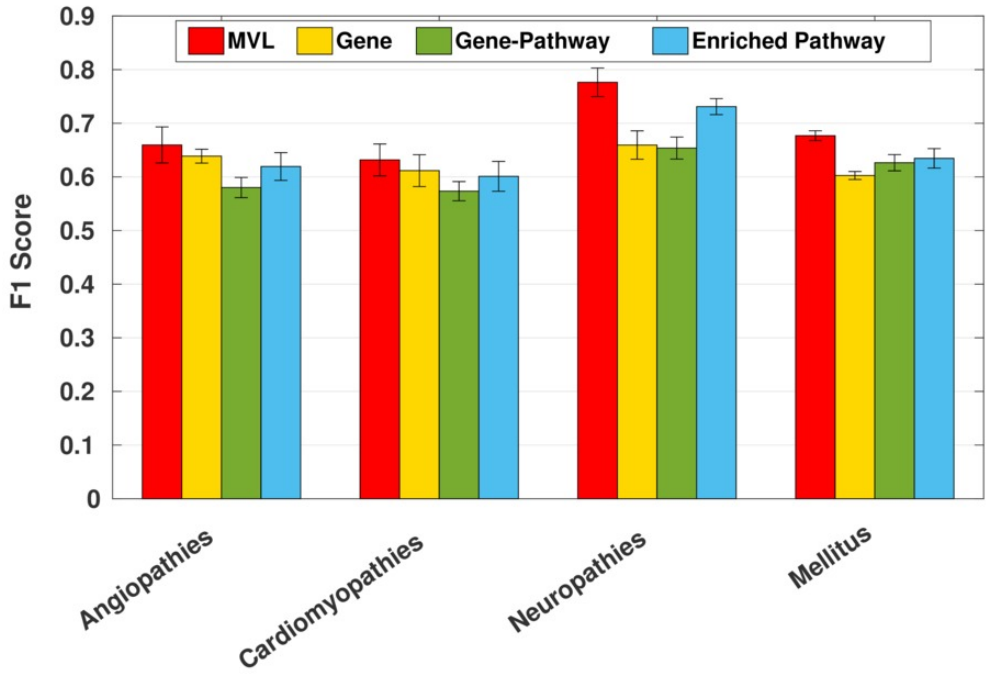


Figure 3.7: Comparison of MVL vs learning on individual features for diabetes diseases

3.5.1 Statistical Significance

Testing for statistical significance emphasizes our belief in the hypothesis that multi-task and multi-view algorithms perform better. This test is of importance especially when the difference in performance of the classifiers in terms of their F1 scores is marginal. Tables 3.5 and 3.6 represent the p-values obtained from the Wilcoxon test at the 95% confidence level for cardiovascular and diabetes diseases respectively. The first two columns in the table represent the two methods that are being compared and the third value represents the p-value obtained for the two methods using the Wilcoxon rank test. Our null hypothesis H_0 denotes that the algorithms are the same and our alter-

nate hypothesis H_a argues that the algorithms are different. Since we chose the significance level threshold to be 0.05, we reject the null hypothesis for a p-value ≤ 0.05. We performed the ranked test between MVL and CFS-LR (represents single task learning), MVMTL and MVL, MVMTL and MTL and lastly MTL and single task learning methods (kNN and RF). As an example, the p-value associated with the testing of MVL and CFS-LR for cardiovascular diseases is 0.003 which is ≤ 0.05 and so we reject the null hypothesis that the two algorithms are similar. From the tables we can see that the suggested multi-task and multi-view learning methods significantly outperform the other methods mentioned in our work. The scores reported here are based on the paired test between two methods across all the T tasks.

Table 3.5: Wilcoxon ranked test among the methods for cardiovascular diseases

Method 1	**Method 2**	**p-value**
MVL	CFS-LR	0.003
MVMTL	MTL	0.040
MVMTL	MVL	0.038
MTL	kNN	$4.35e^{-10}$
MTL	RF	$6.88e^{-8}$

Table 3.6: Wilcoxon ranked test among the methods for diabetes diseases

Method 1	**Method 2**	**p-value**
MVL	CFS-LR	0.046
MVMTL	MTL	0.002
MVMTL	MVL	0.015
MTL	kNN	$7.98e^{-4}$
MTL	RF	$2.51e^{-5}$

3.6 Conclusion

In this empirical study, we explored a new direction of multi-view and multi-task learning algorithms for studying drug-disease associations. These algorithms provide new avenues to be explored for applications in cheminformatics where there is a need to utilize better computational models in addition to another important aspect of integrating data in a useful manner. The advantages of using these methods are (i) unifying information from multiple resources (ii) learning similar tasks to find new associations.

Multi-view learning has been the major study in this paper due to its powerful capabilities to accommodate and extract useful information from multiple heterogeneous sources. With researchers finding every extra piece of information crucial in understanding the behavior of drugs that would help in understanding their associations, there is always a need to come up with methods that will jointly learn these type of data. We showed that each feature space might perform better in some occasions but the MVL algorithm consistently performs better thereby eliminating the need to assess and iterate through the different feature spaces. For example, the drug-gene associations might provide better predictive performance than the drug-enriched pathway associations. But it cannot be affirmatively said that this would be the case. Also, other information could hold key clues that might provide additional insight. The aim is therefore to maximize the predictive learning capabilities of algorithm to effectively infer associations. The other biggest advantage of MVL is the utilization of unlabeled samples. Since the learning involves minimizing the view disagreement on the unlabeled samples, the algorithm proves to be powerful in capitalizing on this information.

Multi-task learning on the other hand has been adopted by the bioinformatics community mostly in terms of modeling the multitarget property of drugs thereby improving the potential to find new targets. Multitarget problems are dealt in a multilabel setting which differs from MTL where similar diseases can be learned jointly. This setting is particularly useful when we have limited number of samples to learn from. The "extra" information from the other tasks have shown

to combat the problem of small sample sizes. In this study, we have shown that the joint learning of related diseases benefit the overall performance of each disease. By combining the advantages of MVL and MTL, the scope of integrating data as well as finding new potential targets is tremendous and our statistical testing also reinforces our hypotheses. The MVMTL algorithm explained in this study has the capability to also model task relationship which means it can form clusters of similar tasks/diseases. As part of our ongoing research, the next step would be to integrate additional information such as drug-side effects relationships and also take advantage of cross-disease information in order to further improve predicted drug-disease associations.

Chapter 4

Weighted Multi-view Learning for Predicting Drug-Disease Associations

4.1 Introduction

The paradigm of drug discovery has moved from finding new drugs that exhibit therapeutic properties for a disease to reusing existing approved drugs for a newer disease. The association between a drug and a disease involves a complex network of targets and pathways. In order to provide new insights, there has been a constant need for sophisticated tools that have the potential to discover new associations from the underlying drugs-disease interactions. In addition to computational tools, there has been an explosion of data available in terms of drugs, disease and their activity profiles. On one hand, researchers have been using existing machine learning tools that have shown great promise in predicting associations but on the other hand there has been a void in exploiting advance machine learning frameworks to handle this kind of data integration. In this paper, we propose a new learning framework called weighted multi-view classification that is a variant of the well-known Multi-view learning framework. The primary motivation behind this method is that, some descriptors help in prediction better than others. In a multi-view setting each type of descriptor is represented as a view and we hypothesize that not all the views contribute equally to the

final prediction. The proposed learning framework handles this discrepancy by learning weights for each view and the final prediction is the weighted average of the prediction across all the views. The effectiveness of the method especially to predict drug-disease associations is demonstrated by comparing it with well-established methods on two different disease data sets.

4.2 Related work

The concept of drug discovery has been around for many decades during which a lot of research in discovering new drugs has advanced the field of medicine by leaps and bounds [40]. The study of interaction between small molecules and bio-molecules has gained importance since it forms the crucial step in drug design [77][149]. However, the entire process of screening compounds to getting the drugs approved is a very time consuming process and researchers turned towards tools and methods that fastened this process. With the rapid accumulation of interaction data in public databases such as BindingDB[87], KEGG[69] and Protein Data Bank[9], computational tools gained popularity to exploit these large databases to build predictive models that helped in speeding up the process [42][57]. Although computational tools do not serve as a substitute to physical testing, they have been a powerful tool with the capability to short list the potential candidates or unravel new insights that have previously been unknown [27].

4.2.1 Drug Repurposing

While discovering new drugs was actively growing, a new concept of reusing existing approved drugs gained popularity for diseases for which it was not originally approved for [92]. Finding new associations between approved drugs and new diseases is advantageous in terms of time, money and other resources used during the development. The concept of one drug - one disease soon moved to one drug - many diseases. One of the popular methods that has been used to achieve reusing old drugs for new diseases is molecular docking [60][90] which takes in the structure of compounds and targets and quantifies their binding. Computational tools and in particular machine

learning algorithms have also played a vital role in establishing such new associations or have also been used as a tool to rank drugs based on their potential to be reused [73][54]. One strategy that was commonly adopted to find such new associations was to construct bipartite graphs between the drug and target and quantify their interactions by a score [24][23]. These scores acted as the labels that were used to build a predictive model or these interactions were modeled as a graph network where the edges were weighted by the interaction scores.

4.2.2 Data Integration and Multi-view Learning

With an explosion of cheminformatics data available in the public databases [30][72] and there has been a lot of excitement in utilizing the abundant information in better understanding the underlying activities of drugs, the targets and pathways they modulate and also gain new insights from these networks in terms of reusing approved drugs for newer diseases. In order to gain usefully from this wealth of information, we need to resort to advance learning methods that more effectively handle and integrate such high dimensional data with a relatively low sample size [142][128][23]. The traditional practice has been to integrate data from different sources and then utilizing an algorithm for either classification or drawing conclusions about associations and interactions. As we integrate more information, there is a possibility of the number of dimensions increasing drastically with the number of samples being relatively the same. In order to derive the utmost benefit from the data, multi-view learning has shown promise in combating this issue by considering each feature space as a view. One of the works that dealt with integrating data from multiple resources includes the work by Yang et al, who used probability matrix factorization to establish a drug-disease association by introducing a scoring method for each drug-tagret-pathway-gene-disease chain [130]. Yu et al., showed the effectiveness of multi-view learning by proposing a framework that mines biological text for gene prioritization [134]. Different vocabularies were considered as different views and they showed that MVL methods performed better than single-view learning. To summarize, each view or feature space might have an underlying useful statistical property that needn't be always preserved while combining them into a single

view.

4.2.3 Weighted Multi-view Learning

Weighted multi-view learning in the literature has been limited to clustering methods. The initial work by Tzortzis et al., explored this gap of unevenly weighting the views for clustering where each view is expressed as a kernel matrix and the final learning is a weighted average of the kernels [113]. This was further extended by Xu et al., who incorporated feature selection in addition to weighting the views [126]. Jiang et al., went on to further weight the features too instead of just performing a feature selection [65]. Denoising in multi-view learning has been another popular method to handle errors in view data especially in image classification that could affect the learning [141]. An image of a white siberian tiger could be wrongly labeled as a zebra and vice versa. In such cases, denoising is a very important strategy to help in learning from such corrupted data [140]. Our motivation is slightly different wherein we assume that not all views are equally important and if there are noisy views, they must have a lower weight in the final prediction. Since weighted multi-view learning has not been explored for classification/regression to the best of our knowledge, we proposed the weighted multi-view learning for classification primarily motivated by the fact that data integration in cheminformatics is highly desirable and needs a systematic approach to handle multiple views and/or noisy data.

For our real data in this study, we have considered the cardiovascular and diabetes diseases since they are two important diseases that are plaguing both developed as well as developing countries [136][105]. Studies have also shown a relation between the two diseases and how people with diabetes are more prone to develop a cardiovascular disease [71]. The availability of data for these two diseases in addition to their importance convinced us to use them for our study.

4.3 Methodology

In this section, we propose a weighted multi-view learning framework that automatically learns the weights of each view so that the final prediction is a weighted sum of the predictions of each view as opposed to the average prediction across all the views.

4.3.1 Notations

In this paper, we used bold uppercase letters to represent a matrix (e.g $\mathbf{X}$) and bold lowercase letters to represent a vector (e.g. $\mathbf{x}$). Greek letters are used to represent regularization parameters (e.g λ), simple lower and uppercase letters are used to represent scalars (e.g x, X). We use the subscript v to denote a view. For example, if we have V views, $\mathbf{X}_v$ represents data $\mathbf{X}$ from view v where $v \in V$. All the vectors are column vectors unless specified.

4.3.2 Overview of the Learning Framework

We formally define the problem setting as follows. Let there be n labeled samples and let their representation in view v be represented by $\mathbf{X}_v \in \mathbb{R}^{n \times d}$ and their labels be represented by $\mathbf{y} \in \{-1,+1\}^{n \times 1}$. We also assume that all the views agree upon the label for a particular observation. The fundamental assumption of co-regularized multi-view learning is that the multiple views involved in prediction are conditionally independent and each view is capable of generating their own prediction model. The other major assumption is that there is class consistency of the labeled samples across all the views. As shown in Equation5.1, the typically used co-regularized multi-view learning considers each view to contribute equally. $f_v(\mathbf{X}_v)$ represents the function learned on the data from view v and the final prediction $f(x)$ is the average of the prediction results from V views. If we consider $f_v(\mathbf{X}_v) = \mathbf{X}_v\mathbf{w}_v$ and using the least square loss function, we have the objective function given by Equation 5.4. The first term is the least square loss function, the second term represents the ℓ_2 norm of the model parameters while λ_1 controls the strength of the norm. The third term represents the view disagreement of different views controlled by the coupling parameter

μ.

$$f(x) = \frac{1}{V} \sum_{v=1}^{V} f_v(\mathbf{X}_v) \tag{4.1}$$

$$\begin{aligned} \min_{w_v} \ & \frac{1}{2} \sum_{v=1}^{V} \left(\mathbf{y} - \frac{\mathbf{X}_v \mathbf{w}_v}{V} \right)^2 + \frac{\lambda_1}{2} \sum_{v=1}^{V} \|\mathbf{w}_v\|^2 + \\ & \frac{\mu}{2} \sum_{v \neq v'}^{V} \|\mathbf{U}_v \mathbf{w}_v - \mathbf{U}_{v'} \mathbf{w}_{v'}\|^2 \end{aligned} \tag{4.2}$$

In our proposed weighted multi-view learning (wMVL), we weight the predictive function of each view by introducing a parameter β_v as shown in Equation (5.2). We additionally add the constraints that $\sum_{v=1}^{V} \beta_v = 1$ and $\beta_v \geq 0$. In addition to the model coefficients, the algorithm automatically learns the weights thereby eliminating the need for any prior knowledge. By including the constraints for the view weights, we obtain the objective function for weighted multi-view learning as given by Equation (4.3).

$$\begin{aligned} \min_{\beta v, w_v} \ & \frac{\beta_v^p}{2} (\mathbf{y} - \mathbf{X}_v \mathbf{w}_v)^2 + \frac{\lambda_1}{2} \sum_{v=1}^{V} \|\mathbf{w}_v\|^2 + \\ & \frac{\mu}{2} \sum_{v \neq v'}^{V} \|\mathbf{U}_v \mathbf{w}_v - \mathbf{U}_{v'} \mathbf{w}_{v'}\|^2 + \lambda_2 (\textstyle\sum_v \beta_v - 1) \end{aligned} \tag{4.3}$$

From Equation 4.3, we see that the proposed method is similar to a weighted least squares excepting that we weight the views in this case rather than the samples given by $\mathbf{y}' = \mathbf{X}'_v \mathbf{w}_v$ where $\mathbf{y}' = \beta_v^{\frac{p}{2}} \mathbf{y}$ and $\mathbf{X}'_v = \beta_v^{\frac{p}{2}} \mathbf{X}_v$. For a new sample, the final prediction by combining all the views is given by Equation 5.2.

$$f(x) = \sum_{v=1}^{V} f_v(\mathbf{X}_v) = \sum_{v=1}^{V} \mathbf{X}'_v \mathbf{w}_v \tag{4.4}$$

We solve the optimization problem of wMVL by alternatively updating $\mathbf{w}_v$ and β_v. The solution for updating the parameters is obtained by computing the partial derivative of the objective function L

with respect to each w_v and β_v.

$$\frac{\partial F}{\partial \mathbf{w}_v} = \beta_v^p \mathbf{X}_v^T (\mathbf{X}_v \mathbf{w}_v - \mathbf{y}) + \lambda_1 \mathbf{w}_v + \mu(V-1)\mathbf{U}_v^T \mathbf{U}_v \mathbf{w}_v - \mu \mathbf{U}_v^T \sum_{v' \neq v} \mathbf{U}_{v'} \mathbf{w}_{v'} \tag{4.5}$$

Rearranging Equation (5.5) and setting it zero, we get a solution for $\mathbf{w}_v$.

$$(\beta_v^p \mathbf{X}_v^T \mathbf{X}_v + \lambda_1 + \mu(V-1)\mathbf{U}_v^T \mathbf{U}_v)\mathbf{w}_v = \beta_v^p \mathbf{X}_v^T \mathbf{y} + \mu \mathbf{U}_v^T \sum_{v' \neq v} \mathbf{U}_{v'} \mathbf{w}_{v'} \tag{4.6}$$

$$\boxed{\mathbf{w}_v = Ft^{-1} St}$$

$$\text{where} \qquad Ft = \beta_v^p \mathbf{X}_v^T \mathbf{X}_v + \lambda_1 + \mu(V-1)\mathbf{U}_v^T \mathbf{U}_v \tag{4.7}$$

$$St = \beta_v^p \mathbf{X}_v^T \mathbf{y} + \mu \mathbf{U}_v^T \sum_{v' \neq v} \mathbf{U}_{v'} \mathbf{w}_{v'}$$

Taking partial derivative of the objective function L with respect to β_v, we get the following equations that is used to update the parameter.

$$\frac{\partial F}{\partial \beta_v} = \frac{p\beta_v^{p-1}}{2}(\mathbf{y} - \mathbf{X}_v \mathbf{w}_v)^2 + \lambda_2 \tag{4.8}$$

Setting Equation (5.8) to zero we get,

$$\beta_v = \left(\frac{2\lambda_2}{p(\mathbf{X}_v \mathbf{w}_v - \mathbf{y})^2} \right)^{1/(p-1)} \tag{4.9}$$

In order to eliminate λ_2 which is one less model parameter to worry about, we substitute (5.9) into the constraint $\sum_{v'=1}^{V} \beta_{v'} = 1$, we get

$$\sum_{v'=1}^{V} \left(\frac{2\lambda_2}{p(\mathbf{X}_v \mathbf{w}_v - \mathbf{y})^2} \right)^{1/(p-1)} = 1 \tag{4.10}$$

Rearranging (5.10), we get

$$(2\lambda_2)^{1/(p-1)} = \frac{1}{\dfrac{1}{\sum\limits_{v'=1}^{V} p(\mathbf{X}_v\mathbf{w}_v - \mathbf{y})^{2/(p-1)}}} \tag{4.11}$$

Substituting (5.11) in (5.9) and rearranging, we get an update rule for β_v

$$\boxed{\beta_v = \frac{1}{\sum\limits_{v'=1}^{V} \left(\dfrac{\mathbf{X}_v\mathbf{w}_v - \mathbf{y}}{\mathbf{X}_{v'}\mathbf{w}_{v'} - \mathbf{y}}\right)^{2/(p-1)}}, \quad p > 1} \tag{4.12}$$

When $p = 1$, the weights are less than 1 given it's constraint of summing to 1. We can get

$$\boxed{\beta_v = \begin{cases} 1, & v = \arg\min\limits_{v'} D_{v'} \\ 0, & \text{otherwise} \end{cases}} \tag{4.13}$$

In Equations (5.12) and (4.13), p is an exponential parameter that controls the sparsity of the view weight vector β. The value of p can be chosen based on the number of views and it's significance is further explained in the experimental section. The idea behind the weight vector is that, the more useful a particular view is, higher is the weight assigned to that view.

4.4 Experimental Study

In this section, we explain in detail the data sets that have been used to test our hypothesis. We used synthetic data as well as real cheminformatics data to study the effectiveness of the proposed method. By performing these experiments, we analyze the role of the weight parameter β as well as the parameter p that controls the sparsity of the weights.

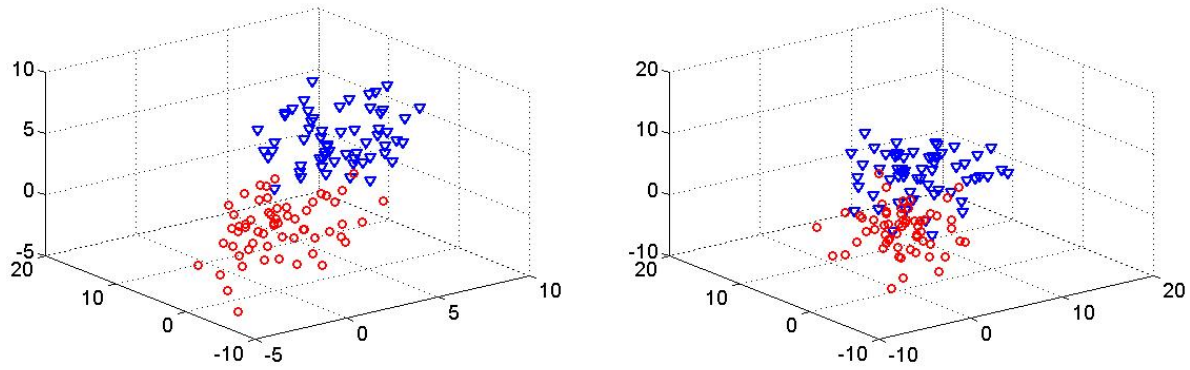
Algorithm 1 Weighted Multi-view Algorithm

Input: $\mathbf{y}$, $\{\mathbf{X}_v\}_{v=1}^{V}$, $\{\mathbf{U}_v\}_{v=1}^{V}$ λ_1, μ, N_{it}, ε
Output: $\{\mathbf{w}_v\}_{v=1}^{V}$, $\{\beta_{v0}\}_{v=1}^{V}$
Initialize $w_{v0} = 0$ and $\beta_v = \frac{1}{V}$ for $v \in [1:V]$
for $iter = 1$ to N_{it} **do**
    for $v = 1$ to V **do**
        Compute Ft as given by Eqs.(7)
        Compute St for every $v' \neq v$ as given by Eqs.(7)
    end for
    Compute $\mathbf{w}_v := Ft^{-1}St$ for each $v \in [1:V]$
    Update β_v using Eqs.(12) or (13) for each $v \in [1:V]$
    $\|\mathbf{w}_v - \mathbf{w}_{v0}\| < \varepsilon$ & $\|\beta_v - \beta_{v0}\| < \varepsilon$
    break
    $\mathbf{w}_{v0} := \mathbf{w}_v$ for each $v \in [1:V]$
    $\beta_{v0} := \beta_v$ for each $v \in [1:V]$
end for
Return $\mathbf{w}_v$ and β_v

4.4.1 Data sets

4.4.1.1 Synthetic Data

To explain the fundamental working of the algorithm, we created a simple synthetic data set with two views. The first view consists of samples from a normal distribution with each class having a different mean and standard deviation. As shown in Figures 4.1a and 4.1b, the first view has the two classes fairly separable and the second view is nothing but a noisy version of view 1. Both the views had 60 samples in each class with 3 features for easy visualization. In the usual regularized multi-view learning setup, the final prediction is the average of the two views (which means each view will have a weight of 0.5) but we can clearly see that view 1 is more useful than view 2. According to our hypothesis, the proposed wMVL will learn a higher weight for view 1 as compared to view 2 thereby weighting the prediction of view 1 more than view 2. The synthetic data is just a simple model to show the effectiveness of the proposed method that can be extended to more views.



(a) View 1 **(b)** View 2 - Noisy

Figure 4.1: The two views of the synthetic data. The circle and the triangle symbols represent the two classes.

4.4.1.2 Drug-disease Data

For our real life data set,we wanted to apply the proposed method to predict drug-disease associations which was the driving motivation for wMVL. Data was primarily collected from the Comparative Toxicogenomics Database (CTD). It is a public database that works on establishing the effects of environmental hazards on human health (insert ref). The database is manually curated to reference direct as well as inferred relationships between drugs, genes and pathways to name a few. The direct associations are curated from published literature and the inferred associations are obtained from these curated information. Figure 7.1 illustrates the CTD database. For example if Gene C and Drug A have a curated association and Drug A and Disease B have a curated association, then CTD defines an inferred relationship between Gene C and Disease B. In addition to CTD, data was also extracted from DrugBank and PubChem which are two widely used databases to extract the structures of drugs. We worked with two different disease data sets namely cardiovascular diseases and diabetic diseases. We chose multiple diseases to show the consistency of our results. The cardiovascular dataset consisted of three diseases - Arrhythmias, Heart Arrest and Heart Failure and the diabetic data set consisted of three diseases - Diabetes Mellitus, Diabetes Neuropathies and Diabetic Cardiomyopathies.

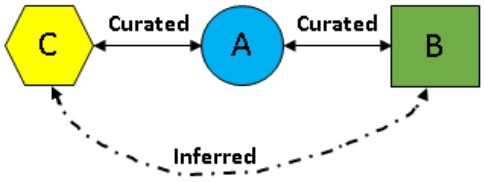


Figure 4.2: CTD

Both the data sets had three types of features/views namely: Structural descriptors, drug-gene associations and drug-pathway associations. The following procedure was used to obtain our final data-feature matrix:

(i) The CTD "Search" option was used to obtain all the chemical-gene of the diseases.

(ii) From the associations obtained, only the associations that had experimental validation with therapeutic, marker|mechanism or both as direct evidence made it to the next stage.

(iii) For each data set (cardio or diabetes) the number of unique drugs and genes were pooled together.

(iv) For these drugs, their pathway information was extracted from CTD.

(v) A drug-gene matrix was constructed with a +1 for an association and a 0 otherwise representing the drugs in the gene feature space.

(vi) A drug-pathway matrix was similarly constructed that formed the second feature space.

(vii) The structures for all the drugs were downloaded from DrugBank and PubChem.

(viii) In order to explore different features, we extracted the EFP fingerprints for the cardiovascular drugs and 2D structures for the diabetes drugs.

(ix) The presence of an association of a disease and a drug was labeled as +1 and -1 otherwise. Table 4.1 shows the number of active and inactive drugs for each disease.

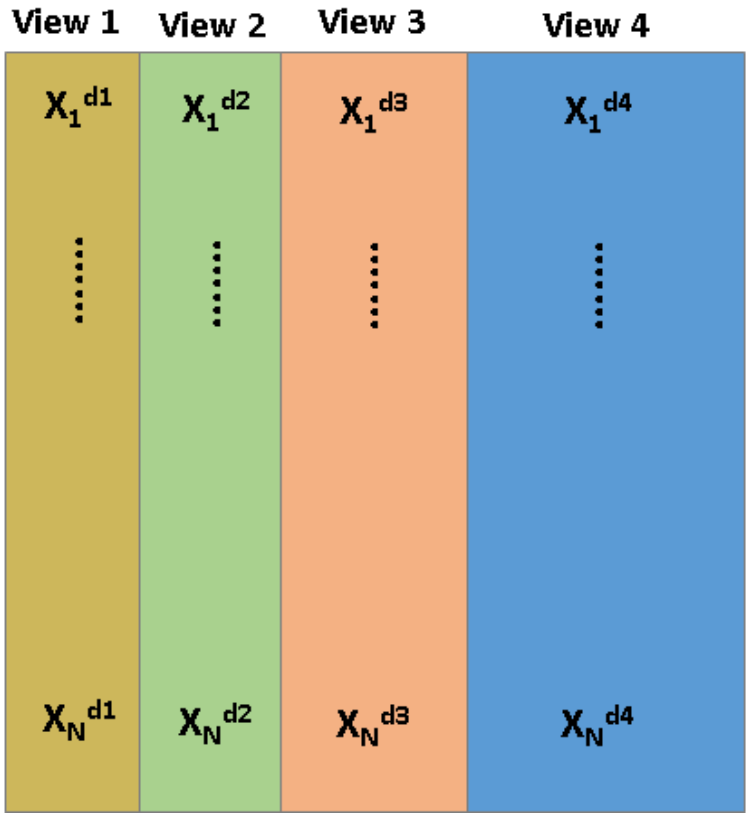


Figure 4.3: A representation of the four views of our drug-disease data set. Each view has N samples where X_i^{dj} represents sample i from view j. The total number of features for each data set in the sum of $d1, d2, d3$ and $d4$

4.4.2 Algorithms/Learning Frameworks

The performance of our proposed wMVL is compared with five other learning methods/algorithms. The primary method with which we study the effectiveness of wMVL is the co-regularized multi-view learning method given by Equation (5.4). We also compare it with other benchmark single view learning (SVL) algorithms such as k-Nearest Neighbors (kNN), Random Forest (RF) and the Regularized Least Squares Method (Ridge). The general rule of thumb for handling heterogeneous data by single task learning algorithms has been to combine them into a single feature space with its cardinality equal to the sum of the features from all the views. In our case, we have three feature spaces namely the structural features (ds), gene features (dg) and the pathway features (dp). All of the four SVL algorithms combine the three spaces as given by Equation (4.14). We call this the Combined Feature Space (CFS).

$$D = ds \bigoplus dg \bigoplus dp \tag{4.14}$$

Disease Name	No. of Drugs	No. of Actives
Aortic Diseases	408	20
Heart Failure		198
Heart Arrest		19
Diabetic Angiopathies	136	24
Diabetic Cardiomyopathies		21
Diabetic Mellitus		45

Table 4.1: Disease Data Characteristics where the no.of drugs represent the total number of drugs and the third column represents the active drugs for each disease

4.4.3 Model Construction and Evaluation

4.4.3.1 Model Construction and Selection

For the multi-view learning methods, 'N' samples are chosen at random from each class of the data set to form the labeled samples. We sample them in such a way that both classes have the same number of positive and negative samples since our focus is not the problem of data imbalance. We also sample approximately N*3 samples to represent the unlabeled data in the MVL setting. The labeled samples are then subjected to a five fold cross validation that results in training and test data that helps in better generalization. The training data is further subjected to a five fold cross validation to get the training and validation data sets that are used to select the model parameters. For the single-view learning methods, we repeat the same process of obtaining our training, validation and test data sets as the MVL methods with the only exception being, SVL does not handle unlabeled samples. The model parameters are selected through a grid search for each method. We varied the number of neighbors for kNN from 3 to 9 in steps of 2. In the case of RF, the only

parameter we considered was the number of features used to make the split and from the literature we considered the square root of the total number of features as a good estimate for the number of features.

4.4.3.2 Model Evaluation

We used the F1 score to measure the performance of the different classifiers.

$$P = \frac{tp}{tp + fp} \tag{4.15}$$

$$R = \frac{tp}{tp + fn} \tag{4.16}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{4.17}$$

tp, fn and fp are true positive, false negative and false positive respectively.

4.5 Results

In this section the results of our experiments performed on the synthetic and real life data sets are presented. The performance of the learning methods have been analyzed and finally we have included the results of statistical testing.

4.5.1 Performance Comparison

4.5.1.1 Synthetic Data Set

The synthetic data set consisted of 60 samples in each class with three features sampled from a normal distribution. As explained earlier, we created two views with view 2 being a noisier version of view 1. The objective function given by Equation (4.3) has an exponential parameter p that is used to adjust the sparsity of the weights across the views. As the value of p is increased, the view weights tend to become more uniform and might degrade the performance of the model. The

choice of p depends on the number of views. As the weights become more uniform, the performance of the classifier will be similar to that of the regular co-regularized multi-view learning.

p Value	Coefficients	
	View 1	View 2
$p = 1$	1	0
$p = 1.5$	0.7926	0.2074
$p = 2$	0.6577	0.3423
$p = 5$	0.5498	0.4502
$p = 10$	0.5307	0.4693

Table 4.2: The view weights learned by wMVL on the synthetic data set with two views

From Table 4.2 we see that when $p = 1$, based on Equation (4.13) the view with the lease error will be assigned a weight 1 and the other views are given a weight 0. For our synthetic data we know that view 1 can definitely make a better prediction than view 2 and hence they are given weights 1 and 0 respectively. As we increase the value of p , the weights start getting more uniform among the views. But we can see that view 1 still has a marginally higher weight than view 2 when $p = 10$ which makes a very convincing argument given the two views. Table 4.3 summarizes the F1 score of the of wMVL and MVL for the five values of p shown in Table 4.2. We can see that when $p = 1$ only one of the views is involved in making a prediction and so the F1 score of wMVL is not better than MVL. But as the value of p is increased, we can observe that the F1 score of wMVL increases and performs better than MVL. We can see that for p values equal to 5 and 10, the weights are more uniform between the two views and yet due to the marginal difference in weights, wMVL performs better than MVL.

p Value	wMVL	MVL	Ridge	kNN	RF
$p = 1$	0.983 ± 0.002	0.973 ± 0.019	0.965 ± 0.011	0.933 ± 0.003	0.826 ± 0.018
$p = 1.5$	0.989 ± 0.012	0.980 ± 0.006	0.973 ± 0.001	0.978 ± 0.011	0.913 ± 0.020
$p = 2$	0.992 ± 0.013	0.966 ± 0.004	0.958 ± 0.019	0.952 ± 0.008	0.934 ± 0.015
$p = 5$	0.991 ± 0.014	0.978 ± 0.010	0.966 ± 0.017	0.963 ± 0.005	0.897 ± 0.026
$p = 10$	0.988 ± 0.021	0.962 ± 0.009	0.964 ± 0.022	0.952 ± 0.006	0.907 ± 0.028

Table 4.3: Comparison of the F1 scores on the Synthetic Set

4.5.1.2 Drug-Disease data set

For the drug-disease data, we have four different views namely, 2D structure, fingerprints, pathway and gene. The MVL would weight all these views equally or in other words, the weight for each view would be 0.25. Tables 4.4 and 4.5 compare the F1 scores of the methods for the two diseases. For wMVL, we had five different p values namely 1, 1.5, 2, 5 and 10. Due to space constraints we present only the results for values 1, 2 and 5 since there was not a very huge difference in the results between $p = 1.5$ and $p = 2$ and between $p = 5$ and $p = 10$. Figures 4.4a and 4.4b show the weight distribution among the views for different p-values. The x-axis represent the p values and the y-axis represent the value of the weights. As an example, if we consider the diabetes disease data set, we can see that when $p = 1$, the 2D descriptors feature space/view is the one to result in least error and hence a weight of 1 is assigned to that view while the rest are assigned a value 0 based on Equation (4.13). As the value of p is increased we can see the weight getting more uniformly distributed across the views. We see a similar behavior of weight distribution for the diabetes data set where the weights become uniform as p increases.

Tables 4.4 and 4.5 represents the average F1 scores of the five different methods and their error variance within parenthesis for the two disease data sets. The method that performed the best is highlighted in bold. As a first observation, we see that the data sets exhibit similar behavior in terms of how the methods perform relative to each other. If we consider Arrhythmias as an example, we see that the average F1 score of multi-view learning algorithms perform better than single-view

Method	Arrhythmias		
	p = 1	**p = 2**	**p = 5**
wMVL	$0.736 \pm 0.0.11$	$\mathbf{0.780 \pm 0.007}$	$\mathbf{0.746 \pm 0.019}$
MVL	$\mathbf{0.752 \pm 0.019}$	0.743 ± 0.005	0.735 ± 0.022
Ridge	0.744 ± 0.009	0.729 ± 0.005	0.724 ± 0.003
kNN	0.715 ± 0.011	0.720 ± 0.009	0.716 ± 0.006
RF	0.742 ± 0.002	0.740 ± 0.003	0.732 ± 0.002
	Heart Arrest		
	p = 1	**p = 2**	**p = 5**
wMVL	$\mathbf{0.678 \pm 0.0184}$	$\mathbf{0.717 \pm 0.0116}$	$\mathbf{0.696 \pm 0.0075}$
MVL	0.665 ± 0.0593	0.672 ± 0.0148	0.664 ± 0.0667
Ridge	0.548 ± 0.0132	0.571 ± 0.0068	0.586 ± 0.0741
kNN	0.654 ± 0.0127	0.664 ± 0.0109	0.653 ± 0.0503
RF	0.636 ± 0.0213	0.657 ± 0.1000	0.640 ± 0.0899
	Heart Failure		
	p = 1	**p = 2**	**p = 5**
wMVL	0.704 ± 0.0058	$\mathbf{0.781 \pm 0.0162}$	$\mathbf{0.746 \pm 0.0231}$
MVL	$\mathbf{0.712 \pm 0.0011}$	0.716 ± 0.0030	0.728 ± 0.0084
Ridge	0.685 ± 0.0185	0.676 ± 0.0083	0.663 ± 0.0250
kNN	0.682 ± 0.0053	0.692 ± 0.0024	0.675 ± 0.0081
RF	0.707 ± 0.0066	0.706 ± 0.0052	0.710 ± 0.0035

Table 4.4: Performance comparison of the average F1 scores of the five methods on the cardiovascular data set

learning methods which work on the CFS. Secondly, we see that when $p = 1$ MVL performs better than wMVL due to the fact that wMVL utilizes only one view. With an increase in the value of p, the F1 score improves with wMVL outperforming MVL. At $p = 5$, wMVL still performed better than all the other methods. In the case of Heart Arrest, we see that wMVL performs better than MVL even for $p = 1$ and we think this only by chance and instead of focusing merely on the F1 score, we think additional views add more insight to the underlying behavior of the drugs. For all the diseases, $p = 2$ gave the best results for wMVL. Since we did not perform the experiments on

the same set of observations for different p values, we have different values for the other methods too and this should not be confused with the other methods being influenced by p. The exponential parameter is used only for wMVL. We see similar observations and trends across the methods for the diabetes diseases.

Method	**Angiopathies**		
	p = 1	**p = 2**	**p = 5**
wMVL	0.587 ± 0.051	$\mathbf{0.617 \pm 0.057}$	$\mathbf{0.669 \pm 0.061}$
MVL	$\mathbf{0.592 \pm 0.012}$	0.610 ± 0.031	0.622 ± 0.020
Ridge	0.521 ± 0.058	0.501 ± 0.002	0.511 ± 0.034
kNN	0.563 ± 0.183	0.573 ± 0.054	0.565 ± 0.146
RF	0.554 ± 0.037	0.572 ± 0.058	0.580 ± 0.104
	Cardiomyopathies		
	p = 1	**p = 2**	**p = 5**
wMVL	0.552 ± 0.054	$\mathbf{0.603 \pm 0.022}$	$\mathbf{0.591 \pm 0.088}$
MVL	$\mathbf{0.567 \pm 0.094}$	0.583 ± 008	0.573 ± 0.039
Ridge	0.529 ± 0.015	0.534 ± 0.076	0.513 ± 0.053
kNN	0.546 ± 0.119	0.543 ± 0.077	0.553 ± 0.091
RF	0.527 ± 0.101	0.506 ± 0.044	0.502 ± 0.041
	Mellitus		
	p = 1	**p = 2**	**p = 5**
wMVL	0.620 ± 0.013	$\mathbf{0.647 \pm 0.053}$	$\mathbf{0.668 \pm 0.036}$
MVL	$\mathbf{0.651 \pm 0.052}$	0.642 ± 0.054	0.650 ± 0.029
Ridge	0.610 ± 0.025	0.633 ± 0.032	0.609 ± 0.031
kNN	0.607 ± 0.068	0.603 ± 0.040	0.634 ± 0.027
RF	0.624 ± 0.0304	0.656 ± 0.020	0.635 ± 0.051

Table 4.5: Performance comparison of the average F1 scores of the five methods on the diabetes data set

4.5.2 Statistical Significance

In addition to the F1 scores, we tested for statistical significance to emphasize that our proposed method is significantly better than the other methods. Table 4.6 shows the p-values obtained from the Wilcoxon paired test at the 95% confidence level for both the data sets. Due to space constraint, we are only presenting the results for the test based on the performance when $p = 2$ and we randomly pick one disease from each data set although all the disease had similar results. We compared wMVL with the rest of the four methods. Our null Hypothesis H_0 is that the two algorithms are same and our alternate hypothesis H_a argues otherwise. We reject the null hypothesis if the p-value ≤ 0.05. From the results, we can see that wMVL significantly outperforms the other methods compared in this study.

Cardiovascular			**Diabetes**		
Method 1	**Method 2**	**p-value**	**Method 1**	**Method 2**	**p-value**
wMVL	MVL	0.023	wMVL	MVL	0.048
wMVL	Ridge	0.007	wMVL	Ridge	0.039
wMVL	kNN	0.026	wMVL	kNN	0.031
wMVL	RF	0.015	wMVL	RF	0.023

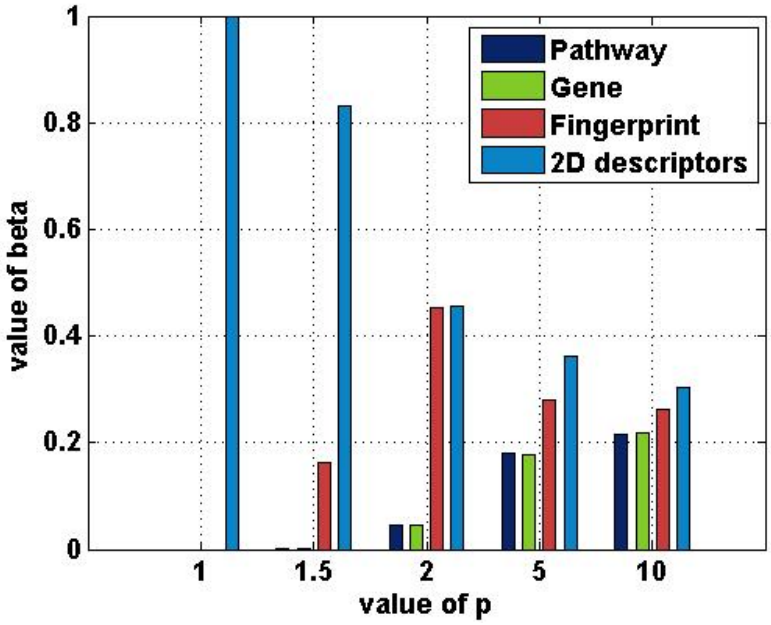
Table 4.6: Wilcoxon ranked test of wMVL with the other methods for the two disease data sets

4.6 Conclusion and Future Work

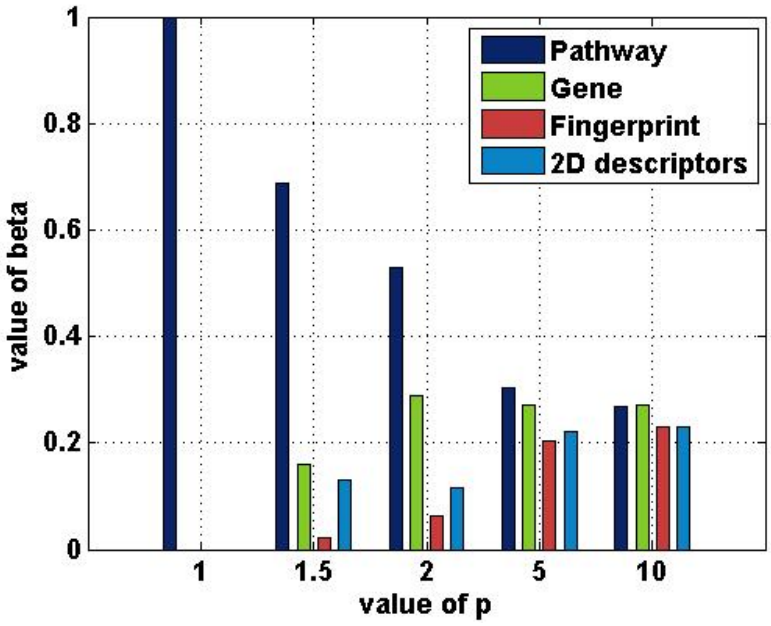
In this paper, we proposed a method called weighted multi-view learning that is similar to a co-regularized multi-view learning framework but weights each view differently. The primary motivation for this work was the need for a framework in cheminformatics that could handle data from multiple sources but at the same time weigh useful data more than noisy data. When there is an increase in the number of features due to additional information, the complexity of the model can drastically increase thereby decreasing the learning model with low sample size. MVL addresses this issue effectively by handling each feature space individually by keeping a relatively low model

complexity but learns useful information from other views. From our experimental results on both synthetic and real life data, we show that wMVL performs better than the co-regularized MVL and single view learning methods. The test for statistical significance further emphasizes that wMVL has a performance capability significantly better than the other traditionally used methods. The algorithm's capability to learn the view weights without the need for prior information makes it all the more attractive and easier to use.

As part of our ongoing work, we are looking to extend the wMVL learning framework to handle multiple tasks since we know that multi-task learning enhances the overall performance of all the tasks when related tasks are learned jointly. In our case we would extend this framework to learn diseases that are related to each other in terms of their properties or their drug activities. We believe that in learning multiple tasks, we overcome the small sample size problem and also help in better learning the drug-disease associations.



(a) Diabetes



(b) Cardiovascular

Figure 4.4: The distribution of the weights across the views controlled by the exponential parameter p

Chapter 5

Multi-task with Weighted Multi-view Learning for Predicting Chemical-Target Interactions

5.1 Introduction

The study of interaction between small molecules and biomolecules, especially proteins has gained importance in the field of design since it forms a crucial step in drug development [3][8]. There is a fast accumulation of protein-chemical interaction data in the public domain such as PubChem, ChEMBL and Protein Data Bank. It was estimated that only 1% of chemical information is stored in public domain [32]. This scenario has changed over the last few years where the accumulation of digitalized data on chemical structures, interaction of chemicals and proteins and chemical genomics has seen an exponential growth. As of October 2016, PubChem has around 1.2 million bioassays, 100 million compounds and 300 million substances. ChEMBL which contains chemical compounds and their bioactivities has around 11 thousand targets, 2 million compounds and 14 million activities. Computational tools for studying PCI has never been a way to replace physical experiments but a way to support them in drug discovery. With the availability of databases con-

taining a large amount of structures and activity information about compounds, researchers have the advantage of exploiting these huge data to develop new methods to effectively predict PCI. The use of machine learning techniques has been drawing great interest in predicting such interactions.

CPI has two important characteristics that make it challenging for the use of machine learning techniques. Firstly, the distribution of known protein interaction data is highly skewed. Secondly, CPIs are very sparse. Bioassays have very less active compounds out of the hundreds of thousands of compounds screened. Establishing a chemical-protein interaction pair has become a key challenge in drug discovery. Since physical experiments have known to be extremely time consuming, ML techniques have been utilized to develop computational methods for CPI prediction.

5.2 Related Work

The research about chemical-target interactions has been around for many years now [3][8], yet the cheminformatics community is constantly on the lookout for tools that help in speeding up the process. Due to the fact that the initial drug-screening process is time consuming phase and very pricey [**?**], researchers brought in the idea of utilizing computational tools that helped in narrowing down the search for potential drugs from hundreds of thousands of molecules to a few thousand molecules (References). This exploitation of computational methods has shown to bring about an improvement in drug design in terms of time and efficiency in identifying potential small molecules and indirectly also reducing the costs involved. Identifying chemical-target interactions are vital in terms of: 1) finding direct relationships between a drug and a target [12]. 2) identifying side effects or adverse drug reactions [139]. 3) establishing a relationship between drugs and diseases [20]. 4) Re-purposing an approved drug to modify a different target which in turn might be therapeutic for a newer disease [24].

5.2.1 Data Integration

Additional data or extra information has always been seen as a boon while trying to gain knowledge. With an explosion of data in the public domain, it is desirable to utilize them to improve the quality of data used in developing prediction models. Most of the databases for computational biology specialize in certain aspects. For example while one database might focus on the side-effect profiles of drugs [76], another database might contain data about their pathway profiles [68]. Utilizing both these databases would prove to be advantageous. A classic way to integrate data from multiple databases or sources is to combine them into a single feature space [23][128]. Although this has been a typical way to fuse heterogeneous data, it might not be a good idea due to the fact that the integration might lead to a very high dimensional data while the sample size remains the same. The other key drawback of such a practice is that, each information space might have an underlying statistical property that might not be preserved while integrating them together. On the other hand, databases can also "teach" or "learn" from each other by acting as supplementary information. The solution for the above stated problems has been addressed by multi-view learning where the views learn from each other [95] on a set of unlabeled samples. In a co-regularized setting, each view has its own model and the views try to minimize their disagreement about the label on a set of unlabeled samples.

5.2.2 Learning Related Targets

Multi-task learning (MTL) has been researched for a long time now for various applications [21][10][53] including computational biology [139][11]. The fundamental assumption of MTL is that when similar tasks are jointly learned together, their overall performance is better than learning each individual task separately [18]. This concept has been used in cheminformatics where similar targets and their drug interactions have been studied [96]. Xing et al., proposed a kernel based MTL approach to model target similarities to improve SAR models. Jintao et al., [139] studied MTL for chemical-protein interaction using a task regularized framework. Similar studies include using MTL for HIV screening [11], prioritizing disease genes [93] and discovering targets

for Alzheimer's [47]. On the other hand, the multi-target scenario where a drug interacts with multiple targets has been handled as a multi-label problem where a classifier is used to train in a "one-versus-rest" binary setting [24]. The other main advantage of MTL is that, the skewed nature of class samples is alleviated by learning from samples of similar tasks.

5.2.3 Multi-task Multi-view Learning

Combining the advantages of both multi-task as well as multi-view learning has not been explored to its full potential in the field of cheminformatics, however, preliminary studies show a promise in exploiting this framework [20][139]. Zhang et al., used the multi-task multi-view framework to predict adverse drug reactions (ADRs). A quantitative relationship between drug structures, drug-protein profiles and drug-ADRs (multi-view) was studied by learning multiple ADRs together (multi-task). Similarly Chandrasekaran et al., performed an empirical study using the MTMVL to predict drug-disease associations by learning related diseases as well as combining data from varied sources results and have shown the effectiveness of this framework. But one underlying assumption in this has been the assumption that each source of data contributes equally to the learning process. This assumption might sometimes be too strong. In order to mitigate this problem, weighted multi-view learning (wMVL) methods have been proposed where each view is weighted based on its prediction power. Initial weighted multi-view learning concentrated on clustering methods that helped in learning the view weights as well as the model parameters without any prior knowledge about the views. Chandrasekaran et al., further proposed a weighted multi-view framework for classification and demonstrated its effectiveness in predicting drug-disease associations. In this paper, we propose a method called Multi-task with weighted Multi-view Learning (MTwMVL) framework that leverages the advantages of wMVL as well as multi-task learning. The proposed method is effective in the following ways:

- The multi-view learning part takes care of the high dimensional data that is accrued due to the fusion of multiple sources.

- Weighting the views based on their prediction power further improves MVL.
- The multi-task component of the framework addresses the problem of low sample size by learning related tasks jointly. The skewed nature of cheminformatics data is a major challenge utilizing machine learning tools and MTL overcomes this hurdle.

5.3 Method

In this section, we propose a multi-task with weighted multi-view learning (MTwMVL) framework that jointly learns similar related tasks and learns from multiple views in a weighted way where the learning framework automatically learns the weights of each view for each task. The final prediction for a given task is the weighted sum of the predictions of each view as opposed to the average prediction across all the views.

5.3.1 Notations

In this paper, we used bold uppercase letters to represent a matrix (e.g $\mathbf{X}$) and bold lowercase letters to represent a vector (e.g. $\mathbf{x}$). Greek letters are used to represent regularization parameters (e.g λ), simple lower and uppercase letters are used to represent scalars (e.g x, X). We use the subscript v to denote a view. For example, if we have V views, $\mathbf{X}_v$ represents data $\mathbf{X}$ from view v where $v \in V$. All the vectors are column vectors unless specified.

5.3.2 Overview of Learning Framework

Let n be the number of labeled samples in view v and task t, denoted by $\mathbf{X}_t^v \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \{-1,+1\}^{n \times 1}$ be their corresponding labels. We represent the unlabeled data of task t and view v as $\mathbf{U}_t^v \in \mathbb{R}^{n \times d}$. For a given task t, the prediction on a sample $\mathbf{x}_t^t$ is given by the average prediction of all the views as given by Equation 5.1.

$$f_t(\mathbf{X}_t) = \frac{1}{V}\sum_{v=1}^{V} f^v(x^v) \tag{5.1}$$

For our proposed method, we define the prediction for task t as the weighted average of all the views as given by Equation 5.2.

$$f_t(\mathbf{X}_t) = \sum_{v=1}^{V} \beta_t^v f^v(x^v) = \beta_t^v \mathbf{X}_t^v \mathbf{w}_t^v \tag{5.2}$$

The objective function of multi-task multi-view algorithm proposed by Zhang et al., is given by Equation 5.3. The first term is the weighted least squares error of the models on each task and view and the second term regularizes the model coefficients for each view of a given task. For a particular task, the third term minimizes the view disagreement between two different views controlled by the parameter μ. Minimizing this term enforces the views to agree with each other on the unlabeled samples as much as possible. Similarly, the fourth term minimizes the disagreement between two tasks for a given view.

$$\begin{aligned}\min_{w_v} \ & \frac{1}{2}\sum_{t=1}^{T}\sum_{v=1}^{V}\left\|\mathbf{y}_t - \frac{\mathbf{X}_t^v \mathbf{w}_t^v}{V}\right\|^2 + \frac{\lambda_1}{2}\sum_{v=1}^{V}\|\mathbf{w}_t^v\|^2 + \\ & \frac{\mu}{2}\sum_{v\neq v'}^{V}\left\|\mathbf{U}_t^v \mathbf{w}_t^v - \mathbf{U}_t^{v'} \mathbf{w}_t^{v'}\right\|^2 + \frac{\gamma}{2}\sum_{t\neq t'}^{T}\left\|\mathbf{w}_t^v - \mathbf{w}_{t'}^v\right\|^2\end{aligned} \tag{5.3}$$

We modify the above objective function to accommodate the weighting of the views by introducing β_t^v. For a view that gives low error, a higher weight is assigned and for a view with high error, a lower weight is assigned to the view. We further impose the following constraints: for a given task, $\sum_{v=1}^{V} \beta_t^v = 1$ and $\beta_t^v >= 0$. Similarly we also enforce that for a given task, the view weights should be similar as given by the fifth term in Equation 5.4. We aim to jointly minimize Equation 5.4 by alternately optimizing the parameters $\mathbf{w}_t^v$ and β_t^v.

$$\min_{w_t^v,\beta_t^v} \frac{1}{2}\sum_{t=1}^{T}\sum_{v=1}^{V}\beta_t^{v^2}\|\mathbf{y}_t-\mathbf{X}_t^v\mathbf{w}_t^v\|^2+\frac{\lambda_1}{2}\sum_{v=1}^{V}\|\mathbf{w}_t^v\|^2+$$
$$\frac{\mu}{2}\sum_{v\neq v'}^{V}\left\|\mathbf{U}_t^v\mathbf{w}_t^v-\mathbf{U}_t^{v'}\mathbf{w}_t^{v'}\right\|^2+\frac{\gamma}{2}\sum_{t\neq t'}^{T}\left\|\mathbf{w}_t^v-\mathbf{w}_{t'}^v\right\|^2+ \tag{5.4}$$
$$\frac{\lambda_3}{2}\sum_{t\neq t'}^{T}\left\|\beta_t^v-\beta_{t'}^v\right\|^2+\lambda_2(\sum_{v=1}^{V}\beta_t^v-1)$$

We differentiate the above equation with respect to each $\mathbf{w}_t^v$ and equate it to zero.

$$\frac{\partial F}{\partial \mathbf{w}_t^v}=\beta_t^{v^2}\mathbf{X}_t^{vT}(\mathbf{X}_t^v\mathbf{w}_t^v-\mathbf{y}_t)+\lambda_1\mathbf{w}_t^v+\mu(V-1)\mathbf{U}_t^{vT}\mathbf{U}_t^v\mathbf{w}_t^v$$
$$-\mu\mathbf{U}_t^{vT}\sum_{v'\neq v}\mathbf{U}_t^{v'}\mathbf{w}_t^{v'}+\gamma(T-1)\mathbf{w}_t^v-\gamma\sum_{t'\neq t}\mathbf{w}_{t'}^v \tag{5.5}$$

By setting Equation 5.5 to 0 and rearranging the terms, we get the following equations.

$$\boxed{A_t^v\mathbf{w}_t^v+\sum_{v'\neq v}B_t^{v'}\mathbf{w}_t^{v'}+\sum_{t'\neq t}C_{t'}^v\mathbf{w}_{t'}^v=D}$$

$$A_t^v=\lambda_1+\gamma(T-1)+\mu(V-1)\mathbf{U}_t^{vT}\mathbf{U}_t^v+\beta_t^{v^2}\mathbf{X}_t^{vT}\mathbf{X}_t^v$$
$$B_t^{v'}=-\mu\mathbf{U}_t^{vT}\mathbf{U}_t^{v'} \tag{5.6}$$
$$C_{t'}^v=-\gamma I$$
$$D=\beta_t^{v^2}\mathbf{X}_t^{vT}\mathbf{y}_t$$

The above sets of equations solve for each $\mathbf{w}_t^v$ jointly by computing the sets of linear equations.

We similarly solve for the view weights by differentiating Equation 5.5 with respect to each β_t^v and setting it to zero.

$$\frac{\partial F}{\partial \beta_t^v}=(\mathbf{y}_t-\mathbf{X}_t^v\mathbf{w}_t^v)^2\beta_t^v+\lambda_3(T-1)\beta_t^v-\lambda_3\sum_{t'\neq t}\beta_{t'}^v+\lambda_2 \tag{5.7}$$

By setting Equation 5.7 to zero, we solve for β_t^v given by Equation 5.8.

$$\beta_t^v = \frac{\lambda_3 \sum_{t' \neq t} \beta_{t'}^v - \lambda_2}{(\mathbf{y}_t - \mathbf{X}_t^v \mathbf{w}_t^v)^2 + \lambda_3 (T-1)} \tag{5.8}$$

Substituting Equation 5.8 in the constraint, $\sum_{v=1}^{V} \beta_t^v = 1$, we get the following equation where $D_t^v = (\mathbf{y}_t - \mathbf{X}_t^v \mathbf{w}_t^v)^2$.

$$\sum_{a=1}^{V} \frac{\lambda_3 \sum_{t' \neq t} \beta_{t'}^a - \lambda_2}{D_t^a + \lambda_3 (T-1)} = 1 \tag{5.9}$$

$$(\lambda_3 \sum_{t' \neq t} \beta_{t'}^v - \lambda_2) \left(\sum_{a=1}^{V} \frac{1}{(\mathbf{y}_t - \mathbf{X}_t^a \mathbf{w}_t^a)^2 - \lambda_3 (T-1)} \right) = 1 \tag{5.10}$$

$$(\lambda_3 \sum_{t' \neq t} \beta_{t'}^v - \lambda_2) = \frac{1}{\left(\sum_{a=1}^{V} \frac{1}{(\mathbf{y}_t - \mathbf{X}_t^a \mathbf{w}_t^a)^2 - \lambda_3 (T-1)} \right)} \tag{5.11}$$

Substituting Equation 5.11 into Equation 5.9, we get an update for β_t^v.

$$\boxed{\beta_t^v = \frac{1}{\sum_{a=1}^{V} \left(\frac{(\mathbf{y}_t - \mathbf{X}_t^v \mathbf{w}_t^v)^2 - \lambda_3 (T-1)}{(\mathbf{y}_t - \mathbf{X}_t^a \mathbf{w}_t^a)^2 - \lambda_3 (T-1)} \right)}} \tag{5.12}$$

Equations 5.6 and 5.12 give the update rules for the model co-efficients and the view weights respectively. Initially, all the views are weighted equally (1/V) and the weights are automatically learned using the update rule. The advantage of this method is that it eliminates the need for a prior knowledge of the view weights.

Algorithm 2 Multi-task with Weighted Multi-view Algorithm

Input: $\mathbf{y}_t$, $\{\mathbf{X}_t^v\}_{v=1,t=1}^{V,T}$, $\{\mathbf{U}_t^v\}_{v=1,t=1}^{V,T}$ λ_1, μ, N_{it}, ε
Output: $\{\mathbf{w}_t^v\}_{v=1,t=1}^{V,T}$, $\{\beta_{t0}^{v0}\}_{v=1,t=1}^{V,T}$
Initialize $w_{t0}^{v0} = 0$ and $\beta_t^v = \frac{1}{V}$ for $t \in [1:T]$ and $v \in [1:V]$
for $iter = 1$ to N_{it} **do**
    for $t = 1$ to T **do**
        for $v = 1$ to V **do**
            Compute Ft as given by Eqs.(7)
            Compute St for every $v' \neq v$ as given by Eqs.(7)
        end for
        Compute $\mathbf{w}_t^v := Ft^{-1}St$ for each $v \in [1:V]$
        Update β_t^v using Eqs.(12) or (13) for each $v \in [1:V]$
        $\left\|\mathbf{w}_t^v - \mathbf{w}_{t0}^{v0}\right\| < \varepsilon$ & $\left\|\beta_t^v - \beta_{t0}^{v0}\right\| < \varepsilon$
        break
        $\mathbf{w}_{t0}^{v0} := \mathbf{w}_t^v$ for each $v \in [1:V]$ and $t \in [1:T]$
        $\beta_{t0}^{v0} := \beta_t^v$ for each $v \in [1:V]$ and $t \in [1:T]$
    end for
end for
Return $\mathbf{w}_t^v$ and β_t^v

5.4 Experimental Study

5.4.1 Data Sets

5.4.1.1 Data Preprocessing

Data was collected from ChEMBL database [51] that contains manually curated information regarding bioactive molecules that have the potential to be drugs. We picked two subfamilies from the GPCR family namely Dopamines and Histamines. For the Dopamines we chose D_1, D_2 and D_3 receptors and for Histamines we chose H_1, H_2 and H_3 receptors. We first search for the target (e.g. Dopamine D_1) and download their bioactivities. The files contained information of the chemicals that were tested against the D_1 receptor. As a preprocessing phase, we then adopted the following criteria to filter the raw data to obtain the final dataset.

1. Records with a value of 1 for the pot/duplicate field were removed.

2. Only records from the B,A and F assays were retained.

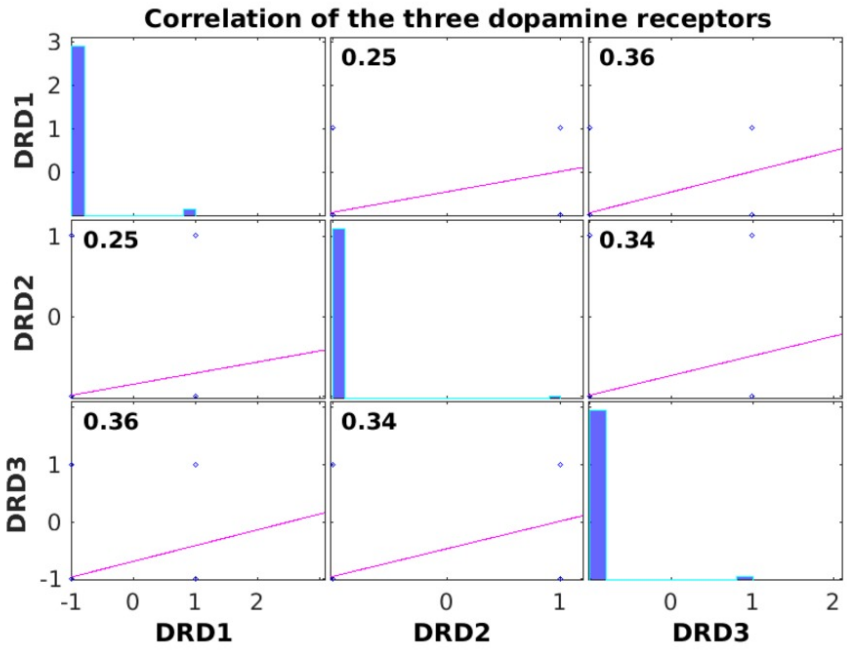


Figure 5.1: Label correlation of the three dopamine receptors - DRD1, DRD2 and DRD3

3. Records with a molecular weight < 150 and > 900 were filtered out.

4. chemicals that did not have a confidence score of 9 were removed further.

5. All the chemicals with an absence of pKi value were filtered out.

6. Some chemicals had multiple activities reported. Samples that had a PChEMBL value that differed by more than 1 were excluded and the average values of the remaining records was used as the PChEMBL value for that particular chemical.

7. Chemicals with a pKi value <=5 were counted as active interactions (label +1) and a pKi > 5 was accounted as an absence of interaction (label -1) with the target of interest.

Figures 5.1 and 5.2 are the correlation plots of dopamine and histamine receptors respectively. The correlation was calculated for the proteins within their families. Just for the correlation study, we filtered out the chemicals (if any) common to all three receptors (eg: DRD1, DRD2, DRD3) and calculated the correlation on their labels. The dopamine receptors had 201 chemicals common

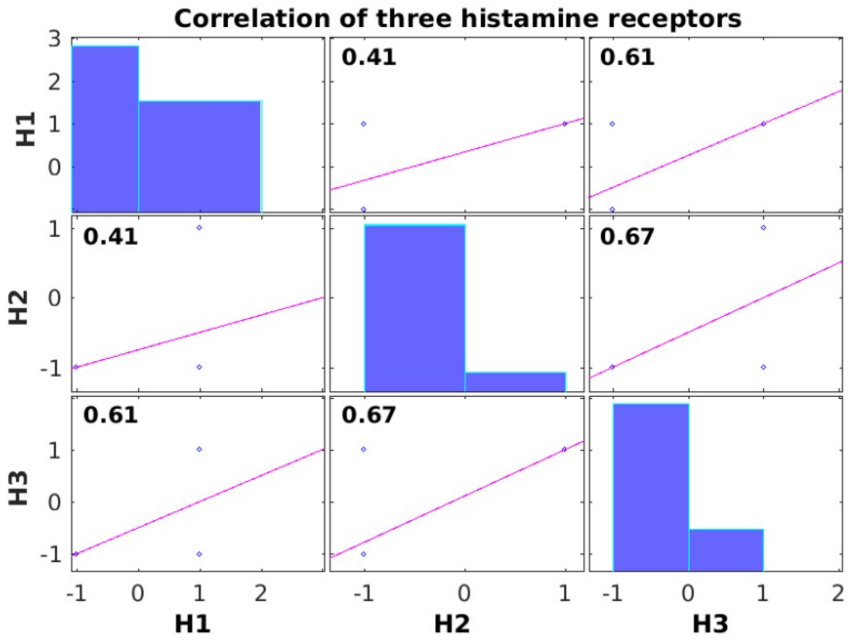


Figure 5.2: Label correlation of the three histamine receptors - H1, H2 and H3

between DRD1, DRD2 and DRD3. Similarly the histamine receptors had just 10 chemicals common between H1, H2 and H3.

5.4.2 Feature/View Construction

We constructed three views for each compound in the data set. Alexious et al., [74] showed the potential of the molecular descriptor space capturing different properties of a compound. In reference to that work, we considered three types of molecular descriptors - 2D structural features, Extended Circular Fingerprints (ECFP) and pharmacophore based descriptor GpiDAPH3 and hence each target had three views. The 2D descriptors had 192 features, the ECFP descriptors had 1024 features and the GpiDAPH3 descriptors had varying number of features for each target in the subfamilies. We modified the third pharmacophore view by standardizing across targets within each subfamily. If two or more features had greater than 95% correlation, we represented the entire group with a randomly picked feature from the group. The 2D descriptors and the pharmocophore descriptors were calculated using MOE software [116]. For e.g. if D_1 had a1 and a2 pharmocophore features,

Table 5.1: The table represents the total number of chemicals for each target and the number of active and inactive interactions.

Target	Number of chemicals	Number of Actives	Number of Inactives
DRD1	505	16	489
DRD2	2514	53	2461
DRD3	1715	41	1674
H1	424	15	409
H2	95	8	87
H3	2239	18	2221

Table 5.2: The number of features of the three views of the two GPCR families

Family	Number of Features		
	2D	ECFP	GpiDAPH3
Dopamines	192	1024	6128
Histamines	192	1024	4879

D_2 had a3 and D_3 had a1, we constructed a binary matrix with a1,a2 and a3 as the three features for view 3 of Dopamines and marked a +1 for compounds with a particular feature and 0 otherwise.

5.4.3 Model Construction and Evaluation

We choose N samples at random from each task by making sure we select equal samples from both classes since we are not studying the effects of class imbalance. In addition to the N samples, we additionally choose random samples to form the unlabeled samples for MVL. The size of the

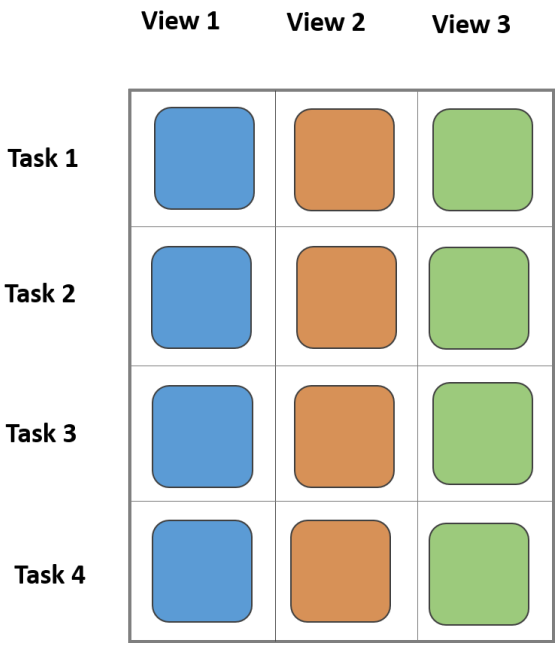


Figure 5.3: A graphical representation of multiple tasks and multiple views. Each view has N samples and each view has d_1, d_2 and d_3 number of total features respectively. A sample from view 1 and task 2 is represented as $\mathbf{x}_2^1$

unlabeled set is generally $N*3$. We subject the data to a five fold cross validation to obtain the training and test set. We further perform a five fold cross validation on the training set to get the training and validation sets. We use the training and validation sets for a grid search to obtain the optimal hyperparameters for each algorithm. This entire process is repeated 10 times and the performance of the classifiers are reported as the average of all the 50 models.

The F1 score is used to measure the performance of all the models. P and R denote the precision and recall values and tp, tn, fp and fn represent true positive, true negative, false positive and false negative respectively.

$$P = \frac{tp}{tp+fp} \tag{5.13}$$

$$R = \frac{tp}{tp+fn} \tag{5.14}$$

$$F1 = \frac{2*P*R}{P+R} \tag{5.15}$$

Table 5.3: Summary of the model parameters for each learning method

Method	**Hyperparameters**
MTL	λ_1, γ
MVL	λ_1, μ
MVMTL	λ_1, γ, μ
wMVL	λ_1, λ_2, γ, μ
MTwMVL	λ_1, λ_2, λ_3,γ, μ
Ridge	λ_1

5.4.4 Performance Comparison

We compared the performance of the proposed MTwMVL method with the following methods.

1. By setting $\mu = 0$ and without introducing β, the objective function reduces to a regularized **multi-task, single-view learning**.

$$\min_{w_t} \ \frac{1}{2}\sum_{t=1}^{T} \|\mathbf{y}_t - \mathbf{X}_t\mathbf{w}_t\|^2 + \frac{\lambda_1}{2}\|\mathbf{w}_t\|^2 + \frac{\gamma}{2}\sum_{t\neq t'}^{T} \|\mathbf{w}_t - \mathbf{w}_{t'}\|^2 + \tag{5.16}$$

2. By setting $\gamma = 0$ and by not introducing β, the objective function reduces to a co-regularized **multi-view, single-task learning**.

$$\min_{w^v,} \ \frac{1}{2}\sum_{v=1}^{V} \left\|\mathbf{y} - \frac{\mathbf{X}^v\mathbf{w}^v}{V}\right\|^2 + \frac{\lambda_1}{2}\sum_{v=1}^{V} \|\mathbf{w}^v\|^2 + \frac{\mu}{2}\sum_{v\neq v'}^{V} \left\|\mathbf{U}^v\mathbf{w}^v - \mathbf{U}^{v'}\mathbf{w}^{v'}\right\|^2 \tag{5.17}$$

3. By introducing a weighting parameter β to the co=regularized MVL, we get the objective function of weighted-Multi-view Learning (wMVL) which is again a **multi-view, single-task learning**.

$$\min_{w^v,\beta^v} \ \frac{1}{2}\sum_{v=1}^{V}\beta^{v^2}\left\|\mathbf{y}-\mathbf{X}^v\mathbf{w}^v\right\|^2+\frac{\lambda_1}{2}\sum_{v=1}^{V}\left\|\mathbf{w}^v\right\|^2+ \\ \frac{\mu}{2}\sum_{v\neq v'}^{V}\left\|\mathbf{U}^v\mathbf{w}^v-\mathbf{U}^{v'}\mathbf{w}^{v'}\right\|^2+\lambda_2(\sum_{v=1}^{V}\beta^v-1) \tag{5.18}$$

4. By not introducing β and setting λ_2 and λ_3 to 0, we get the **multi-task multi-view learning** where all the views are weighted equally.

$$\min_{w_t^v} \ \frac{1}{2}\sum_{t=1}^{T}\sum_{v=1}^{V}\left\|\mathbf{y}_t-\frac{\mathbf{X}_t^t\mathbf{w}_t^v}{V}\right\|^2+\frac{\lambda_1}{2}\sum_{v=1}^{V}\left\|\mathbf{w}_t^v\right\|^2+ \\ \frac{\mu}{2}\sum_{v\neq v'}^{V}\left\|\mathbf{U}_t^v\mathbf{w}_t^v-\mathbf{U}_t^{v'}\mathbf{w}_t^{v'}\right\|^2+\frac{\gamma}{2}\sum_{t\neq t'}^{T}\left\|\mathbf{w}_t^v-\mathbf{w}_{t'}^v\right\|^2 \tag{5.19}$$

5. By setting γ and μ to 0 and by not introducing β, the function reduces to ridge regression that represents **single-task, single-view learning**. Each task/target is subjected to the following objective function.

$$\min_{w} \ \frac{1}{2}\left\|\mathbf{y}-\mathbf{X}\mathbf{w}\right\|^2+\frac{\lambda_1}{2}\left\|\mathbf{w}\right\|^2 \tag{5.20}$$

5.5 Result Discussion

In this section, we test the proposed method with two datasets from ChEMBL and compare its performance with the other algorithms explained in our experiment section. For all the algorithms, we calculate the F1 score as a measure of the classifiers' performance. For both the Dopamine and Histamine data, the number of tasks T is 3 and the number of views V is 3. Tables 5.4 and 5.5 show the F1 scores (and the variance) of our experiments. The rows of the tables represent the learning methods and the columns represent the targets. For the single-task single-view algorithm,

we chose ridge regression since the base classifier is similar to the proposed method and it is only fair to compare learning methods that are similar. The method with the highest F1 score with respect to each target is represented in bold.

We first take the dopamine data as an example to discuss the results given in Table 5.4. As a first remark, we see that Ridge regression has the least F1 score and MTwMVL has the highest F1 score. This can be attributed to the fact that Ridge puts together all the three feature spaces and considers them as a single view. With low sample size and high dimensionality, the performance of the classifier could be affected. Secondly, we observe that by adopting a multi-view framework, the F1 score is considerably higher. Since the framework considers each feature space as a separate view and the views learn from each other, the performance of the classifier is better than Ridge. The performance of the MVL is further improved by wMVL which weights the three views based on their prediction power. Instead of weighting the three views equally, wMVL learns a weight for each view by assigning highest weight for the view with least error and least weight for the view with the highest error. Both the methods however still learn each of the targets separately.

Table 5.4: The average F1 score of the five learning methods on the Dopamine data set

Methods	**Dopamines**		
	D_1	D_1	D_3
Ridge	0.602±0.048	0.592±0.089	0.376±0.024
MVL	0.715±0.041	0.683±0.053	0.476±0.060
wMVL	0.735±0.021	0.716±0.039	0.502±0.067
MTL	0.769±0.053	0.752±0.025	0.747±0.049
MTMVL	0.770±0.012	0.783±0.003	0.762±0.038
MTwMVL	**0.823±0.022**	**0.830±0.015**	**0.787±0.027**

Observing the F1 score of MTL, we see that the F1 score is better than the MVL learning methods. This is because, as shown in Table 5.2, the number of chemicals having a true positive

Table 5.5: The average F1 score of the five learning methods on the Histamine data set

Methods	**Histamines**		
	H_1	H_2	H_3
Ridge	0.616±0.056	0.629±0.002	0.510±0.005
MVL	0.693±0.043	0.675±0.010	0.581±0.027
wMVL	0.704±0.053	0.714±0.024	0.600±0.016
wMTL	0.748±0.029	0.732±0.035	0.655±0.042
MTMVL	0.776±0.022	0.768±0.051	0.688±0.37
MTwMVL	**0.817±0.023**	**0.794±0.017**	**0.728±0.026**

interaction with the targets is very low. Since the main advantage of MTL is to learn similar tasks together, the tasks utilize the positive interactions from the other two dopamine receptors too. This improves the performance of the classifier as compared to Ridge. The increase in the F1 scores of all the three dopamine receptors between Ridge and MTL proves that multi-task learning is a promising method when individual tasks have very low positive or negative (or both) sample sizes. By combining the advantages of MTL and MVL, the MTMVL performs better than if both of them were applied on the data separately. In spite of learning from related tasks, the number of features can still be very high when all the features are combined into a single view. As a result, we see that MTMVL performs better than Ridge, MVL and MTL methods. Lastly, the proposed method that extends the MTMVL into a weighted framework improves the performance of the model further by combing wMVL and MTL.

Similar to the dopamine data set, the results for the histamine data exhibit similar trends in the performance of the six different learning methods. The Ridge classifier has the least F1 score and MTwMVL has the highest F1 score, thus showing the efficiency of our method over other relevant algorithms. In all of our experiments, if a particular iteration gave an NaN as the F1 score, we ran an extra iteration to replace the NaN values. Similarly when the experiments were performed during each fold, the random samples that were chosen were the same across all the six learning

methods.

5.6 Conclusion

In this paper, we proposed a multi-task with weighted multi-view framework to predict the interactions of chemicals and targets. The motivation behind weighting the views comes from the fact that different views can have varied predictive power. The second motivation is that, as more descriptors/information are used to represent the compounds, the dimensions can drastically increase for relatively the same number of samples. In order to mitigate this problem, multi-view learning partitions each feature space as a view. On the other hand, multi-task learning addresses the problem of low sample size by jointly learning similar tasks. The proposed method proves to be a unified framework that addresses both the problems stated above. Our systematic comparison of the performances of the proposed method and relevant learning algorithms show the promise of MTwMVL. As a part of our future work, we would extend the current method to handle heterogeneous tasks. Similar tasks would share parameters while dissimilar tasks would not have a common representation. Multi-task learning has already been used to tackle the problem of automatically learning task relationships. The other factor that would be incorporated is to handle missing values. Some feature spaces might have missing values and improvising the MTwMVL to handle missing data would be an advantageous improvement.

Chapter 6

Literature Survey on Truth discovery

6.1 Introduction

In the age of information burst, we have voluminous data available across different platforms. Analysing such data for building recommendation systems , decision making or to build predictive models has been going on for many years now. The challenge recently has been the inconsistency of facts/truths regarding a certain object across multiple platforms. For example, data is available across many mediums such as Twitter, Facebook, Instagram (or other social networks), forums (such as Quora or Stack exchange) and blogs. When information from different sources are aggregated to make a decision or build products, the truth needs to better detected. For example, a user on Facebook might have their location as "California" but have their location on Twitter as "Texas". Only one of these two could be true in most cases unless there is an exception. In the presence of a conflict, the performance of a model built on such data is pushed to be poorer which is not desirable. In order to overcome this discrepancy, finding the truth about the conflicting data has gained importance in the last few years. It is easy to discard conflicting data but sometimes useful information from that source might be lost. On the other hand, the most common and intuitive way to resolve this conflict is to take a majority voting or averaging. The main drawback of adopting such a method is that sources might copy from each other or there might be just one source with

the most recent update about the truth and is more valuable than the other sources and hence voting or averaging might disregard such data. Another drawback is that in a voting/averaging process, every source is given equal weightage. Due to this, the effect of quantity of the data overtakes the quality of the truth to be detected. To combat these issues, truth discovery or truth finding has gained rapid popularity in the aim to find fact(s) about objects [36][39][50][81][85][144][132]. Li et al., have done a very exhaustive literature survey on truth discovery methods addressing their various facets.

6.2 Facets of Truth Discovery

6.2.1 Common Strategies

When multiple sources provide different values for an object, the common strategy used to find the most likely value would be to take a majority voting or the average in the case of continuous information. As discussed earlier, if there are multiple sources claiming a false value, the majority voting would result in declaring that as the true value. The strength of truth discovery methods lie in the fact that they have the ability to output a minority value as the truth. The main principle that is followed is to estimate the reliability of the sources based on the data given. The reliability of a source is estimated based on the values for all the objects and properties it provides and each source is assigned a weight 'w_i', where $i = 1 : s$, the total number of sources.

There are many challenges associated with truth discovery which has given rise to a lot of different methods. Each method has addressed different challenges involved in truth discovery. Among the different aspects of truth discovery, a few of them are in terms of the input data, the appropriate assumptions about the output/truth, the possibility of a correlation or some sort of relationships between the sources. We summarize a few aspects here to show our understanding and perspective of the field.

6.2.2 Input Data and Preprocessing

The input data from different sources need preprocessing most of the times. If all the sources have the same value claimed about an object, it indicates a lack of conflict and so the true value is clearly established. These records are generally removed and it has also been shown that removing these records in fact improves the effectiveness of the methods [144]. One method however overlooks this and assumes that there might be other unknown outputs [102]. The input data across the sources are mostly in different formats. Hence it is essential to standardize them since the truth in order to compare the sources [84]. For example, some sources might list name in a "last name, first name" format and some other might include the middle name too. In terms of numeric features, one source might have it in different units compared to the rest.

Data collected from certain sources might contain several records for the same object. This is common in platforms that are open to public can alter the information. In these cases, the data with the latest time stamp is considered although this can come with its own problems. In the absence of a time stamp, a well deviced rule should be used [104]. Most of the work deals with static data but in some scenarios, data is dynamic and changes over time. This means that the model has to be recomputed each time which might not be feasible. Some methods address this and have developed methods for streaming data [85][118]. A new challenge in terms of extracting data has been to mine data from unstructured databases. Typically, structured data from relational databases have been a lot more easy to handle [50][36] but they seem to lack some extra information that unstructured data provide. [34][91].

Most of the methods address truth discovery based on the assumption that the ground truth is not known any of the objects. Although this might be true in most cases, there has been a branch of work where the authors consider that there is some amount ground truth that is known [133][88][38]. With the help of the labeled objects, truth discovery is approached in a semi-supervised setting. It has been shown that a few labels go a long way in helping learn the truth of

the unlabeled objects.

6.2.3 Estimating Source Reliability

A big portion of research in truth discovery is with respect to finding source reliability. The first line of thought is that sources are totally independent in collecting information about an object. This implies that the truth would be the same across the sources but the false data would be different depending on what basis and condition each source collected the information [50][82][133]. The second type of methods assume that the source presents the truth on all the objects with the same degree of reliability or with the same probability. Although this has been shown to be true in most applications, it might be a strong assumption for certain fields [132][144]. Unlike previous assumptions, some methods learn multiple reliability scores for each source. Gupta et al., demonstrated that the objects could first be clustered and a reliability score could be estimated for each object set [56]. As pointed out earlier, it might be too strict to assume that a source is reliable for all the attributes of an object. Yin et l., estimated reliability scores separately for each attribute thereby resulting in multiple reliability scores for a source [133].

A lot of methods in truth discovery assume that sources are not independent and have some level of dependency on each other [35][36][37][107]. When sources copy from each other, they are also bound to have the same false information regarding the objects. It is useful to detect this copying relationship between sources which would capture the common errors between them. This would however not be possible if sources copy from a correct source. Each source has its own information plus some copied information. Dong et al., used a Bayesian inference approach to directly detect the copying relationship [36] where an iterative method is used to update the copying relationship and the truth alternatively. In methods dealing with dynamic data, Dong et al., used a Hidden Markov Model to capture the copying relationship [37] by using snapshots of data and the output is an evolving value of the truths. Pochampally et al., added a different perspective to model the source dependency by studying the correlations among them [103]. The method models

correlations using joint precision and joint recall and the probability of an observation to be true is inferred using Bayesian analysis. In some exceptions like the work pointed out by Wang et al., the sources cite where they copied from [119]. For example, a news website could reference their information from a blog or another news portal.

A pitfall in estimating source reliabilities is the way in which they are initialized as a starting point. Most of them assign equal weights to all the sources as a common way to begin the estimation but this doesn't prove to be effective always. Assigning equal weights would result in the truth being updated as the majority of the values claimed by the different sources. This works when the majority of the values represent the truth. If majority of the sources contain false values, assigning equal view weights would not make sense. To overcome this, some researchers make use of a small set of labeled data [133]. Some others model source similarity as a prior knowledge so that the initial point of estimation is more accurate [80]. Unlike what was expected, studies show that more sources do not always and necessarily mean better capability to learn the truth and in fact some corrupt sources might degrade the performance [38]. Sarma et al., proposed a method to select a subset of sources based on cost constraints formulized as an optimization problem [107]. A different perspective to this whole argument was made by Li et al., who showed that a bad source also contributes towards truth discovery by setting negative weights to them. This in turn with a high probability infer what is wrong information. The truth about an object was in most cases considered independent in terms of its attributes but this need not be true always. For example, the date of birth and age could be related and need not be considered independent and this knowing relation could improve truth finding [102].

The relation between objects could also be a temporal or a spacial one. Another important challenge in truth discovery is the type of data present. Invariably the data falls into one of the two categories namely, categorical or numerical. Most of the methods are capable of handling only one type of data. Li et al., [82] proposed a unified framework which could work with heterogeneous

data to discover the truth. The authors use appropriate loss functions and formulized the truth of a object as the weighted sum of the objects from each source. The weights of the sources and the truth were alternatively optimized.

6.2.4 Assumptions

Having discussed about the input data and source reliability, we now turn our focus towards the assumptions and strategies adopted in the development of truth discovery in terms of their claimed values and output. Some methods assume that every object has only one true value and if the object votes for a value, then it opposes all the other values. This strategy of complimentary voting was demonstrated by [50][145]. This might not be true always and so some work propose the possibility of multiple truths [103][145]. For example, a course might be lectured by more than one professor or an album might have multiple singer. Under these assumptions, instead of just the source accuracy, precision and recall values were also used to detect the multiple truths. Zhi et al., put forth an interesting idea of including "unknown" as a common truth to all the objects apart from truth claimed by the object [146]. This helped scenarios in which there was no truth. It was impacted the optimization of finding the truth if the constraint was the relation between the sources. With the addition of this unknown truth, the sources has a relation if there was no truth common between them.

Yin et al., demonstrated the need for interpreting the claimed values [132] in a more relaxed setting. For example, if source 1 claimed a value of Mr. X's property as 3 million, source 2 claimed a value of 3.1 million and source 3 claimed a value of 8 million, then source 2 is considered trustworthy with a high probability given that source 1 is true whereas the value claimed by source 3 is false. Some methods output a label for each truth claimed by the sources. They output either a true or false label for each truth [144]. Some other methods output a score for each claimed value and the truth is deduced using post processing rules. Most or all of the methods use accuracy as a performance measure to arrive at the truth of an object. While accuracy is common for categorical

data, mean square error is widely used for continuous data.

6.2.5 Templates of Popular Frameworks

6.2.5.1 Notations

The common jargon and definitions used by most researchers in this field is as follows: We have different sources 's' that contribute some information or value 'v' about an object 'o'. The truth of an object is 'v^*'. Each source might be given a weight 'w' that indirectly tells us the reliability of a source. Hence an observation is an object 'o' from source 's' and that assigns a value 'v' as the information of the object. Each object can have more than one property for which the truth needs to be determined. For example, let us say that we want to determine the truth about a person's height, weight and home location, there are multiple sources that provide this information. We could for simplicity consider the DMV and hospital records as two sources 's_1' and 's_2'. The object here is the person and we have to determine the truth for three properties namely, height, weight and home location. The two sources provide a value for each of the property for the object. It has to also be noted that every source need not always provide a value for the object.

6.2.5.2 Methods

The central theme of truth discovery has been to discover truth from minorities. The three common approaches to all the methods in the literature can be divided into two main categories namely (i) probabilistic frameworks and (ii) iterative methods [82]. The first method is based on probabilistic graphical models (PGM) that uses a likelihood function of the general form given by Equation 6.1. Each value claimed by a source v_o^s is generated by the source weight w_s and corresponding truth v_o^*.

$$\prod_{s=1}^{S} p(w_s|\beta) \prod_{o=1}^{O} (p(v_o^*|\alpha) \prod_{s=1}^{S} p(v_0^s|v_0^*, w_s) \tag{6.1}$$

The iterative methods are straightforward where the source weights and the truth computation are alternatively optimized until convergence. While optimizing one of them, the other parameter is considered to be fixed. The truth is computed as a weighted inference like voting. Equation 6.2 represents a general form of the objective function used to estimate source weights and truth alternatively until convergence. $f(v_o^s, v_o^*)$ is the loss function that is chosen appropriately based on the input data.

$$\arg\min_{v_o^*, w_s} \sum_{o=1}^{O} \sum_{s=1}^{S} w_s . f(v_o^s, v_o^*) \tag{6.2}$$

Some of the methods that address various aspect(s) of truth discovery are already considered as benchmark methods in this area. The algorithms might be greatly diverse and hence comparing them on a single scale is not possible. We summarize here, some of the widely used methods. 2-estimates is a method that is based on the single truth assumption using complimentary voting and an extended version of the same called 3-estimates additionally models the difficulty of obtaining the truth of an object [50]. LTM [132] is a method based on PGM which detects multiple truths while TruthFinder is a Bayesian model that iteratively estimates source weights and truths [145].

6.3 Applications of Truth Discovery

People post reviews about medicines, physicians, books, restaurants and various other topics. These reviews however lack quality if we were to look at just one online portal. practically, the audience read multiple platforms to form an opinion on what might be true. In applications such as healthcare, it is often essential to know the truth since a wrong piece of information can mislead the audience. Discovering the truth helps both patients, doctors as well as the pharmaceutical industry [94]. Crowd sensing is a field which can be a high motivating application for truth discovery. With the popularity of social media, a lot of information is updated on social platforms. Unfortunately

most of them are outdated and act as mere noise. For example, when a natural disaster hits a particular geographical location, people post about the supplies needed and this information is copied by several sources. Unfortunately such data are not updated and keep floating around even much later. Learning from such highly noisy data can be very challenging and a few studies have addressed it [120][121]. Conflict also arises when different people accrue data from different platforms. There can be inconsistencies in labelling between two people and given a time frame, different people collect different sizes of data. In the end, truth discovery can be used to find the truth from a bunch of incomplete data [125].

6.4 Conclusion

In spite the numerous frameworks that have been put forth by various groups, there are still huge gaps that need to be addressed. Truth discovery in itself poses a lot of challenges like the ones we summarized here. Hence there is tremendous scope for improvement in terms of data, methodology and applications that might have specific challenges to name a few. All the work until now have addressed only a couple of aspects and it is desirable to build a framework that unifies as many factors as possible. Some applications cheminformatics have very specific challenges such as dealing with conflicting data within a single source. These conflicts are not time based where we could eliminate records based on their time stamp. They are results of physical experiments and due to external conditions or some errors, the end values differ widely. Although these records are often discarded, it is desirable to estimate the truth given that these experiments come at a cost and time. Extracting useful information from them would prove to be valuable.

Truth discovery has also been explored within a very limited scope in terms of translating it to a learning problem. Although it might be complicated, a framework on those lines would open new avenues in this field. In the next chapter, we propose a framework for truth discovery that loosely adapts from multi-view multi-task learning frameworks.

Chapter 7

A Semi-supervised Approach for Truth Discovery

7.1 Introduction

We are currently in the era of data explosion where there is so much data all around but we are not sure if they are all true. In many situations when we seek to look for a particular information, we are bombarded with facts from multiple sources. For example, if we search for the population of Kansas, we are able to get information from Wikipedia, the census bureau and from election campaigns. In most cases, all the three sources do not claim the same value for the question in place. This conflicting information serves as a motivation for truth discovery which aims at finding the truth about an object from heterogeneous sources.

An intuitive way of resolving such conflicts is to find the mean across the sources if the data is numerical or get a majority voting on them if the data is categorical. This straightforward method does not work very well when the number of false data is more. This method indirectly biases the truth towards the "majority is right" idea. In the real world, many sources copy from each other. During this process, if the false information is duplicated, our voting/averaging method would fail.

Hence it is utmost important to asses the quality of each source in terms of their reliability.

Several features of truth discovery have been explored with more research being done constantly to improve the estimation of source reliability. Here we present a semi-supervised setting to truth discovery where we assume that a small set of ground truths is known. The proposed work incorporates multiple data types like the CRH framework proposed by Li et al., [82]. We fundamentally model the problem to comprise of two different tasks (entries with known and unknwon ground truths) that learn the source reliabilities from each other. Our experimental study shows that the proposed method achieves better accuracy than similar methods.

7.2 Related Work

The central theme of truth discovery has been to estimate the reliability of sources that ultimately help in finding the truth as closely as possible. The setting and constraints under which these reliabilities are estimated are diverse and we summarize a few methods that closely relate to our problem setting. We divide the literature based on the availability of ground truth namely (i) unsupervised and (ii) semi-supervised approaches. In a setting where no ground truth is known, Li et al [82] proposed a unified framework called CRH that estimates the truth as a weighted sum of the sources by learning the weights and truths alternately without any prior knowledge on neither of them. The CRH framework also has the capacity to handle numeric as well as categorical data which none of the previous methods had. Other benchmark methods such as 2-Estimates and 3-Estimates brought in a direction to systematically extract truth in the presence of conflict. These methods however dealt only with categorical data. The 2-Estimates worked on an assumption that there can be only one truth for each entry.

A semi-supervised truth discovery framework proposed by Yin et al., demonstrated the effectiveness of having ground truth, even if small in number went a long way in more accurately finding

the truth from conflicting sources [133]. They closely adapt the semi-supervised graph learning approach that tries to predict the labeled samples in addition to the structure of the graph. All of the approaches focus only on one type of data except CRH which contributes mainly in handling heterogeneous data which has shown that exploiting both types of data is more effective in studying the reliability of the sources than estimating source reliability using just one type of data.

7.3 Notations

In this paper, we used bold uppercase letters to represent a matrix (e.g. $\mathbf{X}$) and bold lowercase letters to represent a vector (e.g. $\mathbf{x}$). Greek letters are used to represent Langrange multipliers (e.g. λ_1), simple lower and uppercase letters ar eused to represent scalars (e.g. x,X). We use the subscript s to denote a source. For example, if we have S sources, $\mathbf{X}_s$ represents data $\mathbf{X}$ from source s where $s \in S$. All the vectors are column vectors unless specified.

7.4 Methodology

In this section, we describe the framework of our proposed methods that aim at learning the truths from heterogeneous sources on multiple reported facts.

7.4.1 Problem Setting

We formally define the problem setting as follows. Every observation is called an object and each object can have one or more properties. For example, "John" is an object and "height" and "weight" are the properties of the object. The value for each property could be provided by one or more sources. The value of an object reported by different sources need not necessarily agree with each other which is the motivation for truth discovery. We represent the data from source 's' as $\mathbf{X}_s$ where $s \in S$. 'x_{ij}^s' denotes the j^{th} property of object 'i' from source 's'. For both our methods proposed, we assume that a small set of truths are known and the truths are known at random. So

we might know the true value for the j^{th} property of an i^{th} object from source s. We do not assume that we know all the properties of an object.

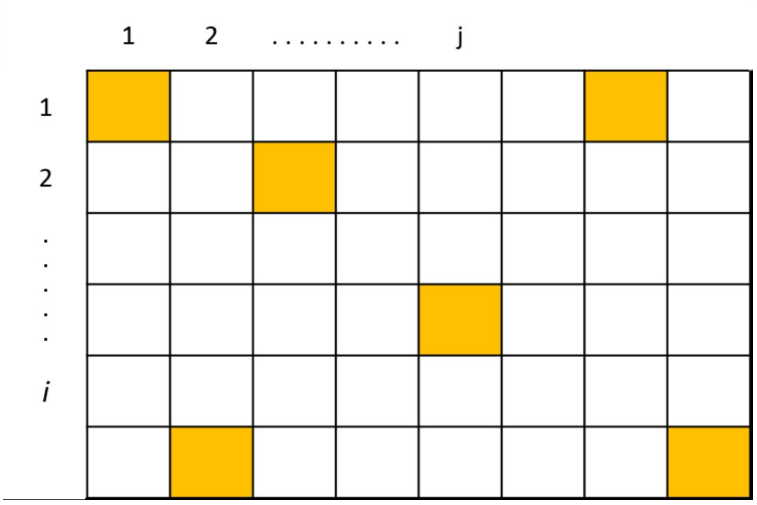


Figure 7.1: The figure represents our problem setting where the known ground truths are shaded in yellow. The rows represent the objects and the columns represent the properties.

7.4.2 Learning with Strong and Weak Truths

The first method we propose is called 'Truth Discovery using Strong and Weak Truths' - LSWT. We draw the inspiration for this method from the concept of multi-task learning in machine learning. The hypothesis is that learning closely related tasks together improves the overall performances of all the tasks. We adopt a similar idea of considering the objects with known ground truths and the objects with unknown ground truths as two different but related tasks. Instead of combining the two types of objects together and weighting the sources, we weight each "task" separately but add a constraint that the weights of the two tasks should be similar.

We first partition the data into objects with known ground truths X and objects without any known ground truth U. We use two different parameters β and α to weight X and U respectively. The corresponding truths of the two tasks are represented by X^* and U^*. While X^* are the ground truths, U^* are initialized with the mean value for continuous data and the with the majority voting for categorical data.

$$\min_{X^*,U^*,\beta_s,\alpha_s} f(X^*,U^*,\beta_s,\alpha_s) = \frac{1}{2}\sum_{s=1}^{S}\beta_s^2\sum_{n_1=1}^{N_1}\sum_{m=1}^{M}L(x_{nm}^*,x_{nm}^s)$$

$$+\frac{1}{2}\sum_{s=1}^{S}\alpha_s^2\sum_{n_2=1}^{N_2}\sum_{m=1}^{M}L(u_{nm}^*,u_{nm}^s)+\frac{\gamma}{2}\sum_{s=1}^{S}(\beta_s-\alpha_s)^2 \tag{7.1}$$

$$\text{s.t.}\sum_{s=1}^{S}\beta_s=1 \text{ and } \sum_{s=1}^{S}\alpha_s=1$$

In the above equation, $L(\cdot)$ defines the loss function and is based on the data type of the object's property. We incorporate the two constraints using Lagrange multipliers as given in Equation 7.2

$$\min_{X^*,U^*,\beta_s,\alpha_s} f(X^*,U^*,\beta_s,\alpha_s) = \frac{1}{2}\sum_{s=1}^{S}\beta_s^2\sum_{n_1=1}^{N_1}\sum_{m=1}^{M}L(x_{nm}^*,x_{nm}^s)$$

$$+\frac{1}{2}\sum_{s=1}^{S}\alpha_s^2\sum_{n_2=1}^{N_2}\sum_{m=1}^{M}L(u_{nm}^*,u_{nm}^s)+\frac{\gamma}{2}\sum_{s=1}^{S}(\beta_s-\alpha_s)^2 \tag{7.2}$$

$$+\lambda_1(\sum_{s=1}^{S}\beta_s-1)+\lambda_2(\sum_{s=1}^{S}\alpha_s-1)$$

7.4.3 Computing Source Weights

The optimization for this framework is performed by alternately updating the source weights and truths by solving Equation 7.2. With known ground truths X^*, we solve for β by taking partial derivative with respect to each β_s and setting it to zero.

$$\frac{\partial f}{\partial \beta_s} = \beta_s\sum_{n_1=1}^{N_1}\sum_{m=1}^{M}L(x_{nm}^*,x_{nm}^s)+\gamma(\beta_s-\alpha_s)+\lambda_1 \tag{7.3}$$

By setting Equation 7.3 to zero and rearranging the equation, we get:

$$\beta_s = \frac{\gamma\alpha_s-\lambda_1}{\sum_{n_1=1}^{N_1}\sum_{m=1}^{M}L(x_{nm}^*,x_{nm}^s)+\gamma} \tag{7.4}$$

By substituting Equation 7.4 into the constraint $\sum_{s'}\beta_{s'}=1$ we get the following equation,

$$\gamma\alpha_s - \lambda_1 = \sum_{s'=1}^{S} \frac{1}{\sum_{n_1=1}^{N_1}\sum_{m=1}^{M} L(x^*_{nm}, x^{s'}_{nm}) + \gamma} \tag{7.5}$$

Substituting Equation 7.5 in 7.4, we get an expression for β_s

$$\boxed{\beta_s = \frac{1}{\sum_{s'=1}^{S} \frac{\sum_{n_1=1}^{N_1}\sum_{m=1}^{M} L(x^*_{nm}, x^{s}_{nm}) + \gamma}{\sum_{n_1=1}^{N_1}\sum_{m=1}^{M} L(x^*_{nm}, x^{s'}_{nm}) + \gamma}}} \tag{7.6}$$

We can similarly obtain an expression for α_s by differentiating Equation 7.2 with respect to α_s and setting it to zero.

$$\boxed{\alpha_s = \frac{1}{\sum_{s'=1}^{S} \frac{\sum_{n_2=1}^{N_2}\sum_{m=1}^{M} L(u^*_{nm}, u^{s}_{nm}) + \gamma}{\sum_{n_2=1}^{N_2}\sum_{m=1}^{M} L(u^*_{nm}, u^{s'}_{nm}) + \gamma}}} \tag{7.7}$$

7.4.4 Computing Truth

As a first step to computing the truth, we first define $L(x^*_{nm}, x^s_{nm})$ for each type of data. For continuous data, the loss function is the normalized square loss of the ground truth and the fact claimed by a particular source.

$$L(x^*_{nm}, x^s_{nm}) = \frac{(x^*_{nm} - x^s_{nm})^2}{std(x^1_{nm}, x^2_{nm}, ..., x^s_{nm})} \tag{7.8}$$

For categorical data, we use a binary vector to represent all the possible values for a particular property of an object. We encode the presence of a value with a 1 and 0 otherwise. For example, if the j^{th} property of the i^{th} object has 5 possible values across s sources, we represent the vector by five bits and if source s claimed the 4^{th} value, then:

$$I_{nm}^{s} = (0,0,0,1,0)^{T} \tag{7.9}$$

Hence the loss function for categorical data is given by,

$$L(x_{nm}^{*}, x_{nm}^{s}) = (I_{nm}^{*} - I_{nm}^{s})^{T}(I_{nm}^{*} - I_{nm}^{s}) \tag{7.10}$$

Similar to computing the source weights, we solve for truth update of each task separately. We first differentiate Equation 7.2 with errors specified in Equations 7.8 and 7.9 with respect to each x_{nm}^{*}.

$$\frac{\partial f}{\partial x_{nm}^{*}} = \sum_{s=1}^{S} \beta_{s}^{2}(x_{nm}^{*} - x_{nm}^{s}) \tag{7.11}$$

By setting Equation 7.10 to zero and rearranging, we get

$$\boxed{x_{nm}^{*} = \frac{\sum_{s=1}^{S} \beta_{s}^{2} x_{nm}^{s}}{\sum_{s=1}^{S} \beta_{s}^{2}}} \tag{7.12}$$

We can similarly show that solution to update $u_{nm}*$ would be:

$$\boxed{u_{nm}^{*} = \frac{\sum_{s=1}^{S} \beta_{s}^{2} u_{nm}^{s}}{\sum_{s=1}^{S} \beta_{s}^{2}}} \tag{7.13}$$

We use the expressions from Equations 7.6, 7.7, 7.12 and 7.13 to update the source weights and truths of the two tasks until convergence.

7.4.5 Updating Truth

Updating numeric data is pretty straightforward. We update categorical data by the following manner. If we have two sources and the source weights are 0.3 and 0.7 respectively, and if the

binary vectors are (0,1) and (1,0) then the truth for the property would be the value with a higher probability. In this case, $\frac{(0,1)*0.3^2+(1,0)*0.7^2}{0.3^2+0.7^2} = (0.845, 0.155)$. So the truth is updated by the first value with a probability of 0.845.

Algorithm 3 Learning with Strong and Weak Truths

Input: $\mathbf{X}^*$, $\mathbf{U}^*$, $\{\mathbf{X}^s\}_{s=1}^S$, $\{\mathbf{U}^s\}_{s=1}^S$, γ, N_{it}, ε
Output: $\mathbf{X}^*$, $\mathbf{U}^*$
$\alpha_{s0} = \frac{1}{S}$, $\beta_{s0} = \frac{1}{S}$
for $iter$ = 1 to N_{it} **do**
    for s = 1 to S **do**
        Compute β_s as given by Equation 7.6
        Compute α_s as given by Equation 7.7
        Compute X^* as given by Equation 7.12
        Compute U^* as given by Equation 7.12
    end for
    $\|\beta_s - \beta_{s0}\| < \varepsilon$ & $\|\alpha_s - \alpha_{s0}\| < \varepsilon$
    break
    $\beta_{s0} := \beta_s$ for each $s \in [1 : S]$
    $\alpha_{s0} := \alpha_s$ for each $s \in [1 : S]$
end for
Return $\mathbf{X}^*$ and $\mathbf{U}^*$

7.4.6 Choice of Source Weight Computation

The regularization we have enforced in our method constraints the sum of the weights to be equal to 1. This constraint does not allow dissimilar sources to have a wide variation in source weights especially if the number of sources are in the thousands. Li et al, in their CRH framework propose a regularization that maps the source weights in the 0-1 range to a 0-inf range. This allows to magnify the differences between the sources. We briefly present that regularization for our framework that involves the two sets of data that are related to each other.

Parameters t_s and r_s are introduced where $t_s = \exp(-\beta_s)$ and $r_s = \exp(-\alpha_s)$. To accommodate this regularization, we rewrite Equation 7.2 in terms of the new parameters.

$$\min_{X^*,U^*,\beta_s,\alpha_s} f(X^*,U^*,\beta_s,\alpha_s) = \sum_{s=1}^{S} -\log t_s \sum_{n_1=1}^{N_1}\sum_{m=1}^{M} L(x^*_{nm},x^s_{nm})$$

$$+ \sum_{s=1}^{S} -\log r_s \sum_{n_2=1}^{N_2}\sum_{m=1}^{M} L(u^*_{nm},u^s_{nm}) + \gamma \sum_{s=1}^{S} |\log t_s - \log r_s| \tag{7.14}$$

$$+\lambda_1(\sum_{s=1}^{S} t_s - 1) + \lambda_2(\sum_{s=1}^{S} r_s - 1)$$

To solve for the source weights we differentiate Equation 7.14 with respect to t_k and set the derivative to zero, we get:

$$\sum_{n_1=1}^{N_1}\sum_{m=1}^{M} L(x^*_{nm},x^s_{nm}) + \gamma = \lambda_1 t_s \tag{7.15}$$

$$t_s = \frac{\sum_{n_1=1}^{N_1}\sum_{m=1}^{M} L(x^*_{nm},x^s_{nm}) + \gamma}{\lambda_1} \tag{7.16}$$

Based on the constrain that $\sum_{s=1}^{S} t_s = 1$, we obtain an expression for λ_1

$$\lambda_1 = \sum_{s'=1}^{S}\sum_{n_1=1}^{N_1}\sum_{m=1}^{M} L(x^*_{nm},x^{s'}_{nm}) + \gamma \tag{7.17}$$

By substituting λ_1 in Equation 7.16, we get:

$$\boxed{\beta_s = -\log\left(\frac{\sum_{n_1=1}^{N_1}\sum_{m=1}^{M} L(x^*_{nm},x^s_{nm}) + \gamma}{\sum_{s'=1}^{S}\sum_{n_1=1}^{N_1}\sum_{m=1}^{M} L(x^*_{nm},x^{s'}_{nm}) + \gamma}\right)} \tag{7.18}$$

We can obtain a similar solution for r_s.

$$\boxed{\alpha_s = -\log\left(\frac{\sum_{n_2=1}^{N_2}\sum_{m=1}^{M} L(u^*_{nm},u^s_{nm}) + \gamma}{\sum_{s'=1}^{S}\sum_{n_2=1}^{N_2}\sum_{m=1}^{M} L(u^*_{nm},u^{s'}_{nm}) + \gamma}\right)} \tag{7.19}$$

7.5 Experiments

7.5.1 Real-world Data Sets

To demonstrate the effectiveness of the proposed methods, we consider two sets of data (i) synthetic and (ii) real-world data. The details of each data set is explained below.

7.5.1.1 Simulated Data Set

We utilized two data sets from the UCI repository namely the Adult Data Set and the Bank Data Set. Both these data sets contain numeric as well as categorical data set. Although the data sets have a label and are typically used for classification, we ignore the labels and use just the data. The Adult data has 32561 objects and 14 properties. The Bank data has 45211 objects and 17 attributes. The former had missing values whereas the latter did not. The actual data was regarded as the ground truth and we generated four different versions of the data with conflict with the ground truth. These four versions are considered as four different sources.

We artificially alter the truth for each source in such a way that the first source has very few altered ground truths and the second has more conflicts than the first and so on. This would mean that the algorithm would learn a higher weight for source 1 compared to the other three sources. Source 4 would have the least weight with sources 2 and 3 having weights in between 1 and 4. For continuous data we add Gaussian noise and for categorical data, we randomly alter the truth to create a conflict. The characteristics of both these data sets are tabulated in Table 7.1.

7.5.1.2 Weather Data Set

We got the weather dataset from Qi et al., [] who collected and processed this data for their CRH framework. The data was collected from three websites namely Wunderground, World Weather Online and HAM weather. The three websites were crawled for three days to obtain the forecasts for a few US cities in terms of their high temperature, low temperature and weather condition. Since the three websites were crawled for three days, the resulting data was considered as nine

different sources. This dataset is a good representation for heterogeneous data since it has both numerical (two temperatures) and categorical data (weather condition). Similarly the ground truth for about 20 US cities was obtained from true weather for over a period of 1 month. The ground truth was obtained only for a subset of the samples.

Table 7.1: Statistics of the three data sets

	Weather Data	**Adult Data**	**Bank Data**
# Observations	16038	3646832	5787008
# Entries	2100	455854	723376
# Ground Truth	1740	455854	72336

The table information is as follow: The number of objects in Adult data is 32561 with 14 properties each which makes the number of entries per source equal to 455854. We artificially create 4 views thereby making the total number of observations to be 130244. The ground truth is known for all the observations in the case of Adult and Bank data sets.

7.5.2 Comparison with Other Methods

We compared the proposed algorithm with the following relevant methods that are similar and comparable.

- The baseline voting/average method.
- The CRH framework proposed by Qi et al., where no ground truth is used. The method is initialized with the average/majority voting for numerical/categorical data type respectively.
- Use the CRH framework with a few ground truth labels. In this case, some of the data would have the ground truth as the initializing point while the rest of them would be initialized with the average or voting rule. We call this CRH-SS.

Some other benchmark methods such as 2-Estiamtes, 3-Estimates and TruthFinder have not been compared with since all of them utilize only one data type and the CRH framework has already shown to perform better than these methods.

7.5.3 Performance Measure

For all the methods, the errors on the categorical data and numeric data are estimated based on two different strategies. We use *Error Rate* for categorical data which is the percentage of output different from the ground truth. Similarly for the continuous data, we calculate the distance of the output from the ground truth. We further normalize the error by the variance since each property might have a different range. We then calculate the mean of this normalized error called the *Mean Normalized Absolute Distance*.

7.5.4 Choosing the Ground Truth

This study explores truth discovery in a semi-supervised setting. The CRH framework does not need any ground truth. The CRH-SS and LSWT use ground truth on a small subset of entries. To do this, we randomly select these ground truths to replace the initialized entries. For the set of experiments where we show the impact of ground truth on the accuracy, we vary the number of truths from 10 to 1000. For lower number of ground truths, we tried to make sure we had both numeric and categorical data represented.

7.6 Results

We conduct two sets of experiments. The first one was to study the advantage of having a small set of known ground truths. We compared the performance of the proposed LSWT method to CRH, CRH-SS and Voting/Mean methods. The second experiment shows the effect of the number of ground truths on the performance. We did these experiments for all the three data sets.

Table 7.2: The performance of the methods with 500 known ground truths

Method	Weather		Adult		Bank	
	Error Rate	MNAD	Error Rate	MNAD	Error Rate	MNAD
CRH	0.3983	4.6849	**0.0000**	0.1014	**0.0000**	0.1027
CRH-SS	0.3983	4.6849	**0.0000**	0.1014	**0.0000**	0.1027
LSWT	**0.3660**	**4.2017**	**0.0000**	**0.0693**	**0.0000**	**0.0785**
Voting	0.4845	NA	0.2013	NA	0.2941	NA
Mean	NA	4.7853	NA	0.4632	NA	0.4891

Table 7.2 lists the errors on the three data sets. For this experiment, we chose the number of known ground truths to be 500. In all of the experiments CRH does not use any ground truths. Instead, CRH initializes the truth for its entries with the major voting for categorical entries and with the mean value for its numeric entries. CRH-SS follows the same procedure but we replace some of the entries with the ground truth. For the LSWT method, we group the entries with ground truth as one task and the bigger number of entries with no ground truth as another task whose truths are initialized with the voting/mean method.

For each data, we represent the performance with two different parameters namely the error rate for categorical data and MNAD for the numeric data. First, we take the weather data set that represents a real world data. As a quick observation, we see that the error rate and MNAD values for LSWT is the least compared to the other three methods. We see that on the categorical data, the voting approach performs the worst. This is due to the fact that if majority of the sources had a false value, voting would be in favour of it. Given that the sources may copy from each other, a wrong fact that is copied will bias the result to favour it. The same reason holds good for the MNAD values since the mean just takes a simple average of continuous data and the squared error from the ground truth might be huge. As a second observation, the CRH and CRH-SS perform the

Table 7.3: Effect of known ground truths on error rates for the weather data set

Method	Number of known truths						
	10	25	50	100	200	500	1000
CRH	**0.3983**	0.3983	0.3983	0.3983	0.3983	0.3983	0.3983
CRH+SS	**0.3983**	0.3983	0.3983	0.3983	0.3983	0.3983	0.3983
LSWT	**0.3983**	**0.3948**	**0.3879**	**0.3862**	**0.3759**	**0.3660**	**0.3641**
Voting	0.4845	0.4845	0.4845	0.4845	0.4845	0.4845	0.4845

same way irrespective of how the truth for the entries are initialized. This is because, for a convex formulation, a random starting point would still ensure convergence to the optimum solution.

As for our proposed method, we learn two different weights for the set of entries with ground truth and without ground truth. We enforce that the two weights be similar. So the weights estimated using the ground truths act as a blue print for the weights that are estimated on the entries without ground truth.

We see similar results for the Adult and Bank data sets where eight different versions of the original data set were simulated to reflect eight sources with conflict. We can see that almost all the methods accurately find the truth for the categorical data. This might be due to the fact that, categorical data were just randomly flipped. The continuous data that had Gaussian noise added was more prone to errors as shown in the results.

Table 7.4: Effect of known ground truths on MNAD values for the weather data set

Method	Number of known truths						
	10	25	50	100	200	500	1000
CRH	4.6849	4.6849	4.6849	4.6849	4.6849	4.6849	4.6849
CRH+SS	4.6849	4.6849	4.6849	4.6849	4.6849	4.6849	4.6849
LSWT	**4.5970**	**4.5233**	**4.5157**	**4.4821**	**4.3972**	**4.3948**	**4.2017**
Mean	4.7853	4.7853	4.7853	4.7853	4.7853	4.7853	4.7853

Table 7.5: Effect of known ground truths on error rate and MNAD for Adult and Bank data sets

Known Truths	Adult		Bank	
	Error Rate	MNAD	Error Rate	MNAD
100	0.0000	0.0700	0.0000	0.0796
500	0.0000	0.0693	0.0000	0.0785
1000	0.0000	0.0524	0.0000	0.0692
2000	0.0000	0.04850	0.0000	0.0598

The second set of experiments involve varying the number of ground truths to study its effect on the error. From Tables 7.3 and 7.4 we can notice that as the number of ground truth entries is increased, the error rate and MNAD values decrease for LSWT. When we have as few as 10 known truths, CRH, CRH-SS and LSWT perform similarly. But as the number of known truths is increased, the error rate and MNAD values reduce for LSWT but doesn't have any impact on CRH and CRH-SS due to the reasons discussed previously.

Table 7.5 shows the results for Adult and Bank data sets when the number of ground truths is increased. Since CRH and the Voting/averaging method do not depend on the number of ground

truths, we do not report their scores. CRH-SS on the other hand behaves like CRH for the reasons mentioned previously. We report the scores for all teh methods in table 7.2. In Table 7.5, we report the error for just LSWT. Typical to the weather data set, the MNAD values on the simulated data decrease as the number of ground truths increases. For the simulated data, we got an error rate of 0.0000 on for all the three methods - CRH, CRH-SS and LSWT. This might be due to the fact that categorical data were not altered too much while simulating the data.

7.7 Conclusion

In the field of truth finding, estimation of source reliability is the principal idea that results in efficient discovery of the truths. Among the many challenges that these problems pose the main hurdle has been to estimate parameters of a source by combining different data types. In this study, we propose a truth discovery framework that borrows inspiration from multi-task learning. We hypothesize that in a semi-supervised setting, we define entries with known ground truths and entries with unknown ground truths as two different but closely related tasks. We use two sets of parameters to estimate the source reliability and constraint the parameters to be similar. With experiments on both simulated and real-world data, we show that the proposed method performs better than other similar methods in comparison. We hope to further improve this method by incorporating a way to estimate multiple truths for each entry if any.

Chapter 8

Conclusion and Future Work

This dissertation has explored the scope of semi-supervised approaches for two main applications. We first conducted a detailed preliminary study of using multi-view learning in a semi-supervised setting along with multi-task learning for predicting associations between drugs and diseases. This empirical study opens up new avenues for exploring such algorithms since the use of unlabeled samples is desirable in computational chemistry due to the limited availability of labeled samples and a large number of unlabeled samples.

The first method we explained the weighted multi-view learning framework that weights each feature space so that the final prediction is not the simple average of the views but a weighted average of the views. The key feature of this method is that the weights are estimated without any prior knowledge on the views. The belief is that there might be some feature spaces that have better predictive power than the others and estimating them via view weights is more beneficial.

We then extend the weighted multi-view learning to multi-task with weighted multi-view learning since learning related tasks has shown to improve the overall performance of all the tasks than learning them separately. Previous studies have shown that learning related protein targets or diseases has been beneficial in the learning process. For both of the methods , we evaluated our

algorithms on drug-disease and chemical-target data respectively.

We changed gears in the last chapter to apply semi-supervised learning to truth discovery. As the name suggests, the objective of truth discovery is to mine data from heterogeneous sources that present different facts about a piece of information and find the one that is closest to the actual truth. This problem of truth discovery has been mainly approached from the view point that, estimating the reliability of a source would give an idea about the facts it presents. On the other hand, estimating the source reliability depends on the facts it presents. This challenge of having to learn source reliability along with the truths has led to the proliferation of many methods that treat the many facets of truth discovery. One aspect of estimating the trustworthiness of sources is to be able to leverage multiple data types in the source. Studies have shown that using all the data types together in estimation is more accurate than estimating each of the data type separately. The method we designed handles this in a semi-supervised setting. A lot of methods assume that no ground truth is available but on the other hand, we suggest that a small number of ground truth goes a long way in estimating the source reliabilities.

We borrowed the concept of multi-task learning from machine learning to show that the facts without ground truth and the facts with ground truth could be considered as two tasks that are very closely related. Our experiments with real world and simulated data showed that our problem setting led to more accurate finding of the truth than similar methods that were pitted against it.

Our future work would be in two different direction. The first is to incorporate a way to allow each object to have multiple truths. We need to explore the possibility that each object and its property can have one or more truths and the number of truths need not be the same across all objects and properties. The second direction is to apply the concept of truth discovery to cheminformatics. There are many public databases that have large data influx everyday. For example, experiment results of drug activity with proteins in terms of their IC_{50} value is added to the database. But due to

variations in experiments, the results are not consistent. But a very challenging aspect of such data is that, there are facts about the same object that are conflicting within the same database. Apart from conflicts between sources, we have conflicts within the source too. Modelling this aspect of such databases will substantially increase the number of useful data points that could be extracted from such sources.

The application of truth discovery to the field of computational chemistry and bioinformatics is largely unexplored. A large part solving this problem is also due to the fact that these applications do not follow an agreed upon nomenclature that makes it harder to study them. But addressing this challenge would prove to be a huge milestone in the field by converting the huge data into useful information.

References

[1] Avid M Afzal, Hamse Y Mussa, Richard E Turner, Andreas Bender, and Robert C Glen. A multi-label approach to target prediction taking ligand promiscuity into account. *Journal of cheminformatics*, 7(1):1–14, 2015.

[2] Shivani Agarwal, Deepak Dugar, and Shiladitya Sengupta. Ranking chemical structures for drug discovery: a new machine learning approach. *Journal of chemical information and modeling*, 50(5):716–731, 2010.

[3] D.K. Agrafiotis, W. Cedeno, and V.S. Lobanov. On the use of neural network ensembles in qsar and qspr. *Journal of chemical information and computer sciences*, 42(4):903–911, 2002.

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

[5] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. *Advances in neural information processing systems*, 19:41, 2007.

[6] Homa Azizian, Farzaneh Nabati, Amirhossein Sharifi, Farideh Siavoshi, Mohammad Mahdavi, and Massoud Amanlou. Large-scale virtual screening for the identification of new helicobacter pylori urease inhibitor scaffolds. *Journal of molecular modeling*, 18(7):2917–2927, 2012.

[7] SJ Barrett and WB Langdon. Advances in the application of machine learning techniques in drug discovery, design and development. In *Applications of Soft Computing*, pages 99–110. Springer, 2006.

[8] A. Ben-Hur and W.S. Noble. Kernel methods for predicting protein–protein interactions. *Bioinformatics*, 21(suppl 1):i38–i46, 2005.

[9] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, TN Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[10] Jinbo Bi, Tao Xiong, Shipeng Yu, Murat Dundar, and R Bharat Rao. An improved multi-task learning approach with applications in medical diagnosis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 117–132. Springer, 2008.

[11] Steffen Bickel, Jasmina Bogojeska, Thomas Lengauer, and Tobias Scheffer. Multi-task learning for hiv therapy screening. In *Proceedings of the 25th international conference on Machine learning*, pages 56–63. ACM, 2008.

[12] K. Bleakley and Y. Yamanishi. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*, 25(18):2397–2403, 2009.

[13] Mark S Boguski, Kenneth D Mandl, and Vikas P Sukhatme. Repurposing with a difference. *Science*, 324(5933):1394, 2009.

[14] Maria Laura Bolognesi, Andrea Cavalli, Luca Valgimigli, Manuela Bartolini, Michela Rosini, Vincenza Andrisano, Maurizio Recanatini, and Carlo Melchiorre. Multi-target-directed drug design strategy: from a dual binding site acetylcholinesterase inhibitor to a trifunctional compound against alzheimer?fs disease. *Journal of medicinal chemistry*, 50(26):6446–6449, 2007.

[15] Robert Burbidge, Matthew Trotter, B Buxton, and Sl Holden. Drug design by machine

learning: support vector machines for pharmaceutical data analysis. *Computers & chemistry*, 26(1):5–14, 2001.

[16] Evgeny Byvatov, Uli Fechner, Jens Sadowski, and Gisbert Schneider. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of chemical information and computer sciences*, 43(6):1882–1889, 2003.

[17] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. Drug target identification using side-effect similarity. *Science*, 321(5886):263–266, 2008.

[18] Rich Caruana. Multitask learning. In *Learning to learn*, pages 95–133. Springer, 1998.

[19] Rich Caruana, Shumeet Baluja, Tom Mitchell, et al. Using the future to" sort out" the present: Rankprop and multitask learning for medical risk evaluation. *Advances in neural information processing systems*, pages 959–965, 1996.

[20] Sai Nivedita Chandrasekaran, Alexios Koutsoukas, and Jun Huan. Investigating multiview and multitask learning frameworks for predicting drug-disease associations. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 138–145. ACM, 2016.

[21] O. Chapelle, P. Shivaswamy, S. Vadrevu, K. Weinberger, Y. Zhang, and B. Tseng. Multi-task learning for boosting with application to web search ranking. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1189–1198. ACM, 2010.

[22] J. Chen, J. Liu, and J. Ye. Learning incoherent sparse and low-rank patterns from multiple tasks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4):22, 2012.

[23] Feixiong Cheng, Weihua Li, Zengrui Wu, Xichuan Wang, Chen Zhang, Jie Li, Guixia Liu, and Yun Tang. Prediction of polypharmacological profiles of drugs by the integration of

chemical, side effect, and therapeutic space. *Journal of chemical information and modeling*, 53(4):753–762, 2013.

[24] Feixiong Cheng, Chuang Liu, Jing Jiang, Weiqiang Lu, Weihua Li, Guixia Liu, Weixing Zhou, Jin Huang, and Yun Tang. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol*, 8(5):e1002503, 2012.

[25] Feixiong Cheng and Zhongming Zhao. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *Journal of the American Medical Informatics Association*, 21(e2):e278–e286, 2014.

[26] Annie P Chiang and Atul J Butte. Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clinical pharmacology and therapeutics*, 86(5):507, 2009.

[27] Murat Can Cobanoglu, Chang Liu, Feizhuo Hu, Zolta?n N Oltvai, and Ivet Bahar. Predicting drug–target interactions using probabilistic matrix factorization. *Journal of chemical information and modeling*, 53(12):3399–3409, 2013.

[28] Richard D Cramer, Jeffrey D Bunce, David E Patterson, and Ildiko E Frank. Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional qsar studies. *Molecular Informatics*, 7(1):18–25, 1988.

[29] Peter Csermely, Tamás Korcsmáros, Huba JM Kiss, Gábor London, and Ruth Nussinov. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacology & therapeutics*, 138(3):333–408, 2013.

[30] Allan Peter Davis, Cynthia J Grondin, Kelley Lennon-Hopkins, Cynthia Saraceni-Richards, Daniela Sciaky, Benjamin L King, Thomas C Wiegers, and Carolyn J Mattingly. The comparative toxicogenomics database's 10th year anniversary: update 2015. *Nucleic acids research*, page gku935, 2014.

[31] Spyros N Deftereos, Christos Andronis, Ellen J Friedla, Aris Persidis, and Andreas Persidis. Drug repurposing and adverse event prediction using high-throughput literature analysis. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 3(3):323–334, 2011.

[32] C.M. Dobson. Chemical space and biology. *Nature*, 432(7019):824–828, 2004.

[33] José L Domínguez, Fernando Fernández-Nieto, Marian Castro, Marco Catto, M Rita Paleo, Silvia Porto, F Javier Sardina, José M Brea, Angelo Carotti, M Carmen Villaverde, et al. Computer-aided structure-based design of multitarget leads for alzheimer?fs disease. *Journal of chemical information and modeling*, 55(1):135–148, 2014.

[34] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–610. ACM, 2014.

[35] Xin Luna Dong, Laure Berti-Equille, Yifan Hu, and Divesh Srivastava. Global detection of complex copying relationships between sources. *Proceedings of the VLDB Endowment*, 3(1-2):1358–1369, 2010.

[36] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. Integrating conflicting data: the role of source dependence. *Proceedings of the VLDB Endowment*, 2(1):550–561, 2009.

[37] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. Truth discovery and copying detection in a dynamic world. *Proceedings of the VLDB Endowment*, 2(1):562–573, 2009.

[38] Xin Luna Dong, Barna Saha, and Divesh Srivastava. Less is more: Selecting sources wisely for integration. In *Proceedings of the VLDB Endowment*, volume 6, pages 37–48. VLDB Endowment, 2012.

[39] Xin Luna Dong and Divesh Srivastava. Compact explanation of data fusion decisions.

In *Proceedings of the 22nd international conference on World Wide Web*, pages 379–390. ACM, 2013.

[40] Jürgen Drews. Drug discovery: a historical perspective. *Science*, 287(5460):1960–1964, 2000.

[41] Joel T Dudley, Tarangini Deshpande, and Atul J Butte. Exploiting drug–disease relationships for computational drug repositioning. *Briefings in bioinformatics*, page bbr013, 2011.

[42] T. Eitrich, A. Kless, C. Druska, W. Meyer, and J. Grotendorst. Classification of highly unbalanced cyp450 data of drugs using cost sensitive machine learning techniques. *Journal of chemical information and modeling*, 47(1):92–103, 2007.

[43] Zahra Elmi, Karim Faez, Mohammad Goodarzi, and Nasser Goudarzi. Feature selection method based on fuzzy entropy for regression in qsar studies. *Molecular Physics*, 107(17):1787–1798, 2009.

[44] L Michel Espinoza-Fonseca. The benefits of the multi-target approach in drug design and discovery. *Bioorganic & medicinal chemistry*, 14(4):896–897, 2006.

[45] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117. ACM, 2004.

[46] Jianping Fan, Yuli Gao, and Hangzai Luo. Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation. *IEEE Transactions on Image Processing*, 17(3):407–426, 2008.

[47] Jiansong Fang, Yongjie Li, Rui Liu, Xiaocong Pang, Chao Li, Ranyao Yang, Yangyang He, Wenwen Lian, Ai-Lin Liu, and Guan-Hua Du. Discovery of multitarget-directed ligands against alzheimer?fs disease through systematic prediction of chemical–protein interactions. *Journal of chemical information and modeling*, 55(1):149–164, 2015.

[48] Hongliang Fei and Jun Huan. Structured feature selection and task relationship inference for multi-task learning. *Knowledge and information systems*, 35(2):345–364, 2013.

[49] Stephen B Freedman, Mark Adler, Roopa Seshadri, and Elizabeth C Powell. Oral ondansetron for gastroenteritis in a pediatric emergency department. *New England Journal of Medicine*, 354(16):1698–1705, 2006.

[50] Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. Corroborating information from disagreeing views. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 131–140. ACM, 2010.

[51] Anna Gaulton, Namrata Kale, Gerard JP van Westen, Louisa J Bellis, A Patrícia Bento, Mark Davies, Anne Hersey, George Papadatos, Mark Forster, Philip Wege, et al. The chembl bioactivity database: an update. *Scientific Data, Volume 2, Issue, pp. 150032 (2013).*, 2:150032, 2013.

[52] Hanna Geppert, Martin Vogt, and Jurgen Bajorath. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *Journal of chemical information and modeling*, 50(2):205–216, 2010.

[53] Joumana Ghosn and Yoshua Bengio. Multi-task learning for stock selection. *Advances in Neural Information Processing Systems*, pages 946–952, 1997.

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

[55] Scott M Grundy, Ivor J Benjamin, Gregory L Burke, Alan Chait, Robert H Eckel, Barbara V Howard, William Mitch, Sidney C Smith, and James R Sowers. Diabetes and cardiovascular disease a statement for healthcare professionals from the american heart association. *Circulation*, 100(10):1134–1146, 1999.

[56] Manish Gupta, Yizhou Sun, and Jiawei Han. Trust analysis with clustering. In *Proceedings of the 20th international conference companion on World wide web*, pages 53–54. ACM, 2011.

[57] F. Hammann, H. Gutmann, U. Baumann, C. Helma, and J. Drewe. Classification of cytochrome p450 activities using machine learning methods. *Molecular pharmaceutics*, 6(6):1920–1926, 2009.

[58] Dan He, David Kuhn, and Laxmi Parida. Novel applications of multitask learning and multiple output regression to multiple genetic trait prediction. *Bioinformatics*, 32(12):i37–i43, 2016.

[59] Jingrui He and Rick Lawrence. A graph-based framework for multi-task multi-view learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 25–32, 2011.

[60] Guoping Hu, Xi Li, Xianqiang Sun, Weiqiang Lu, Guixia Liu, Jin Huang, Xu Shen, and Yun Tang. Identification of old drugs as potential inhibitors of hiv-1 integrase–human ledgf/p75 interaction via molecular docking. *Journal of molecular modeling*, 18(12):4995–5003, 2012.

[61] Tyler B Hughes, Na Le Dang, Grover P Miller, and S Joshua Swamidass. Modeling reactivity to biological macromolecules with a deep multitask network. *ACS Central Science*, 2(8):529–537, 2016.

[62] MR Hurle, L Yang, Q Xie, DK Rajpal, P Sanseau, and P Agarwal. Computational drug repositioning: from data to therapeutics. *Clinical Pharmacology & Therapeutics*, 93(4), 2013.

[63] Murat Iskar, Georg Zeller, Xing-Ming Zhao, Vera van Noort, and Peer Bork. Drug discovery in the age of systems biology: the rise of computational approaches for data integration. *Current opinion in biotechnology*, 23(4):609–616, 2012.

[64] Nicolas Jacq, Vincent Breton, Hsin-Yen Chen, Li-Yung Ho, Martin Hofmann, Vinod Kasam, Hurng-Chun Lee, Yannick Legré, Simon C Lin, Astrid Maaß, et al. Virtual screening on large scale grids. *Parallel Computing*, 33(4):289–301, 2007.

[65] Bo Jiang, Feiyue Qiu, and Liping Wang. Multi-view clustering via simultaneous weighting on views and features. *Applied Soft Computing*, 47:304–315, 2016.

[66] Xin Jin, Fuzhen Zhuang, Hui Xiong, Changying Du, Ping Luo, and Qing He. Multi-task multi-view learning for heterogeneous tasks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 441–450. ACM, 2014.

[67] G Joshi-Tope, Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal, GR Gopinath, GR Wu, Lisa Matthews, et al. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl 1):D428–D432, 2005.

[68] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, et al. Kegg for linking genomes to life and the environment. *Nucleic acids research*, 36(suppl 1):D480–D484, 2008.

[69] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, 2017.

[70] Hong Kang, Zhen Sheng, Ruixin Zhu, Qi Huang, Qi Liu, and Zhiwei Cao. Virtual drug screen schema based on multiview similarity integration and ranking aggregation. *Journal of chemical information and modeling*, 52(3):834–843, 2012.

[71] WILLIAM B Kannel and DANIEL L McGee. Diabetes and cardiovascular risk factors: the framingham study. *Circulation*, 59(1):8–13, 1979.

[72] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. Pubchem substance and compound databases. *Nucleic acids research*, page gkv951, 2015.

[73] Sarah L Kinnings, Nina Liu, Peter J Tonge, Richard M Jackson, Lei Xie, and Philip E Bourne. A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *Journal of chemical information and modeling*, 51(2):408–419, 2011.

[74] Alexios Koutsoukas, Shardul Paricharak, Warren RJD Galloway, David R Spring, Adriaan P IJzerman, Robert C Glen, David Marcus, and Andreas Bender. How diverse are diversity assessment methods? a comparative analysis and benchmarking of molecular descriptor space. *Journal of chemical information and modeling*, 54(1):230–242, 2013.

[75] Meghana Kshirsagar, Jaime G Carbonell, Judith Klein-Seetharaman, and Keerthiram Murugesan. Multitask matrix completion for learning protein interactions across diseases. In *International Conference on Research in Computational Molecular Biology*, pages 53–64. Springer, 2016.

[76] Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The sider database of drugs and side effects. *Nucleic acids research*, page gkv1075, 2015.

[77] Irwin D Kuntz. Structure-based strategies for drug design and discovery. *Science*, 257(5073):1078–1082, 1992.

[78] Antonio Lavecchia. Machine-learning approaches in drug discovery: methods and applications. *Drug discovery today*, 20(3):318–331, 2015.

[79] Vivian Law, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam Maciejewski, David Arndt, Michael Wilson, Vanessa Neveu, et al. Drugbank 4.0: shedding new light on drug metabolism. *Nucleic acids research*, 42(D1):D1091–D1097, 2014.

[80] Hongwei Li, Bo Zhao, and Ariel Fuxman. The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing. In *Proceedings of the 23rd international conference on World wide web*, pages 165–176. ACM, 2014.

[81] Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. A confidence-aware approach for truth discovery on long-tail data. *Proceedings of the VLDB Endowment*, 8(4):425–436, 2014.

[82] Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan, and Jiawei Han. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1187–1198. ACM, 2014.

[83] Qingliang Li, Yanli Wang, and Stephen H Bryant. A novel method for mining highly imbalanced high-throughput screening data in pubchem. *Bioinformatics*, 25(24):3310–3316, 2009.

[84] Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. Truth finding on the deep web: Is the problem solved? In *Proceedings of the VLDB Endowment*, volume 6, pages 97–108. VLDB Endowment, 2012.

[85] Yaliang Li, Qi Li, Jing Gao, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. On the discovery of evolving truth. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 675–684. ACM, 2015.

[86] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 64:4–17, 2012.

[87] Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N Jorissen, and Michael K Gilson. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, 35(suppl 1):D198–D201, 2007.

[88] Xuan Liu, Xin Luna Dong, Beng Chin Ooi, and Divesh Srivastava. Online data fusion. *Proceedings of the VLDB Endowment*, 4(11):932–943, 2011.

[89] Ying Liu. Active learning with support vector machine applied to gene expression data for cancer classification. *Journal of chemical information and computer sciences*, 44(6):1936–1941, 2004.

[90] Heng Luo, W Mattes, DL Mendrick, and H Hong. Molecular docking for identification of potential targets for drug repurposing. *Current topics in medicinal chemistry*, 2016.

[91] Fenglong Ma, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei Han. Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 745–754. ACM, 2015.

[92] José L Medina-Franco, Marc A Giulianotti, Gregory S Welmaker, and Richard A Houghten. Shifting from the single to the multitarget paradigm in drug discovery. *Drug discovery today*, 18(9):495–501, 2013.

[93] Fantine Mordelet and Jean-Philippe Vert. Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples. *BMC bioinformatics*, 12(1):1, 2011.

[94] Subhabrata Mukherjee, Gerhard Weikum, and Cristian Danescu-Niculescu-Mizil. People on drugs: credibility of user statements in health communities. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 65–74. ACM, 2014.

[95] Ion Muslea, Steven Minton, and Craig A Knoblock. Active+ semi-supervised learning= robust multi-view learning. In *ICML*, volume 2, pages 435–442, 2002.

[96] Xia Ning, Huzefa Rangwala, and George Karypis. Multi-assay-based structure- activity relationship models: improving structure- activity relationship models by incorporating activity information from related targets. *Journal of chemical information and modeling*, 49(11):2444–2456, 2009.

[97] Xia Ning, Michael Walters, and George Karypisxy. Improved machine learning models for predicting selective compounds. *Journal of chemical information and modeling*, 52(1):38–50, 2011.

[98] Ulf Norinder. Support vector machine models in drug design: applications to drug transport processes and qsar using simplex optimisations and variable selection. *Neurocomputing*, 55(1):337–346, 2003.

[99] TI Oprea and J Mestres. Drug repurposing: far beyond new targets for old drugs. *The AAPS journal*, 14(4):759–763, 2012.

[100] Tudor I Oprea, Sonny Kim Nielsen, Oleg Ursu, Jeremy J Yang, Olivier Taboureau, Stephen L Mathias, Irene Kouskoumvekaki, Larry A Sklar, and Cristian G Bologa. Associating drugs, targets and clinical outcomes into an integrated network affords a new platform for computer-aided drug repurposing. *Molecular informatics*, 30(2-3):100–111, 2011.

[101] Ainslie B Parsons, Renée L Brost, Huiming Ding, Zhijian Li, Chaoying Zhang, Bilal Sheikh, Grant W Brown, Patricia M Kane, Timothy R Hughes, and Charles Boone. Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nature biotechnology*, 22(1):62–69, 2004.

[102] Jeff Pasternack and Dan Roth. Knowing what to believe (when you already know something). In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 877–885. Association for Computational Linguistics, 2010.

[103] Ravali Pochampally, Anish Das Sarma, Xin Luna Dong, Alexandra Meliou, and Divesh

Srivastava. Fusing data with correlations. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 433–444. ACM, 2014.

[104] Theodoros Rekatsinas, Xin Luna Dong, and Divesh Srivastava. Characterizing and selecting fresh data sources. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 919–930. ACM, 2014.

[105] S Sans, H Kesteloot, and D obot Kromhout. The burden of cardiovascular diseases mortality in europe. *European heart journal*, 18(8):1231–1248, 1997.

[106] Philippe Sanseau and Jacob Koehler. Editorial: computational methods for drug repurposing, 2011.

[107] Anish Das Sarma, Xin Luna Dong, and Alon Halevy. Data integration with dependent sources. In *Proceedings of the 14th International Conference on Extending Database Technology*, pages 401–412. ACM, 2011.

[108] Madhavi Sastry, Jeffrey F Lowrie, Steven L Dixon, and Woody Sherman. Large-scale systematic analysis of 2d fingerprint methods and parameters to improve virtual screening enrichments. *Journal of chemical information and modeling*, 50(5):771–784, 2010.

[109] David B Searls. Data integration: challenges for drug discovery. *Nature reviews Drug discovery*, 4(1):45–58, 2005.

[110] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

[111] Tripti Swarnkar, Sergio Nery Simoes, David Correa Martins, Anji Anura, Helena Brentani, Ronaldo Fumio Hashimoto, and Pabitra Mitra. Multiview clustering on ppi network for

gene selection and enrichment from microarray data. In *Bioinformatics and Bioengineering (BIBE), 2014 IEEE International Conference on*, pages 15–22. IEEE, 2014.

[112] Michalis K Titsias and Miguel Lázaro-Gredilla. Spike and slab variational inference for multi-task and multiple kernel learning. In *Advances in neural information processing systems*, pages 2339–2347, 2011.

[113] Grigorios Tzortzis and Aristidis Likas. Kernel-based weighted multi-view clustering. In *2012 IEEE 12th International Conference on Data Mining*, pages 675–684. IEEE, 2012.

[114] Alexandre Varnek and Igor Baskin. Machine learning methods for property prediction in chemoinformatics: quo vadis? *Journal of chemical information and modeling*, 52(6):1413–1437, 2012.

[115] Alexandre Varnek, Cedric Gaudin, Gilles Marcou, Igor Baskin, Anil Kumar Pandey, and Igor V Tetko. Inductive transfer of knowledge: application of multi-task learning and feature net approaches to model tissue-air partition coefficients. *Journal of chemical information and modeling*, 49(1):133–144, 2009.

[116] Santiago Vilar, Giorgio Cozza, and Stefano Moro. Medicinal chemistry and the molecular operating environment (moe): application of qsar and molecular docking to drug discovery. *Current topics in medicinal chemistry*, 8(18):1555–1572, 2008.

[117] Chris L Waller, Ajay Shah, and Matthias Nolte. Strategies to support drug discovery through integration of systems and data. *Drug discovery today*, 12(15):634–639, 2007.

[118] Dong Wang, Tarek Abdelzaher, Lance Kaplan, and Charu C Aggarwal. Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications. In *Distributed Computing Systems (ICDCS), 2013 IEEE 33rd International Conference on*, pages 530–539. IEEE, 2013.

[119] Dong Wang, Md Tanvir Amin, Shen Li, Tarek Abdelzaher, Lance Kaplan, Siyu Gu, Chenji Pan, Hengchang Liu, Charu C Aggarwal, Raghu Ganti, et al. Using humans as sensors: an estimation-theoretic perspective. In *Information Processing in Sensor Networks, IPSN-14 Proceedings of the 13th International Symposium on*, pages 35–46. IEEE, 2014.

[120] Dong Wang, Lance Kaplan, and Tarek F Abdelzaher. Maximum likelihood analysis of conflicting observations in social sensing. *ACM Transactions on Sensor Networks (ToSN)*, 10(2):30, 2014.

[121] Shiguang Wang, Dong Wang, Lu Su, Lance Kaplan, and Tarek F Abdelzaher. Towards cyber-physical systems in social spaces: The data reliability challenge. In *Real-Time Systems Symposium (RTSS), 2014 IEEE*, pages 74–85. IEEE, 2014.

[122] Yong-Hua Wang, Yan Li, Sheng-Li Yang, and Ling Yang. Classification of substrates and inhibitors of p-glycoprotein using unsupervised machine learning approach. *Journal of chemical information and modeling*, 45(3):750–757, 2005.

[123] Manfred K Warmuth, Jun Liao, Gunnar Rätsch, Michael Mathieson, Santosh Putta, and Christian Lemmen. Active learning with support vector machines in the drug discovery process. *Journal of chemical information and computer sciences*, 43(2):667–673, 2003.

[124] Bohdan Waszkowycz, Tim D. J. Perkins, Richard A. Sykes, and Jin Li. Large-scale virtual screening for discovering leads in the postgenomic era. *IBM Systems Journal*, 40(2):360, 2001.

[125] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, pages 2035–2043, 2009.

[126] Yu-Meng Xu, Chang-Dong Wang, and Jian-Huang Lai. Weighted multi-view clustering with feature selection. *Pattern Recognition*, 53:25–35, 2016.

[127] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):i232–i240, 2008.

[128] Yoshihiro Yamanishi, Masaaki Kotera, Minoru Kanehisa, and Susumu Goto. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, 26(12):i246–i254, 2010.

[129] Yoshihiro Yamanishi, Edouard Pauwels, and Masaaki Kotera. Drug side-effect prediction based on the integration of chemical and biological spaces. *Journal of chemical information and modeling*, 52(12):3284–3292, 2012.

[130] Jihong Yang, Zheng Li, Xiaohui Fan, and Yiyu Cheng. Drug–disease association and drug-repositioning predictions in complex diseases using causal inference–probabilistic matrix factorization. *Journal of chemical information and modeling*, 54(9):2562–2569, 2014.

[131] XJ Yao, Annick Panaye, Jean-Pierre Doucet, RS Zhang, HF Chen, MC Liu, ZD Hu, and Bo Tao Fan. Comparative study of qsar/qspr correlations using support vector machines, radial basis function neural networks, and multiple linear regression. *Journal of chemical information and computer sciences*, 44(4):1257–1266, 2004.

[132] Xiaoxin Yin, Jiawei Han, and S Yu Philip. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6):796–808, 2008.

[133] Xiaoxin Yin and Wenzhao Tan. Semi-supervised truth discovery. In *Proceedings of the 20th international conference on World wide web*, pages 217–226. ACM, 2011.

[134] Shi Yu, Léon-Charles Tranchevent, Bart De Moor, and Yves Moreau. Multi-view text mining for disease gene prioritization and clustering. In *Kernel-based Data Fusion for Machine Learning*, pages 109–144. Springer, 2011.

[135] Han Yuan, Ivan Paskov, Hristo Paskov, Alvaro J González, and Christina S Leslie. Multitask learning improves prediction of cancer drug sensitivity. *Scientific reports*, 6, 2016.

[136] Salim Yusuf, Srinath Reddy, Stephanie Ôunpuu, and Sonia Anand. Global burden of cardiovascular diseases part i: general considerations, the epidemiologic transition, risk factors, and impact of urbanization. *Circulation*, 104(22):2746–2753, 2001.

[137] Jintao Zhang. Multi-task and multi-view learning for predicting adverse drug reactions. 2012.

[138] Jintao Zhang and Jun Huan. Inductive multi-task learning with multiple view data. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 543–551. ACM, 2012.

[139] Jintao Zhang, Gerald H Lushington, and Jun Huan. Multi-target protein-chemical interaction prediction using task-regularized and boosted multi-task learning. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pages 60–67. ACM, 2012.

[140] Lei Zhang, Shupeng Wang, Xiaoyu Zhang, Dinggang Shen, and Shuiwang Ji. Collaborative multi-view denoising.

[141] Li Zhang, Sundeep Vaddadi, Hailin Jin, and Shree K Nayar. Multiple view image denoising. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1542–1549. IEEE, 2009.

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

[143] Wei Zhang, Lijuan Ji, Yanan Chen, Kailin Tang, Haiping Wang, Ruixin Zhu, Wei Jia, Zhiwei Cao, and Qi Liu. When drug discovery meets web search: Learning to rank for ligand-based virtual screening. *J. Cheminformatics*, 7:5, 2015.

[144] Bo Zhao and Jiawei Han. A probabilistic model for estimating real-valued truth from conflicting sources. *Proc. of QDB*, 2012.

[145] Bo Zhao, Benjamin IP Rubinstein, Jim Gemmell, and Jiawei Han. A bayesian approach to discovering truth from conflicting sources for data integration. *Proceedings of the VLDB Endowment*, 5(6):550–561, 2012.

[146] Shi Zhi, Bo Zhao, Wenzhu Tong, Jing Gao, Dian Yu, Heng Ji, and Jiawei Han. Modeling truth existence in truth discovery. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1543–1552. ACM, 2015.

[147] Yan-Ping Zhou, Jian-Hui Jiang, Wei-Qi Lin, Hong-Yan Zou, Hai-Long Wu, Guo-Li Shen, and Ru-Qin Yu. Boosting support vector regression in qsar studies of bioactivities of chemical compounds. *european journal of pharmaceutical sciences*, 28(4):344–353, 2006.

[148] Grant R Zimmermann, Joseph Lehar, and Curtis T Keith. Multi-target therapeutics: when the whole is greater than the sum of the parts. *Drug discovery today*, 12(1):34–42, 2007.

[149] Jure Zupan and Johann Gasteiger. *Neural networks in chemistry and drug design*. John Wiley & Sons, Inc., 1999.